

SambaNova DataScale SN30 Overview

April 30, 2024



The image shows a server rack with SambaNova branding. Overlaid on the rack are several blue panels representing different capabilities:

- Top Panel:** A flowchart showing a sequence of steps. Below it are three tabs: **VISION** (with an eye icon), **LANGUAGE** (with a globe icon), and **RECOMMENDATION** (with a thumbs-up icon). The text "SambaNova Dataflow-as-a-Service™" is at the bottom.
- Right Panel:** Text: "APIs, CLI, web browser Python SDK". Below the text are icons representing a network graph and a server rack.
- Bottom-Left Panel:** Title: "Training and Inference". It contains two line graphs. The left graph shows accuracy over 8 epochs, and the right graph shows loss over 8 epochs. The text "DataScale" is written vertically on the right side of this panel.
- Bottom-Right Panel:** Title: "Deployment Options". It shows icons for server racks and a cloud icon. Below the icons are the labels "ON PREMISE" and "CLOUD COLOCATION".

Safe Harbor Statement

The following is intended to outline our general product direction at this time. There is no obligation to update this presentation and the Company's products and direction are always subject to change. This presentation is intended for information purposes only and may not be relied upon for any purchasing, partnership, or other decisions.

Agenda

- Core Technology Stack
- Cardinal SN30 Details
- Dataflow Architecture for Large Deployments
- Other Announcements

SambaNova: Long-term leader in enterprise AI

Snapshot

- Founded in 2017
- Full-stack solution for enterprise AI: AI chips to AI models
- \$1B+ funding raised

