SambaNova DataScale SN30 & Platform Architecture

July 17th, 2024

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

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

The SambaNova Foundation Model Platform

Innovation at every layer of the stack

SambaNova Suite

as-a-SERVICE
Pre-trained Foundation Models

SYSTEMS
DataScale

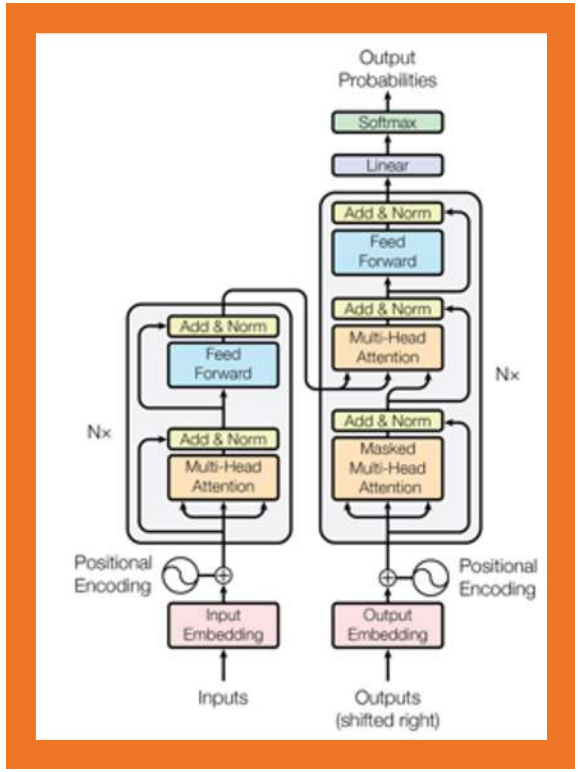
SOFTWARE
SambaFlow™

SILICON
RDU

DataScale®

SambaNova SYSTEMS

AI Is Transforming Software – Models Are The New Code

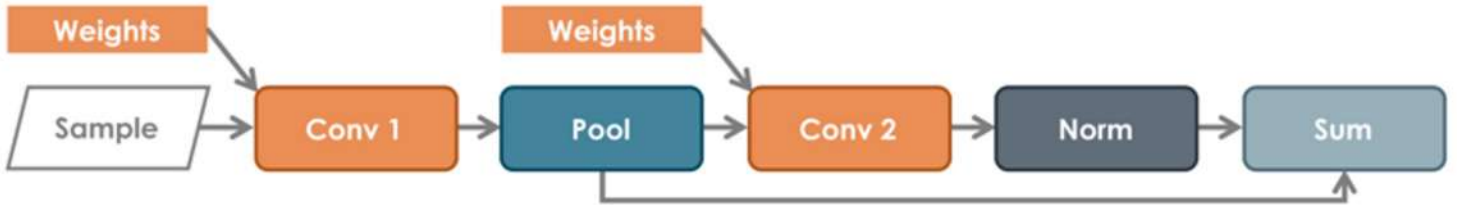


Deep Learning Enablers

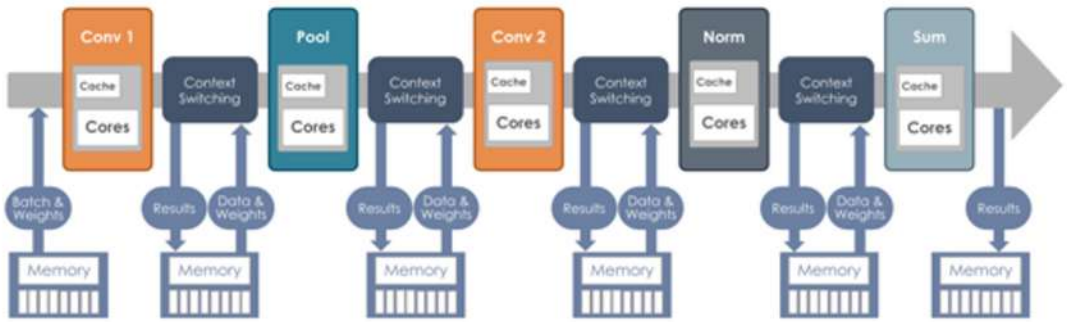
- **Compute**
  - + Commodity – Universally provided.
- **Memory Capacity**
  - + This is huge pain point!
- **Dataflow**
  - + Does not exist in SOTA architectures.
  - + Silently dilutes effective compute!

