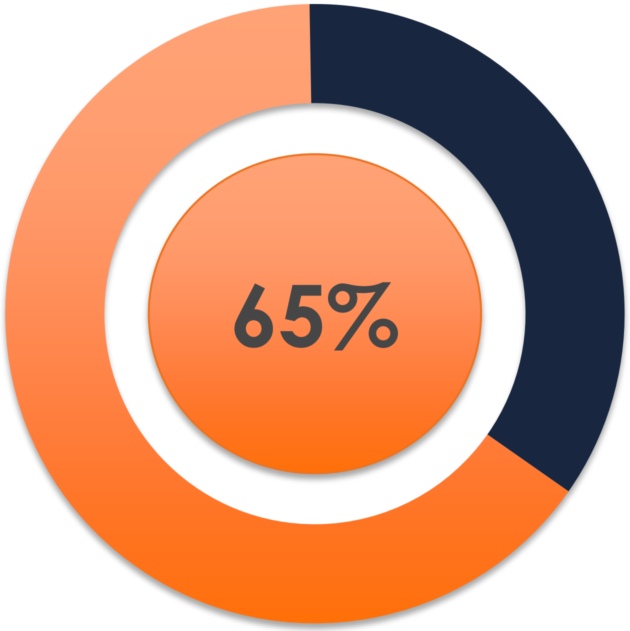
Beyond GPUs: Powering the Next Wave of AI

Anton McGonnell
VP of Product

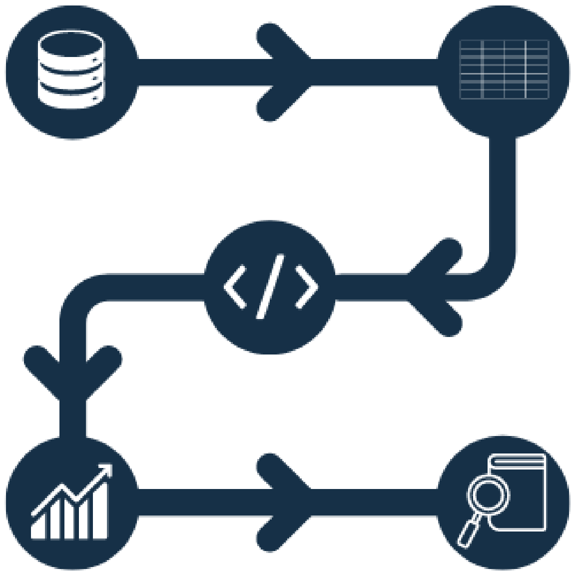
Sept 10, 2024

v 1.0

The Need for *Speed*



Speed and Latency are important criteria for Gen AI Developers – Artificial Analysis