Founded by pioneers in AI



Rodrigo Liang
Co-founder & CEO



Kunle Olukotun
Co-founder & Chief Technologist



Christopher Ré
Co-founder

Sophisticated, long-term investors



The SambaNova Foundation Model Platform

Innovation at every layer of the stack

SambaNova Suite



as-a-SERVICE
Pre-trained Foundation Models

SYSTEMS
DataScale

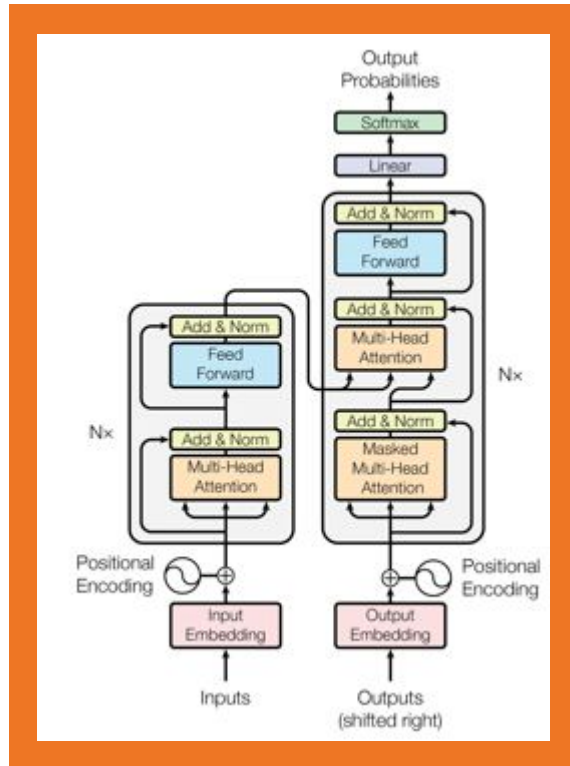
SOFTWARE
SambaFlow™

SILICON
RDU

DataScale®



AI Is Transforming Software – Models Are The New Code

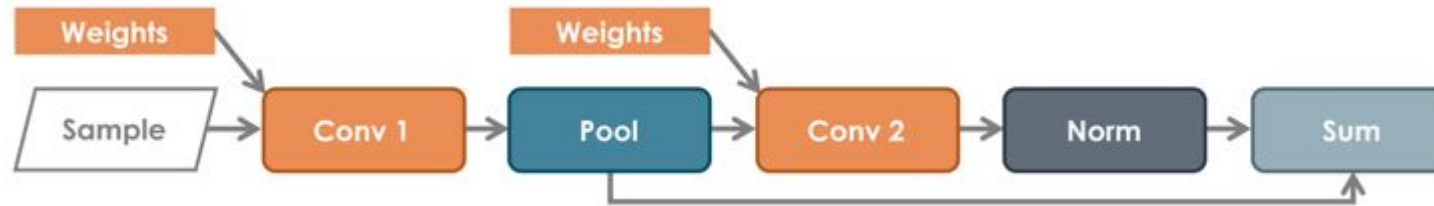


Deep Learning Enablers

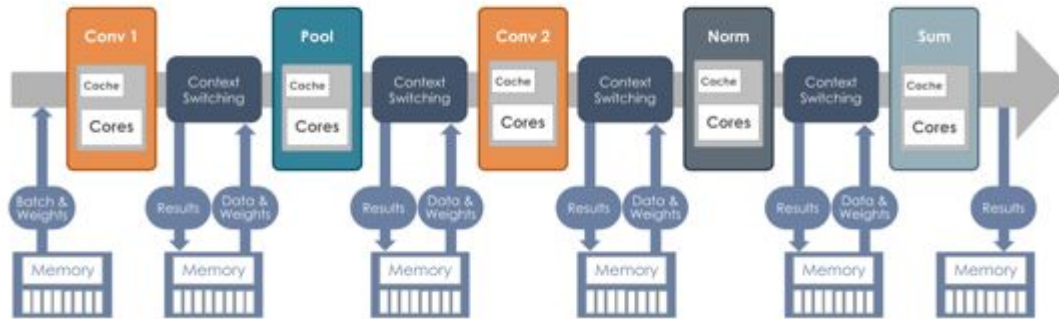
- **Compute**
 - + Commodity – Universally provided.
- **Memory Capacity**
 - + This is huge pain point!
- **Dataflow**
 - + Does not exist in SOTA architectures.
 - + Silently dilutes effective compute!

SambaNova RDA: Compute-Efficiency and Memory-Capacity Using Dataflow

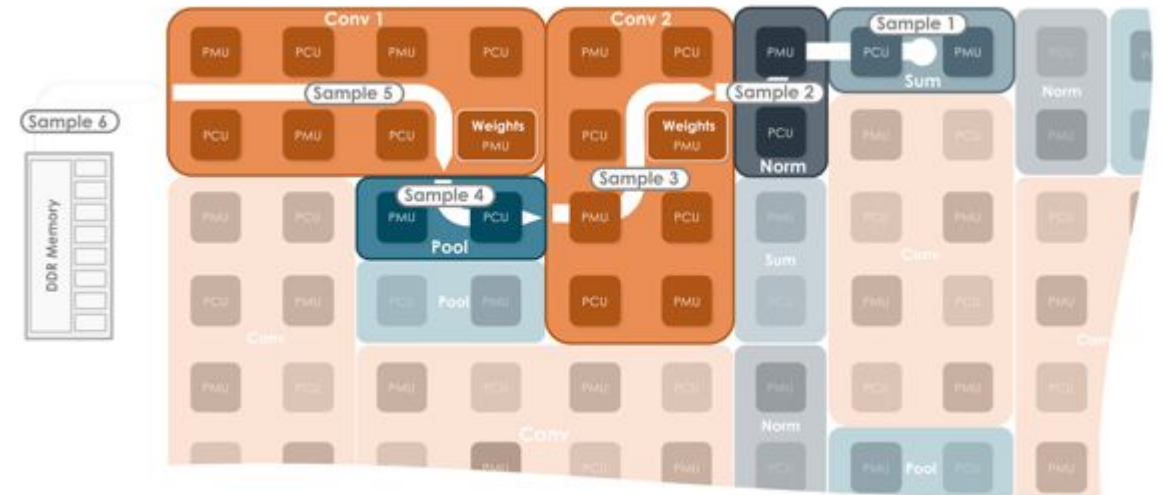
Spatial Dataflow Within an RDU



Simple Convolution Graph



The old way: kernel-by-kernel
Bottlenecked by memory bandwidth
and host overhead



The Dataflow way: Spatial
Eliminates memory traffic and overhead

SambaNova Cardinal SN30 RDU



Cardinal SN30™
Reconfigurable Dataflow Unit™

- 7nm TSMC, 86B transistors
- 102 km of wire
- 640 MB on-chip,
1,024 GB external
- 688 TFLOPS (bf16)
- RDU-Connect™