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 SYSTEMS

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

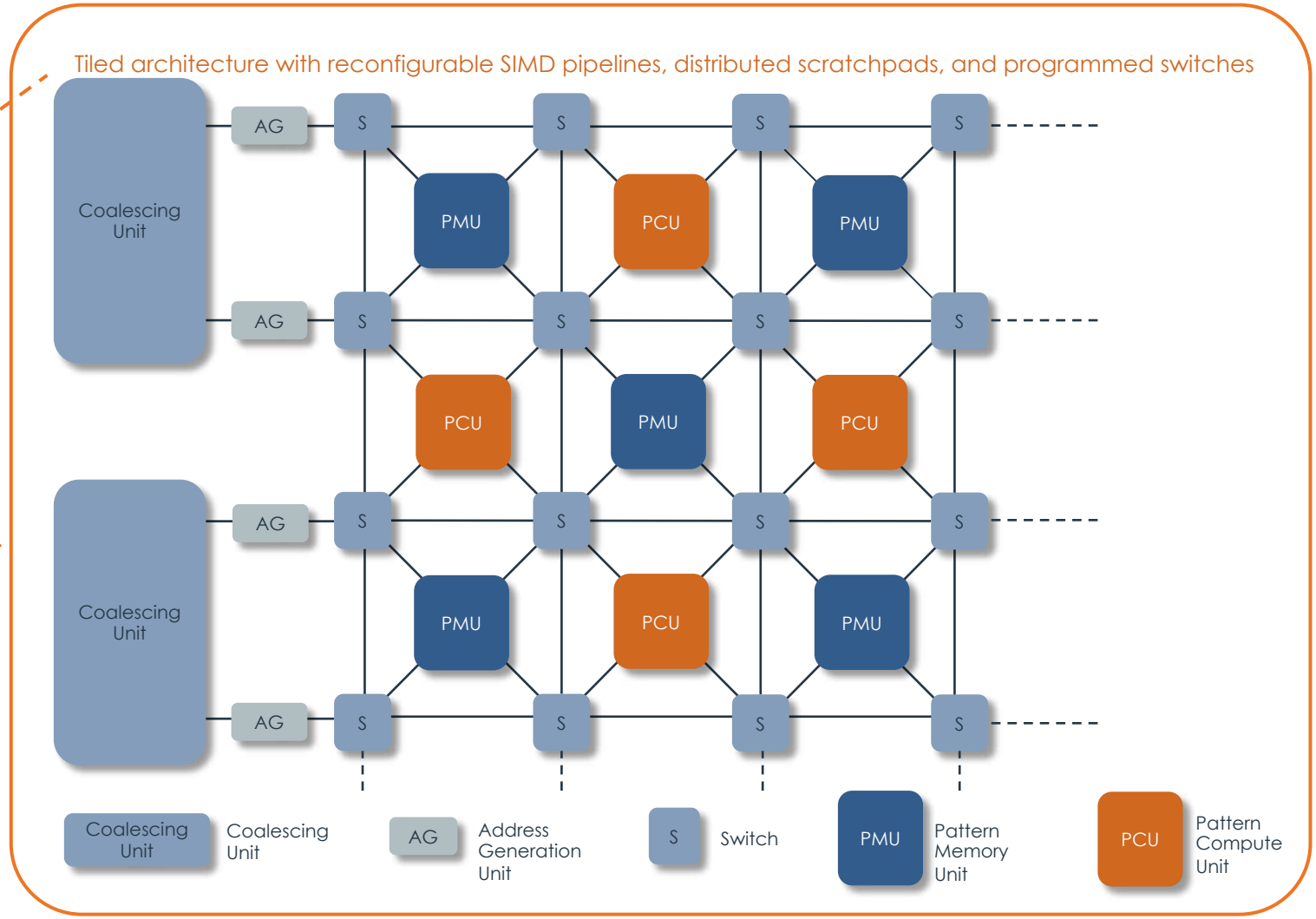
SYSTEMS
DataScale®

SOFTWARE
SambaFlow™

SILICON
RDU

SambaNova SYSTEMS

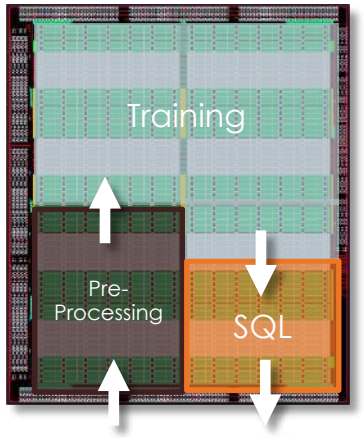
Cardinal SN30: Chip and Architecture Overview



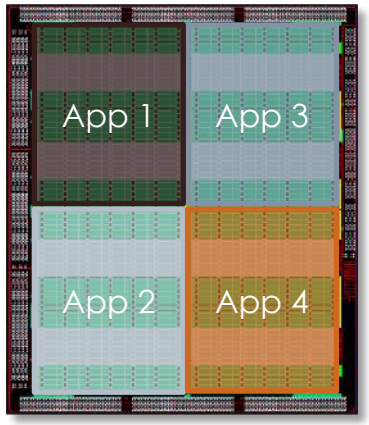
SambaNova Systems Flexibility to Support Key Scenarios