Building Agents Requires Many Models and Faster Real-Time Inference

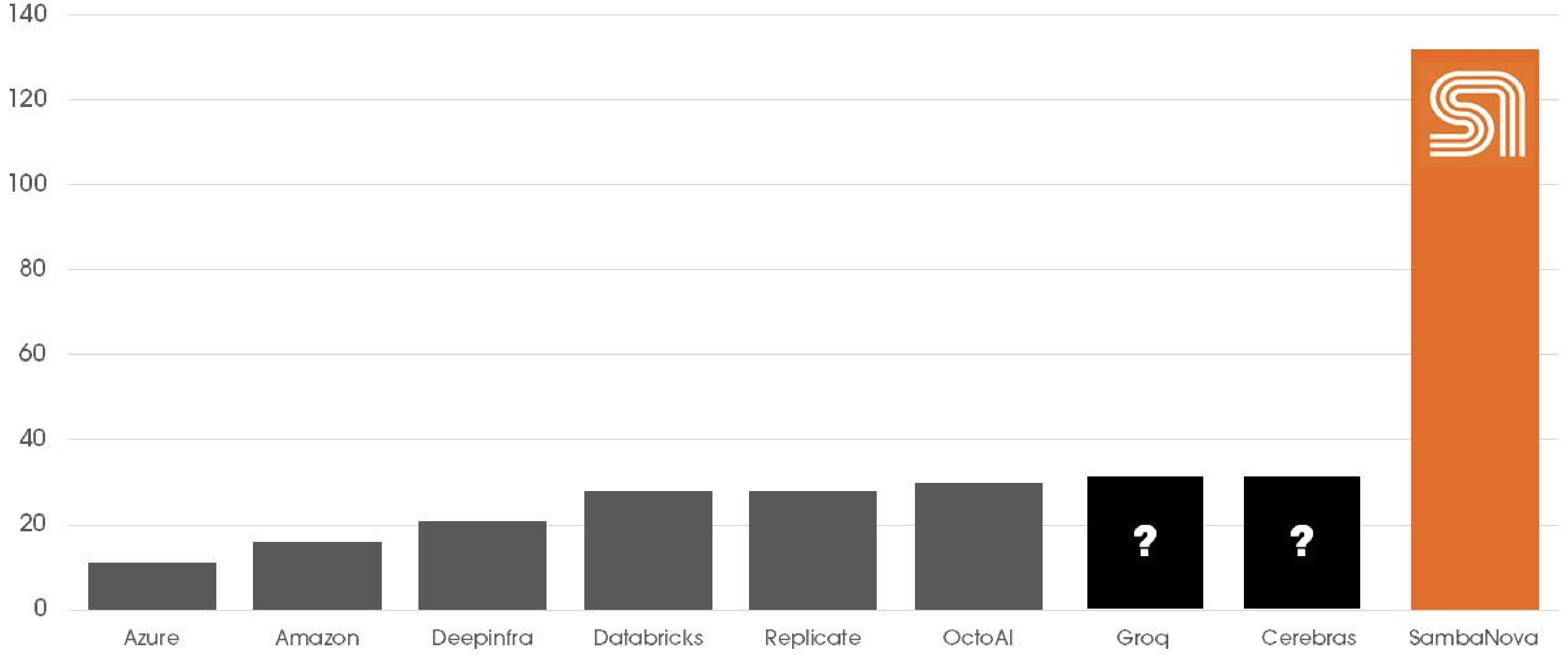
Fast Tokens

The faster, the better

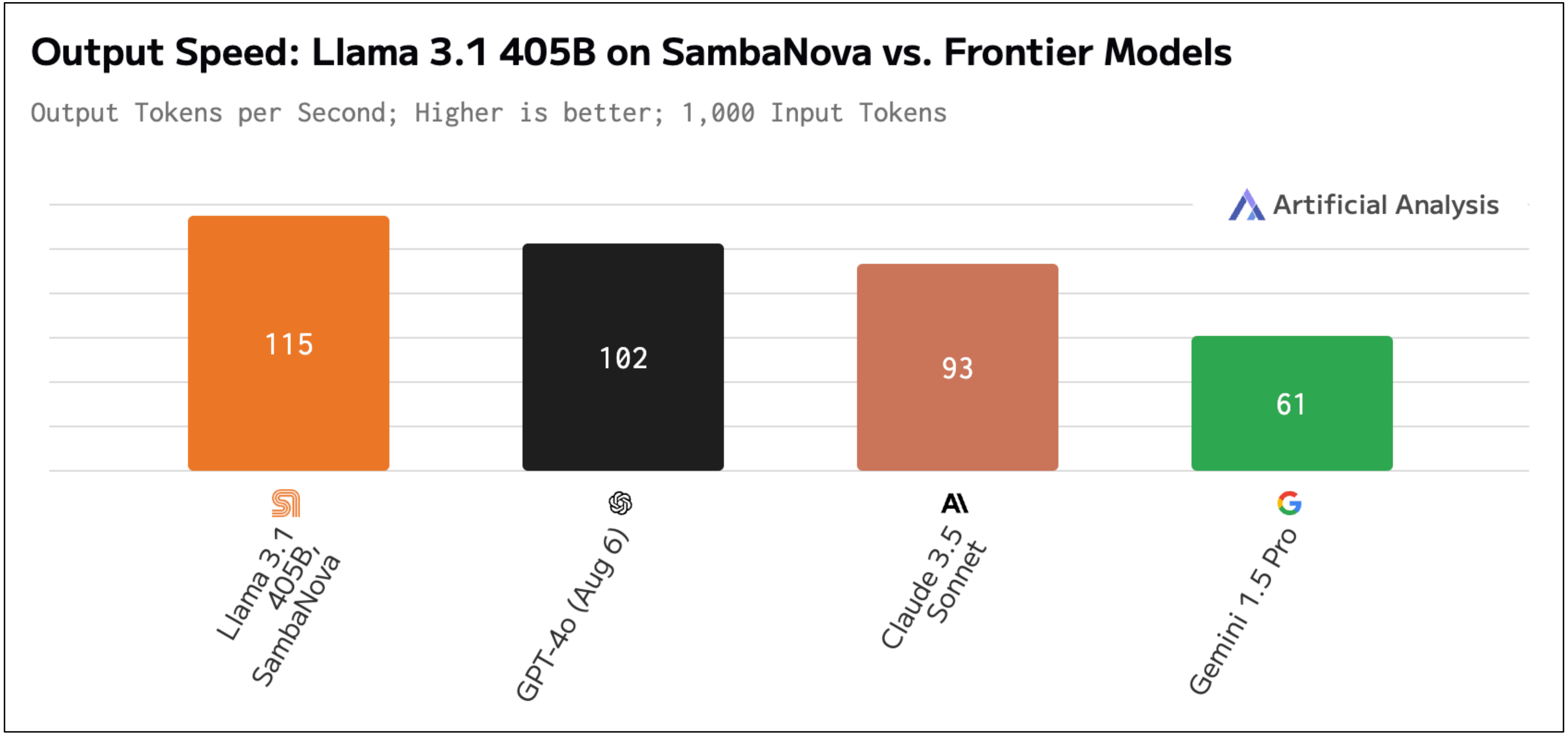
SambaNova Cloud

The Fastest AI Inference on the Best Model

The Fastest Inference on Llama 3.1 405B



405B is the Best Open-Source Model



Faster On All Scales