as-a-SERVICE

Pre-trained Foundation
Models

SYSTEMS

DataScale®

SOFTWARE

SambaFlow™

SILICON

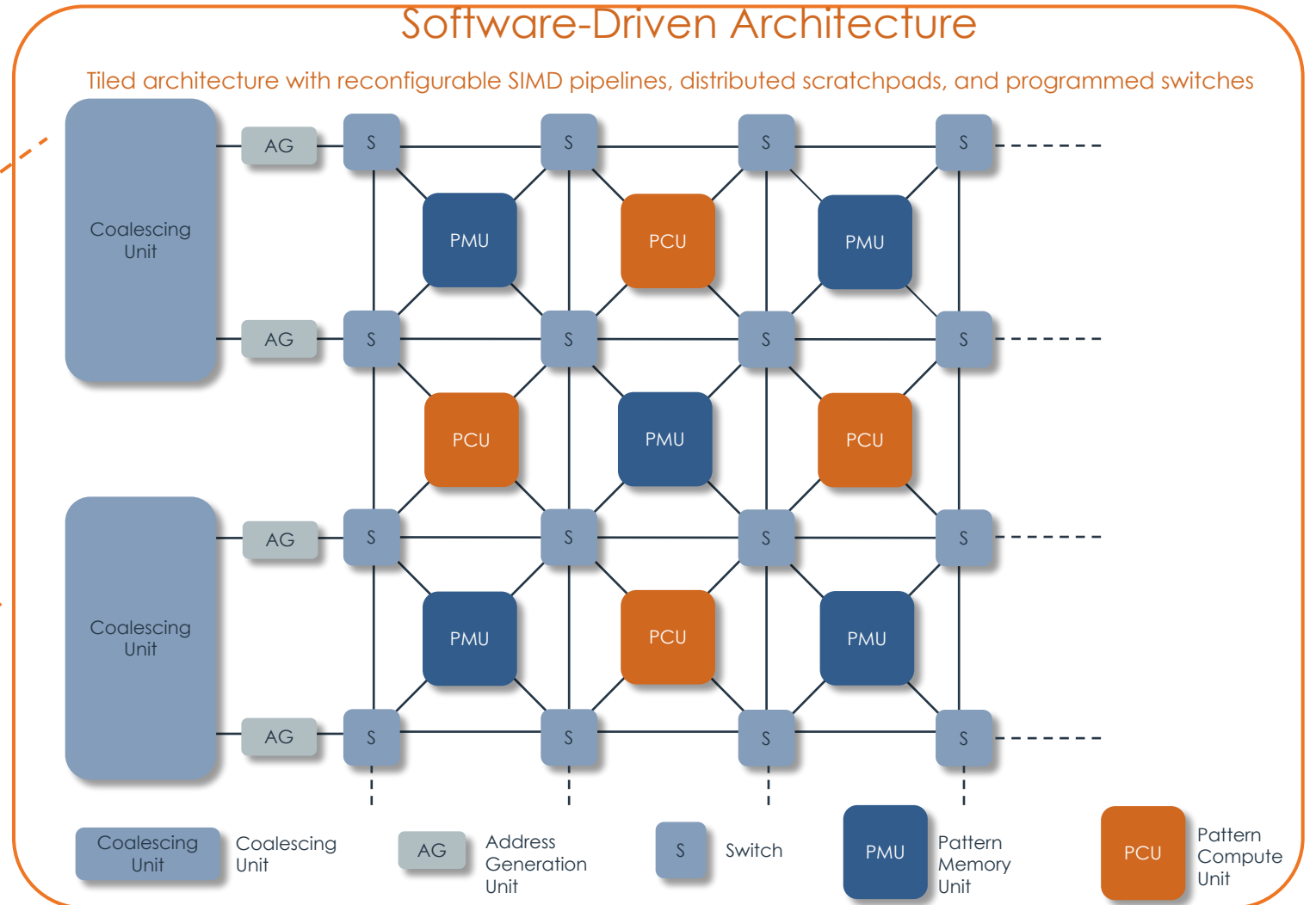
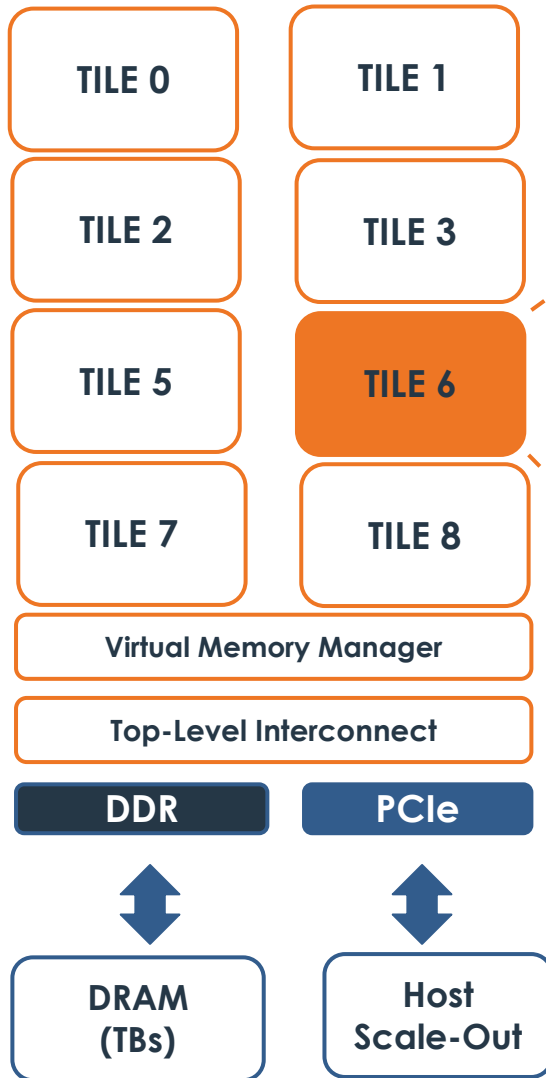
RDU