Multi-tenancy and Multi-user support

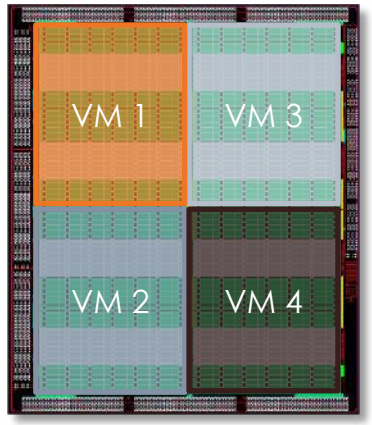
1) High Performance Mixed Workloads



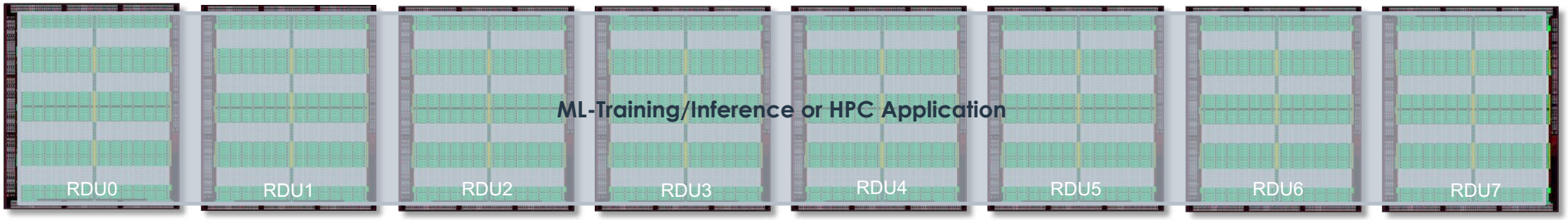
2) Efficient Concurrent Applications



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

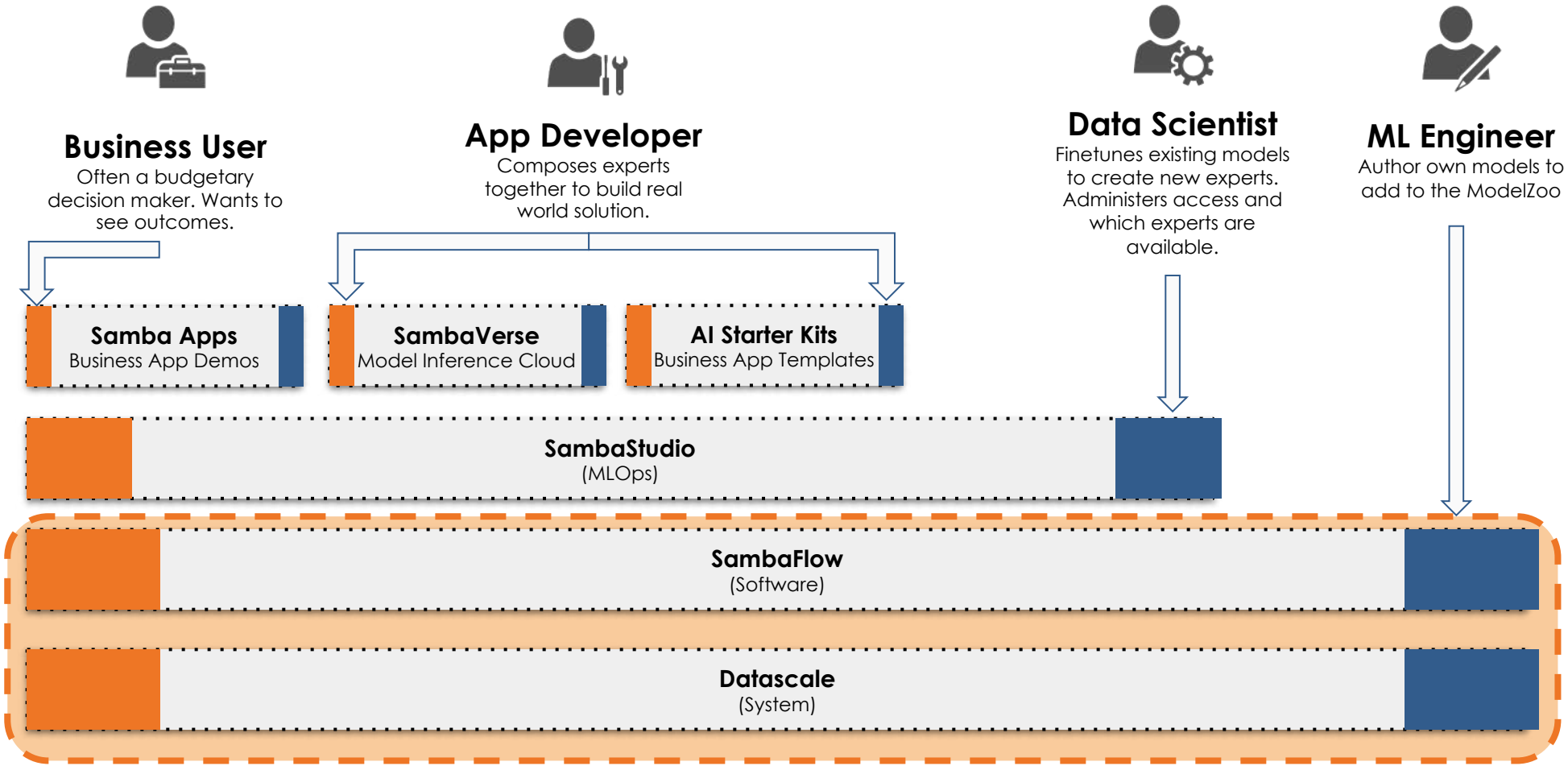
SambaNova SYSTEMS

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

SambaNova SYSTEMS

SambaNova Software Stack

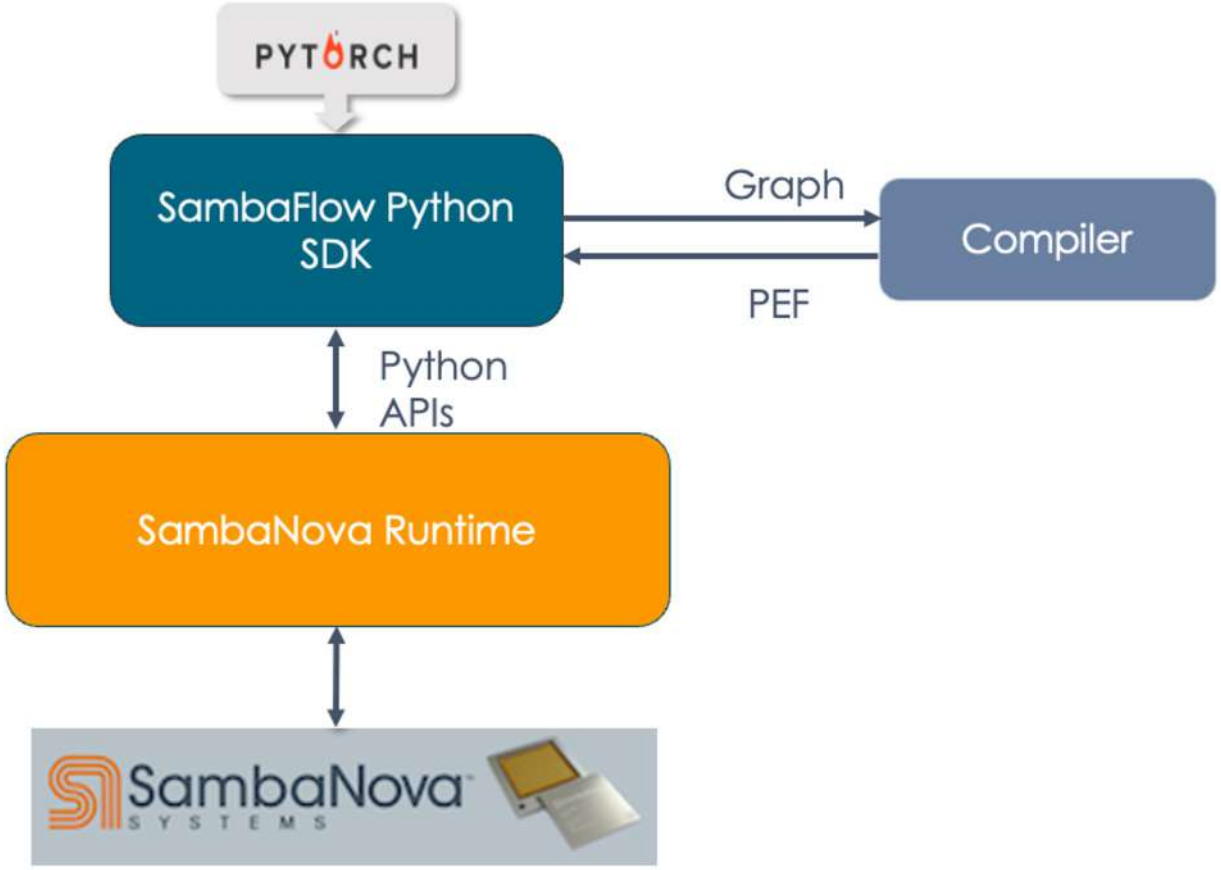


Free,hosted multi-tenant access

Paid, dedicated, single tenant access

SambaFlow

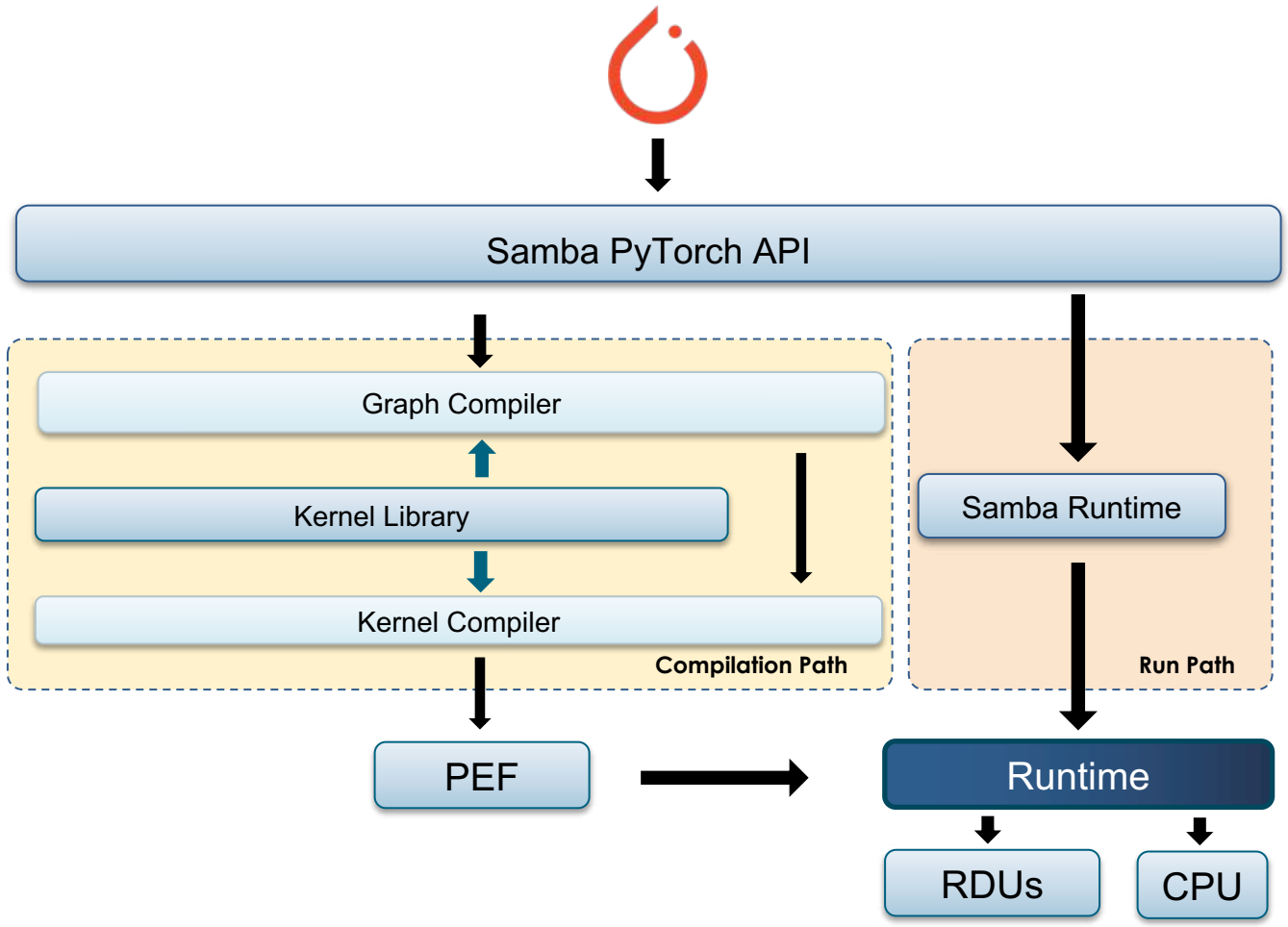
- Supports standard ML frameworks such as Pytorch
- Automatically extracts, optimizes and maps dataflow graphs onto RDUs
  - Achieve high performance without the need for low-level kernel tuning
- A consistent programming model for scaling from 1-RDU to multi system configurations