Cardinal SN30: Chip and Architecture Overview

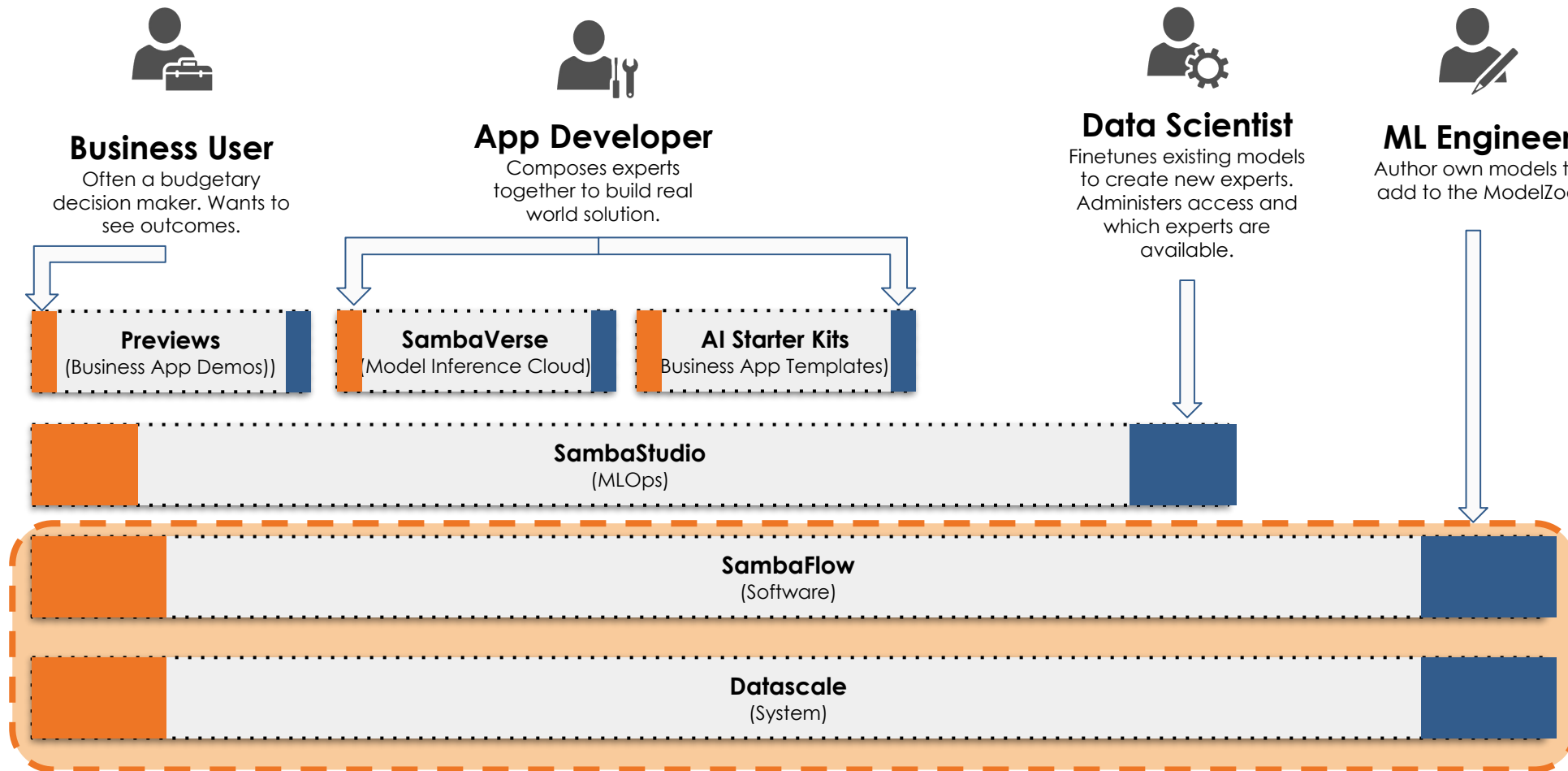


- RDU broken up into 8-tiles
 - 160 PMU and PCUs per tile
 - Additional sub-components like coalescing units (CU) for connectivity to other tiles and off-chip components, switches to set up communication between PMU, PCUs, and CU
- Tile resource management: Combined or independent mode
 - Combined: Combine adjacent to form a larger logical tile for one application
 - Independent: Each tile controlled independently, allows running different applications on separate tiles concurrently.
- Direct access to TBs of DDR4 off-chip memory
- Memory-mapped access to host memory
- Scale-out communication support

Cardinal SN30: Tile



SambaNova Software Stack



as-a-SERVICE
Pre-trained
Foundation Models

SYSTEMS
DataScale®

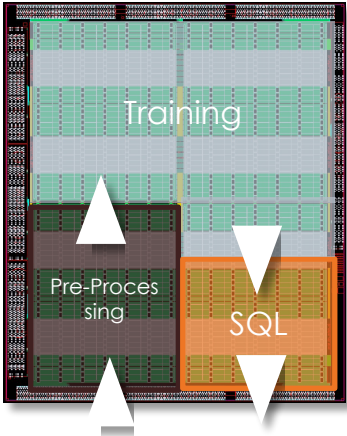
SOFTWARE
SambaFlow™

SILICON
RDU

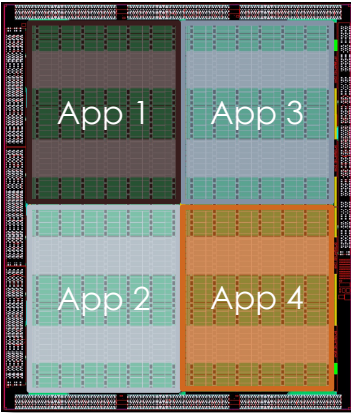
SambaNova Systems Flexibility to Support Key Scenarios

4 RDU deployment examples

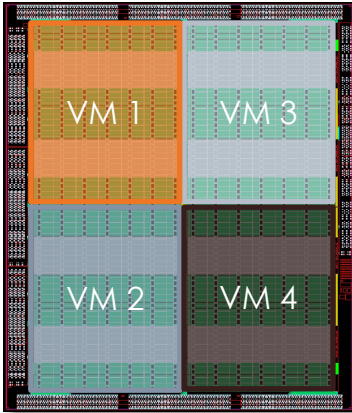
1) High Performance Mixed Workloads



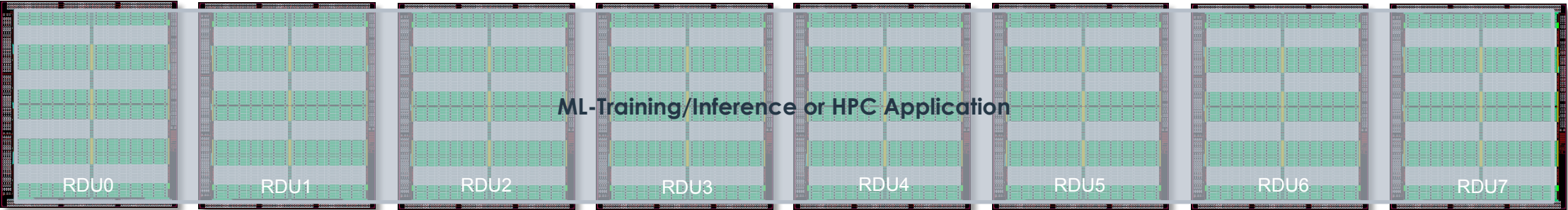
2) Efficient Concurrent Applications



3) Secure Multi-Tenancy



4) Compiler Driven Application Scale-Up



SambaNova DataScale SN30



DataScale SN30

- Rack optimized, integrated system
- 10 RU
- 8S nodes, 8 TB DRAM
- Powered by SambaNova Cardinal SN30™ RDU
- Can be installed in minutes

as-a-SERVICE

Pre-trained Foundation Models

SYSTEMS

DataScale®

SOFTWARE

SambaFlow™

SILICON

RDU

SambaNova DataScale SN30-8 System



- 8 x Cardinal SN30 Reconfigurable Dataflow Unit
- 8 TB total memory (using 64 x 128 GB DDR4 DIMMs)
- 6 x 3.8 TB NVMe (22.8 TB total)
- PCIe Gen4 x16
- Host module

10 Domains

Finance, Legal, Medical, Tabular Data Analysis, Math, Coding, General, API usage, AI Safety

1.3T Parameters

Diverse Set of Tasks

Chat, Text to SQL, Code generation, Moderation, Function/API Calling, Multilinguality, Table Interpretation, Chart QA, Image QA, Writing Assistance, and more

30+ Languages

English, Spanish, French, Japanese, Thai, Arabic, Hungarian, Turkish, Hindi, Russian, and more