- Key components:
  - A Python interface to compile & run models
  - Compiler, intakes a Pytorch graph and outputs a PEF
  - Runtime, custom OS for RDUs



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



Compiler Modes

O0 Operator Mode

- Initial bring up and model testing
- Each operator is run as a separate function
- Some optimizations applied

O1 Module Mode

- Fuse operators into modules for optimization
- Fusion rules defined in YAML files, heuristics automatically applied
- Reusability

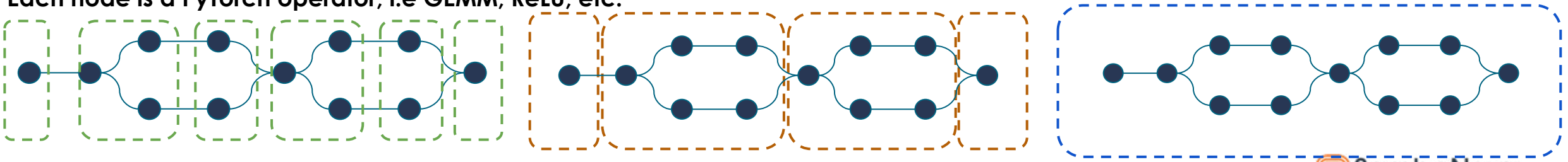
O1 HD (Human Decisions)

- User directed heuristic optimization

O3 Full Graph Mode

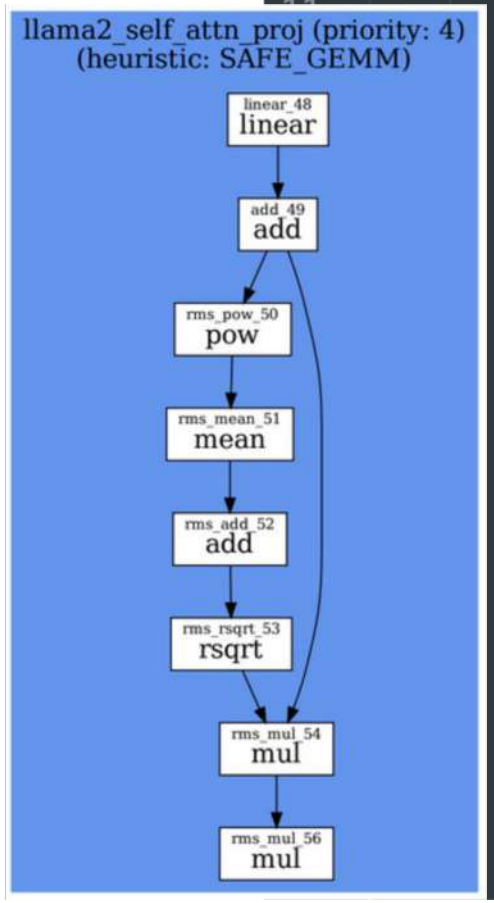
- Fuse and optimize across entire graph
- Configuration specific
- HD files provide expert tuning
- Limited reusability

Each node is a PyTorch operator, i.e GEMM, ReLU, etc.