54 Experts

7 Foundation Model Architectures

Llama, Mistral, Falcon, Bloom, Llava, DePlot, CLIP

Samba-1 CoE

Try It: <https://fast.snova.ai/>

Samba-1

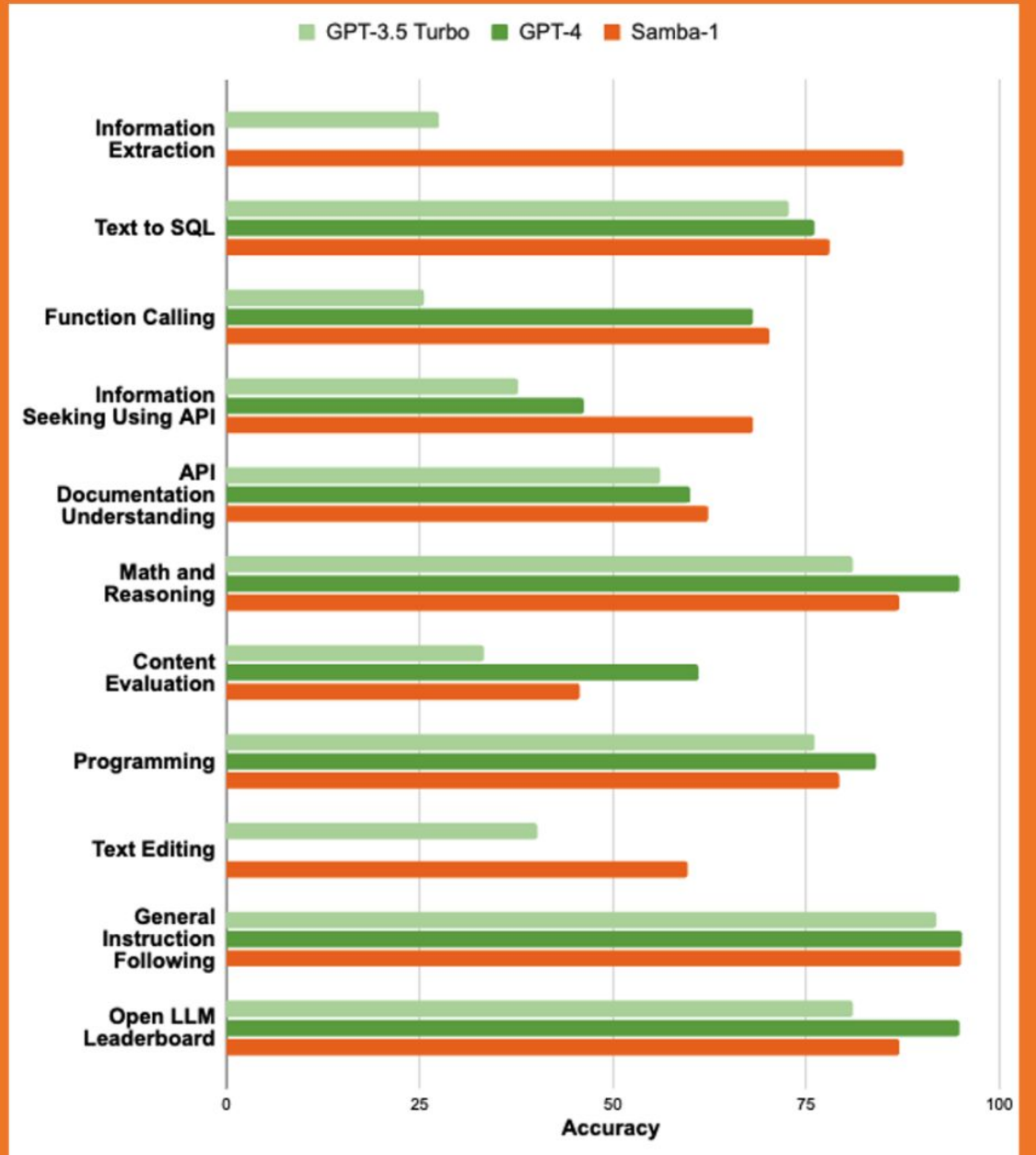
Enterprise-Grade AI Benchmark (EGAI) Tailors Large Language Model (LLM) development to enterprise-specific needs by focusing on benchmarks that matter most to business use cases

Samba-1: Matches or surpasses state-of-the-art closed-source models on EGAI benchmarks. It is a framework that can further adapt to private enterprise data, enhancing performance beyond generic models.

Try It: <https://fast.snova.ai/>

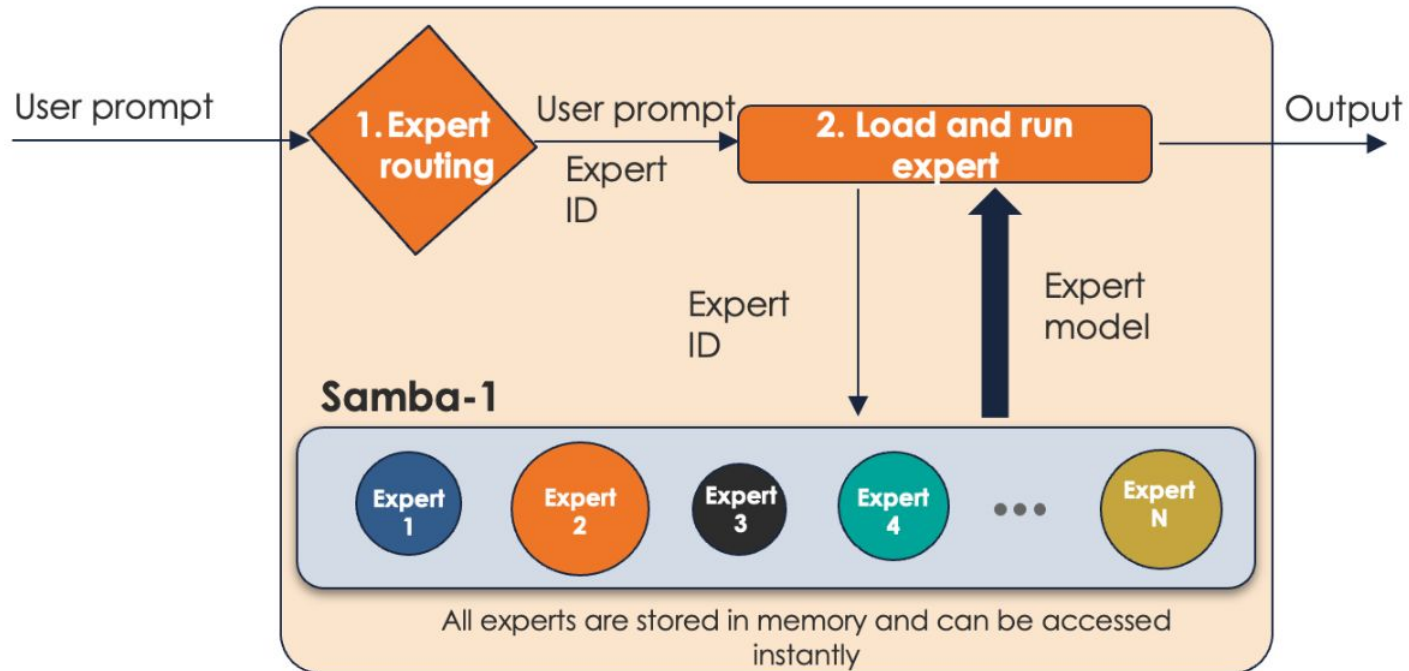
Details:

<https://sambanova.ai/blog/benchmarking-samba-1>



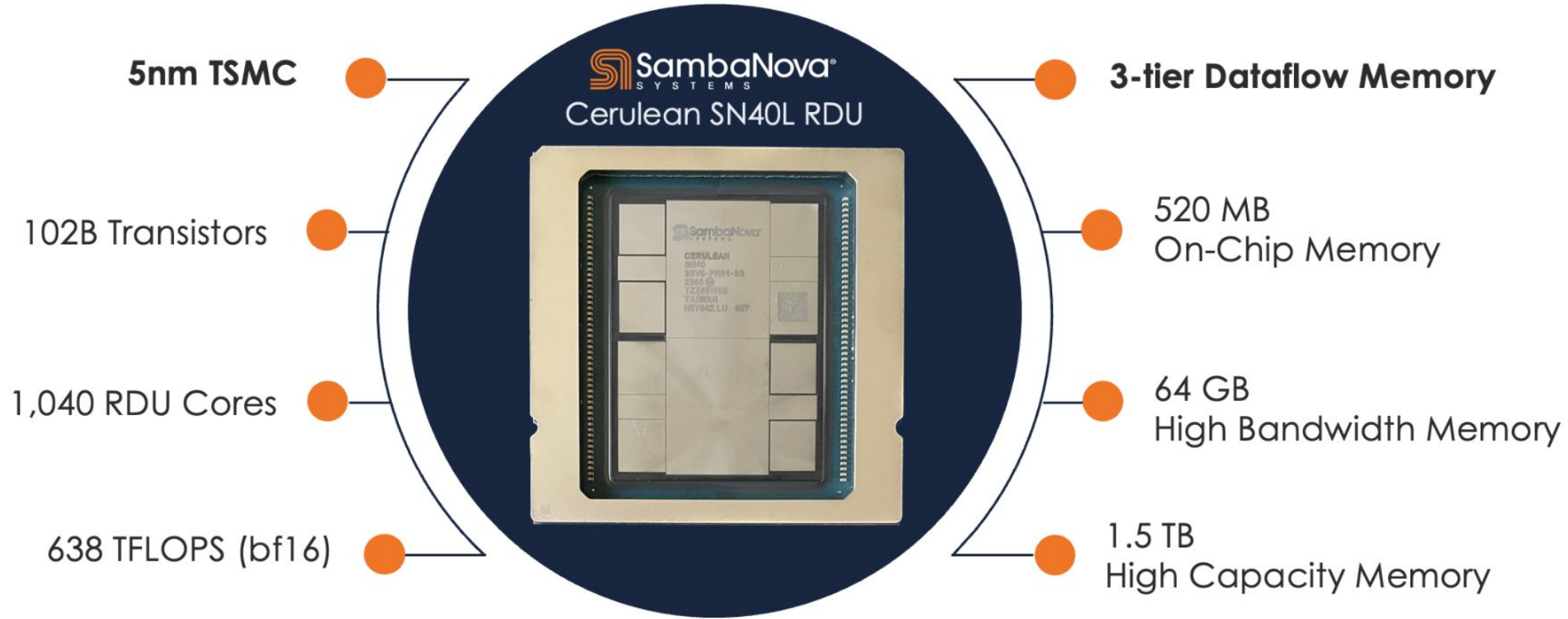
Routing in CoE

A CoE router predicts the best expert(s) for the most accurate response to a prompt