O1 Operator Fusions

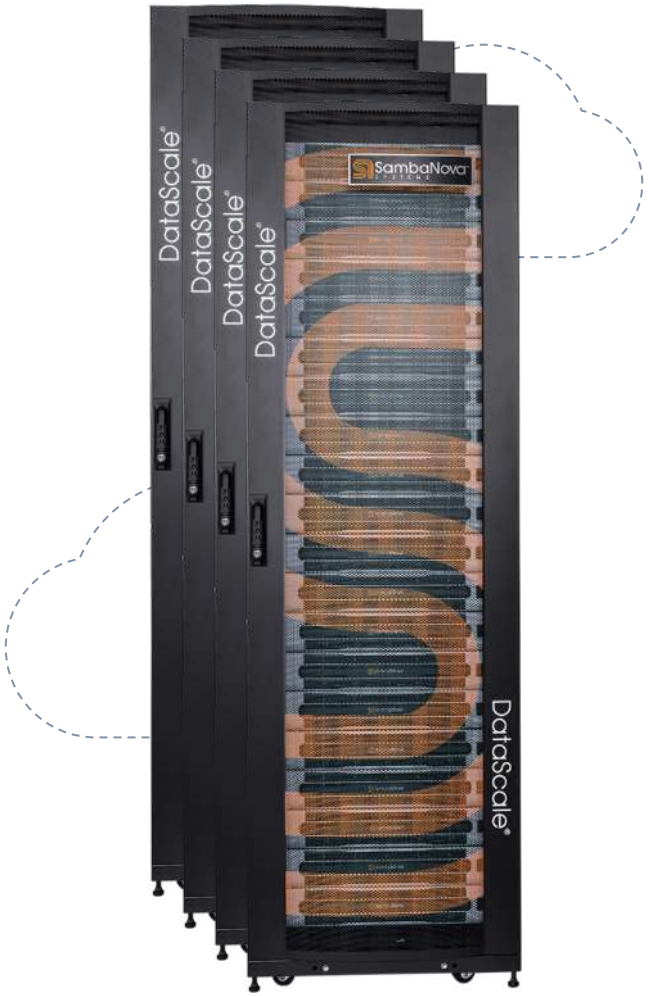
- Patterns of operators to fuse into a dataflow
  - + Users can also define their own patterns in yaml, or define directly in the app
- Each pattern can also specify a "heuristic"
  - + A heuristic is a specific strategy for optimization, put together as a package deal
    - e.g. sharding, tiling, & section cuts
  - + Heuristics are flexible, being applicable to any pattern that meets its requirements




```
llama2_self_attn_proj:
    priority: 4
    heuristic: SAFE_GEMM
    pattern:
        linear_48:
            op_type: linear
            child: add_49
            set m_shard_degree: 4
            set k_shard_degree: 2
        add_49:
            op_type: add
            children:
              - rms_pow_50
              - rms_mul_54
        rms_pow_50:
            op_type: pow
            child: rms_mean_51
        rms_mean_51:
            op_type: mean
            child: rms_add_52
        rms_add_52:
            op_type: add
            child: rms_rsqrt_53
        rms_rsqrt_53:
            op_type: rsqrt
            child: rms_mul_54
        rms_mul_54:
            op_type: mul
            child: rms_mul_56
        rms_mul_56:
            op_type: mul
```

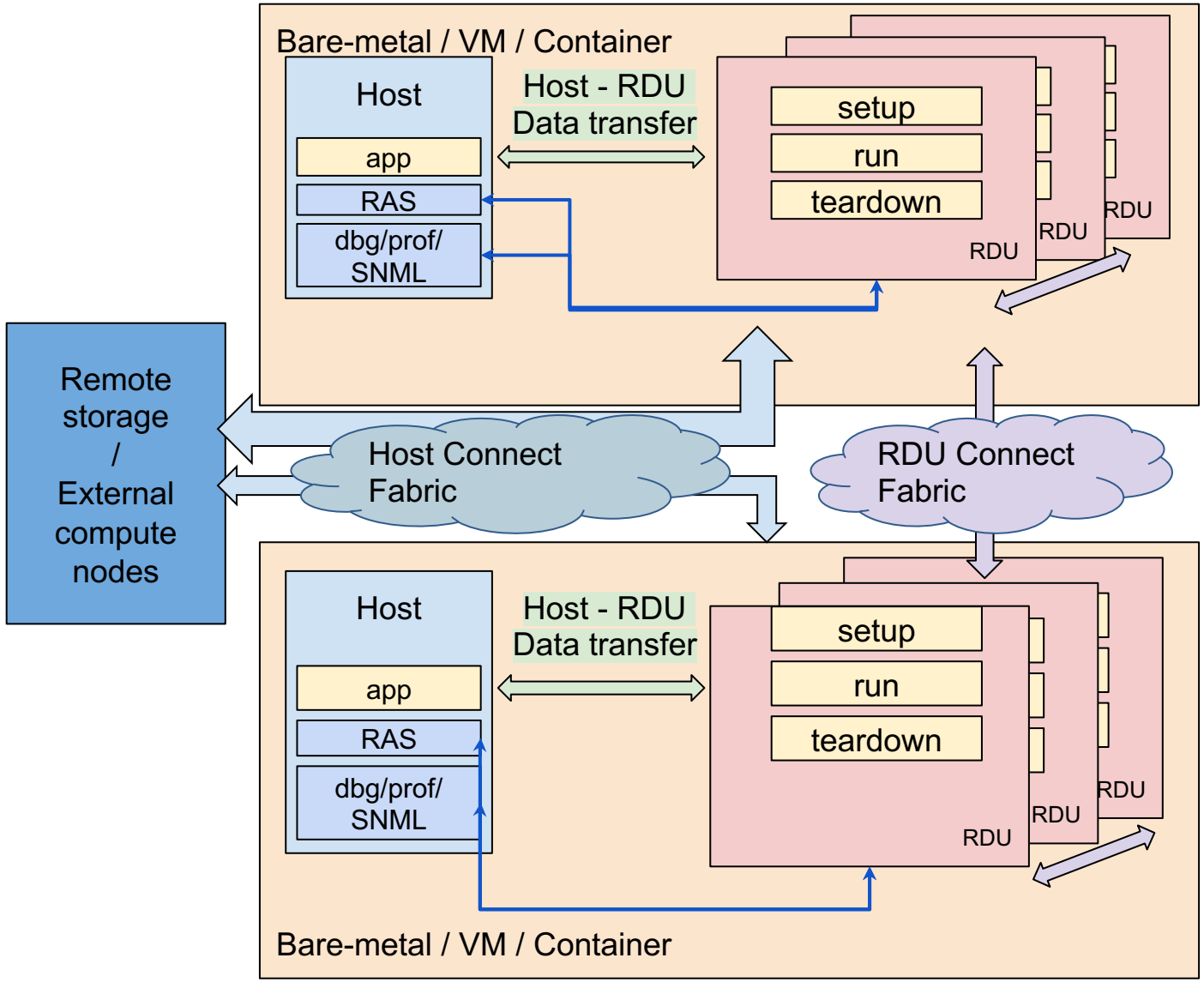
SambaFlow Runtime

- Scalable high-performance runtime stack for SambaNova dataflow distributed systems.
- Operates as an **operating system** for RDUs
  - Manages AI compute, memory, I/O including PCIe and networking
  - Manages application/graph setup, scheduling, execution and tear-down
- Multi-OS support : Ubuntu 20.04.3 LTS, RedHat 8.5
- Minor-version backward compatibility for all Runtime interfaces



Core features of Runtime

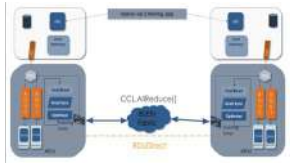
- Model parallel within a node
- Data-Parallel within and across nodes over RDUConnect (Inter-RDU) networking fabric
- Reliability, Availability, Serviceability (RAS)
- Support for external compute nodes and remote storage via host network fabric
- Debugger, performance & system management tool chain
- Language agnostic system management layer (SNML) interface for customers



Enterprise Features of DataScale

Distributed Data Parallel Training

- Distributed Training through Data Parallel
  - Across RDUs, nodes, racks
  - Support >1k RDUs over RDMA
- Algorithm-Topology Library
  - Multi bi-directional ring, all-to-all, Hierarchical allreduce
- Optimized Dataplane
  - High bandwidth over multiple IO fabrics
- Support primitives
  - allreduce, allgather, send, recv, fp32/bf16 mixed



System Reliability, Availability and Serviceability

- Hardware fault/error management
  - DB-based HW fault/error management, provides records of error events, faults
  - Provides a tool interface /opt/sambaflow/bin/snfadm

```
/NODE/XRDU_0/RDU_0/PCIE_8                | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/PCIE_9                | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/PCIE_10               | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/PCIE_11               | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/TILE_0                | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/TILE_1                | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/TILE_2                | N/A                | Present          | Online
/NODE/XRDU_0/RDU_0/TILE_3                | N/A                | Present          | Online
/NODE/XRDU_0/RDU_1                       | 407030B460D05B55   | Present          | Online
/NODE/XRDU_0/RDU_1/DDRCH_0/DIMM_G0       | 22B0D4A            | Present          | Online
/NODE/XRDU_0/RDU_1/DDRCH_0/DIMM_G1       | 22B0E88            | Present          | Online
/NODE/XRDU_0/RDU_1/DDRCH_1/DIMM_H0       | 22B0D45            | Present          | Online
/NODE/XRDU_0/RDU_1/DDRCH_1/DIMM_H1       | 22B0D3A            | Present          | Online
```

Application Diagnostics and Debugging

- Debugging:
  - Slurm_feeder for pef contents
  - Stdout
  - Syslog-based logging
- Observability:
  - Raise exceptions programmatically
  - Syslog-based logging
- Diagnostics:
  - Compute, memory, IO statistics
- SambaTune
  - Gain insight into performance

SambaNova SYSTEMS

Dataflow for Large Deployments

Samba-1 COE SN40L

SambaVerse
SambaApps

SambaNova SYSTEMS

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/

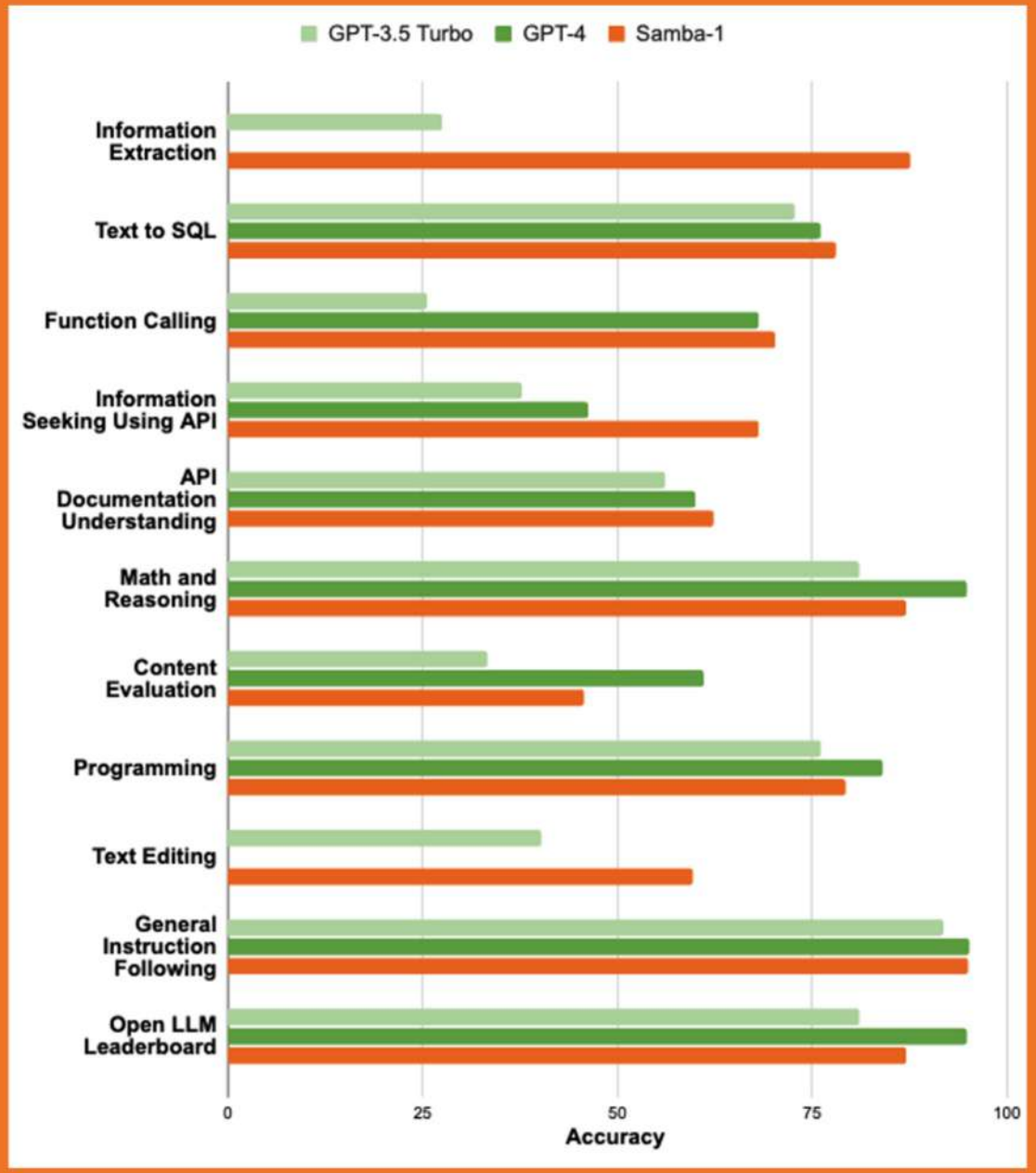
SambaNova SYSTEMS

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 adapts to private enterprise data, enhancing performance beyond generic models.