SN40L: SambaNova's new CoE-optimized RDU

"Cerulean" Architecture-based Reconfigurable Dataflow Unit

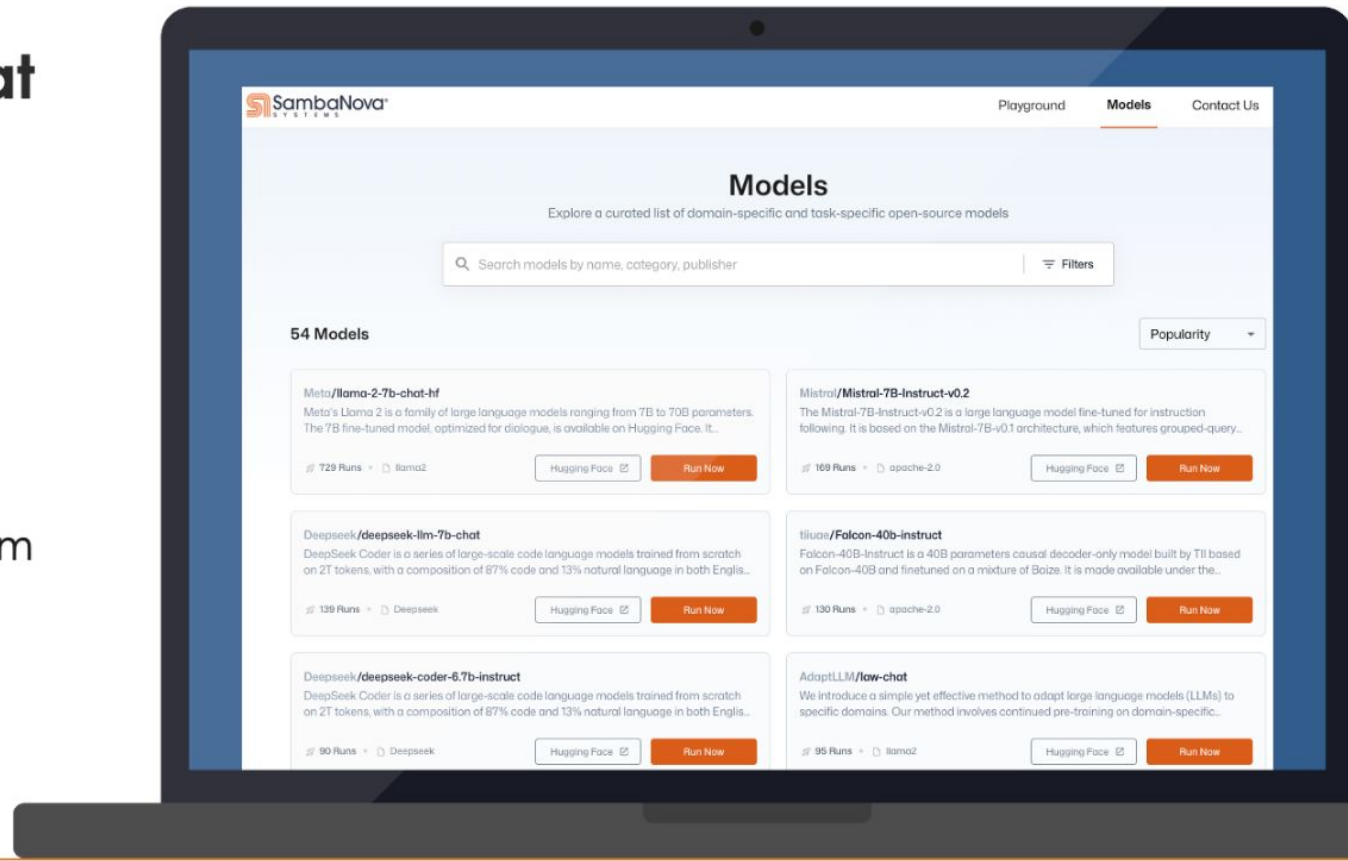


Generative AI Training and Inference

Introducing Sambaverse

The **POWER** of **Samba-1** at a developer's fingertips

- Explore a curated list of top open-source models from Hugging Face
- Test with your Prompts for free
- Find the best fit for your problem statement
- Build complex, multi-expert workflows on top of Samba-1



Introducing Samba Apps

Experience AI-enabled apps powered by the SambaNova Suite **for free!**

Apps available at launch:

SambaChat — Experience the future of conversational AI with a CoE powered assistant

FinSherlock — Unlock insights into the S&P 500 based on each companies 10-K report

DocSage (coming soon) — Extract knowledge from your PDFs!

Samba Apps ^{Beta}


Experience Enterprise grade AI solutions powered By [SambaNova Suite](#) and Samba-1 (CoE)

SambaNova Suite enables you to build, deploy, and manage your own AI solutions using top-performing expert models from the open-source community.

Samba-1 is a 1T parameter Composition of Experts comprised of strategically curated expert models from the open source community that enables anyone to create limitless applications with 10x greater inference performance(10x greater inference performance* for your organization).


We chose a subset of experts to build these Samba Apps, **you can choose them all!**

[GET STARTED](#)


SambaChat ^{Beta}
COE Powered Conversation Symphony


Elevate your conversations with AI-powered assistance.

Spark creativity and generate ideas. Craft compelling content and master the art of dialogue.


FinSherlock ^{Beta}
Financial AI assistant for 10K filings

Your trusted companion in unlocking financial jargon. Easily inquire about 10-K documents of S&P 500 organizations through intuitive Q&A.

Let AI be your guide through the intricate world of finance.


DocSage ^{Beta}
Bring Your Own Data for QnA

Turn PDFs into intelligence. Empower your business with CoE-powered knowledge extraction.

Use AI and retrieval augmented generation (RAG) to do QnA on your uploaded documents.

By using SambaApps you agree to the [Terms of Use](#)

Launched in Beta, continuously improved with user feedback



SambaNova®

S Y S T E M S