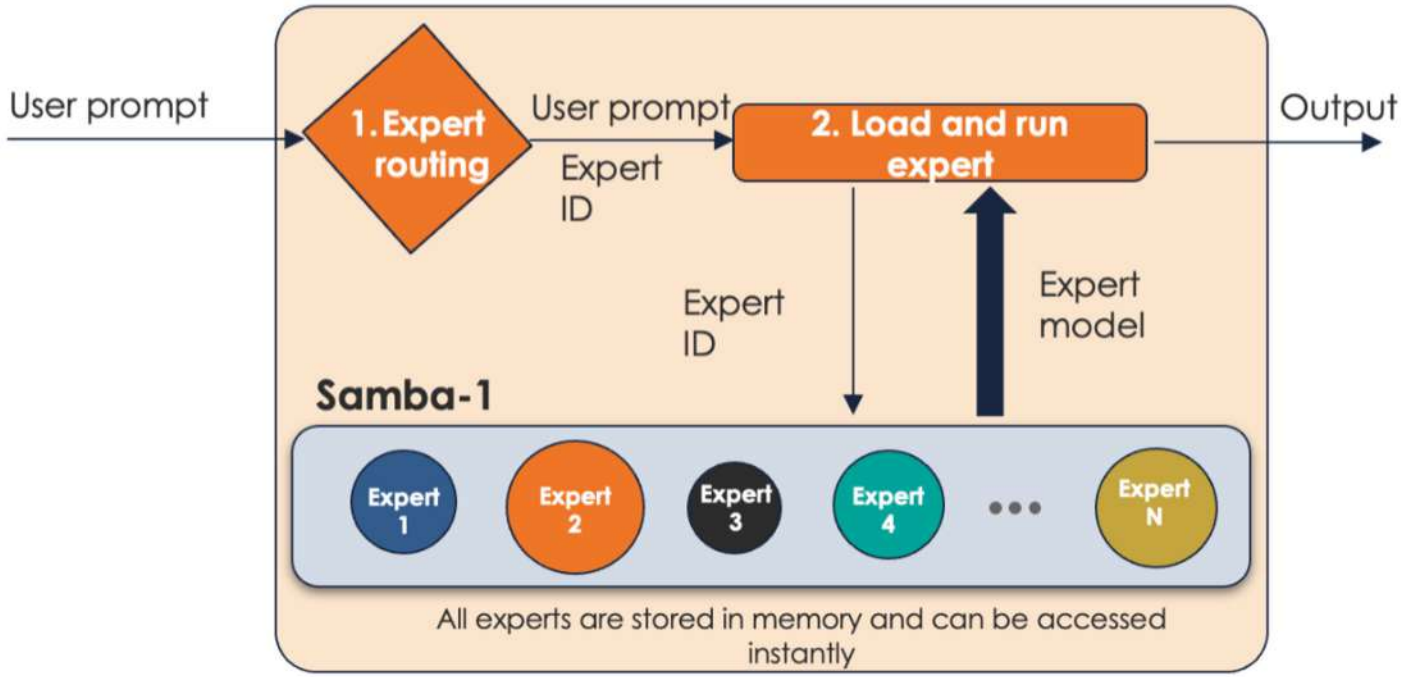
Try It: https://fast.snova.ai/
Details:
https://sambanova.ai/blog/benchmarking-samba-1



SambaNova SYSTEMS

Routing in CoE

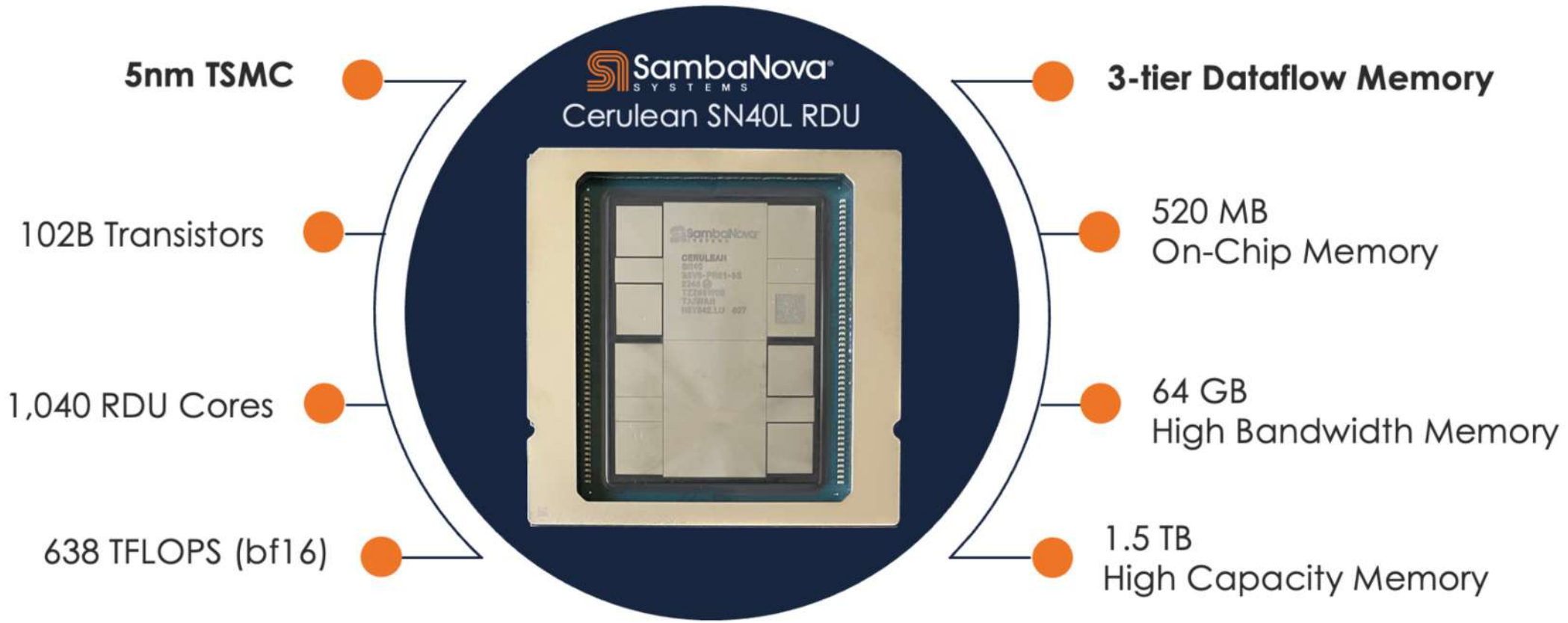
A CoE router predicts the best expert(s) for the most accurate response to a prompt



SambaNova SYSTEMS

SN40L: SambaNova's new CoE-optimized RDU

"Cerulean" Architecture-based Reconfigurable Dataflow Unit



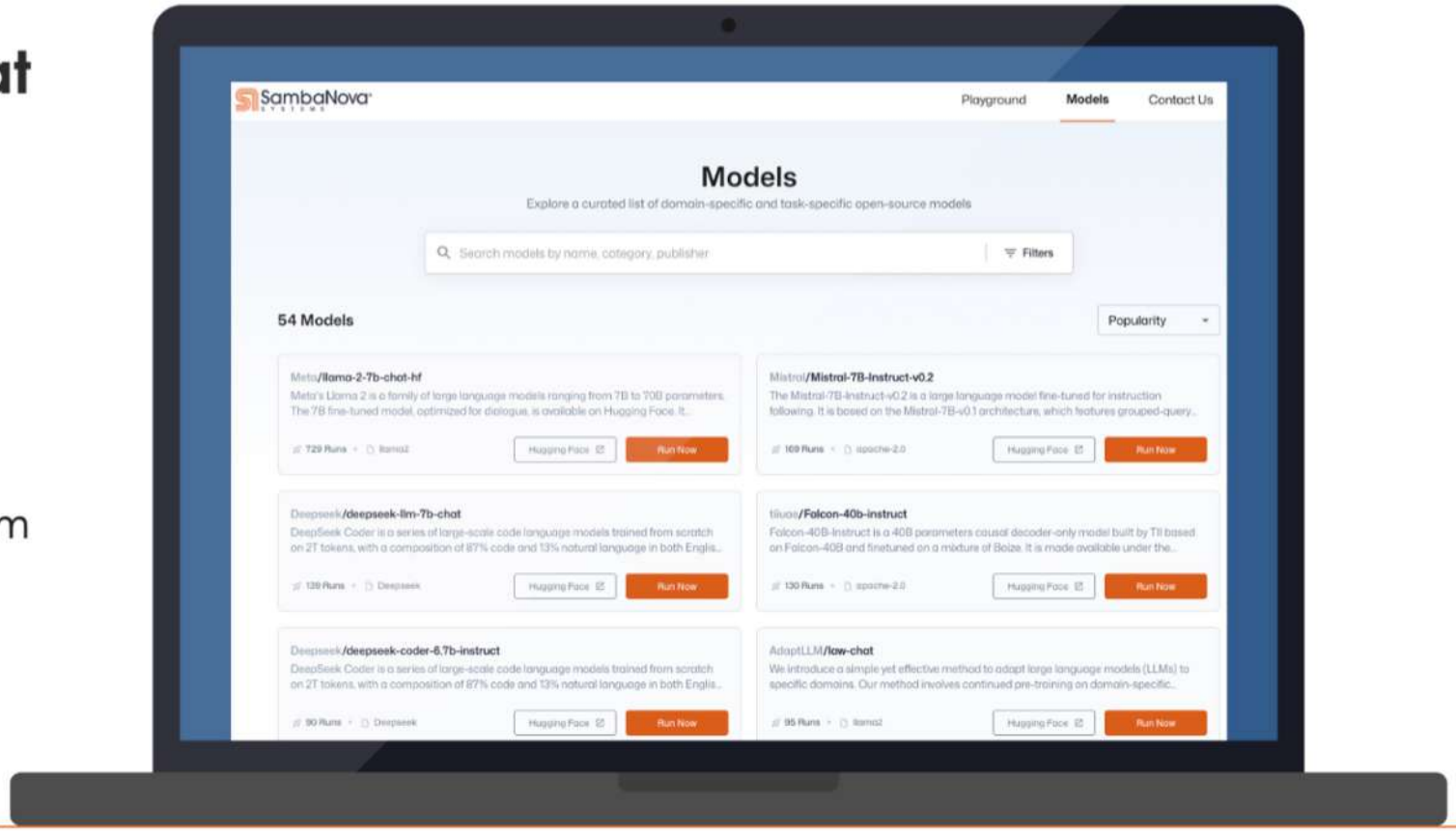
Generative AI Training and Inference

SambaNova SYSTEMS

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



SambaNova SYSTEMS

Introducing Samba Apps

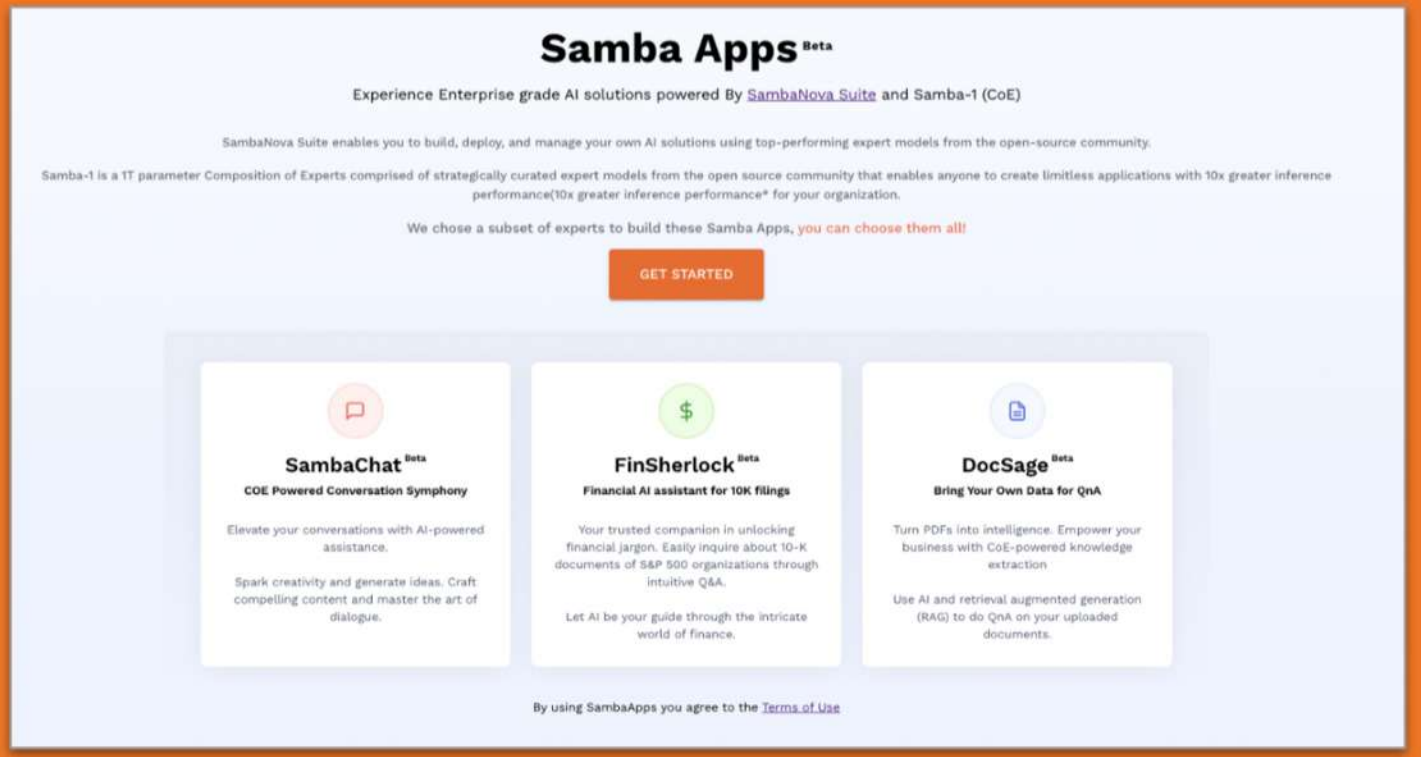
Experience AI-enabled apps powered by the SambaNova Suite **for free!**

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!



Launched in Beta, continuously improved with user feedback

SambaNova SYSTEMS

More Details

- Get more details on Sambanova Public Docs
  - SambaFlow developer documentation
- Contact Sambanova Support team
  - help@sambanova.ai
- Go to the Support Portal
  - support.sambanova.ai


SambaNova®
SYSTEMS