10X Faster Than GPUs

	SambaNova RDUs	Nvidia GPUs
Llama 3.1 8B 16-bit	**1066**	93
Llama 3.1 70B 16-bit	**570**	32
Llama 3.1 405B 16-bit	**132**	14

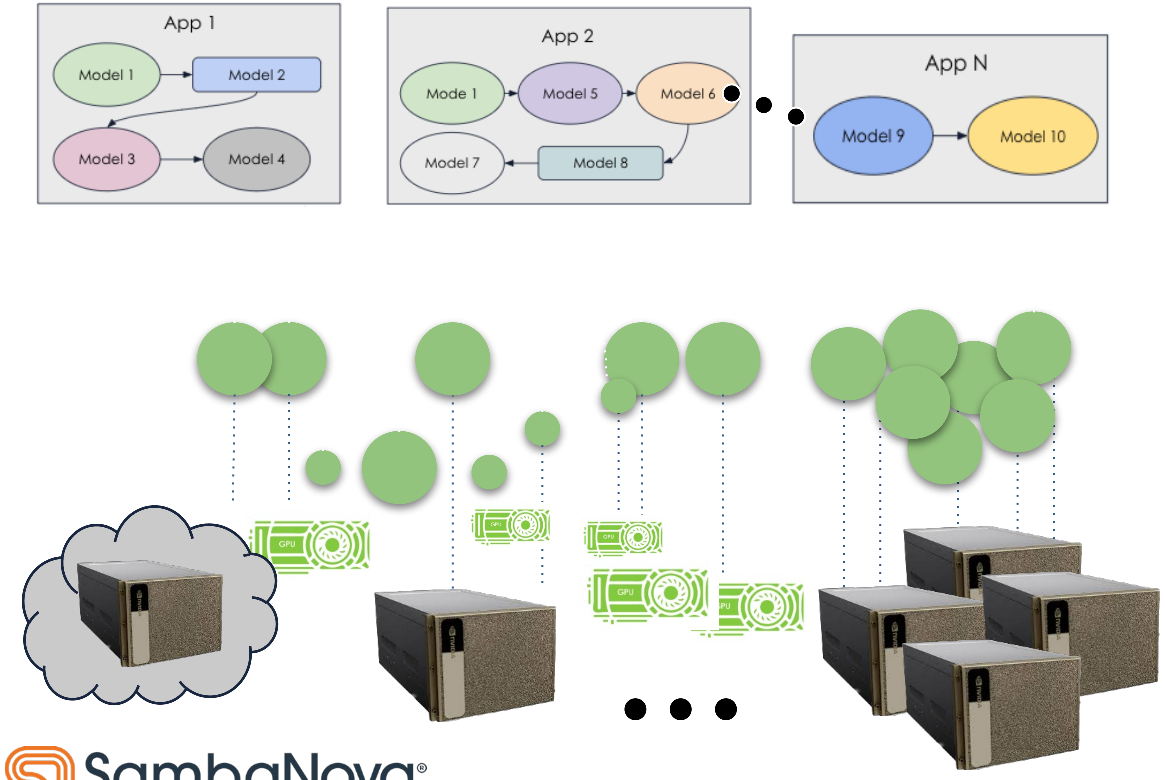
Tokens/Second/User

No Number of GPUs Can Achieve RDU Performance

A Fundamental Shift of Models Deployment at Scale

Traditional GPU Systems

Individual model endpoints

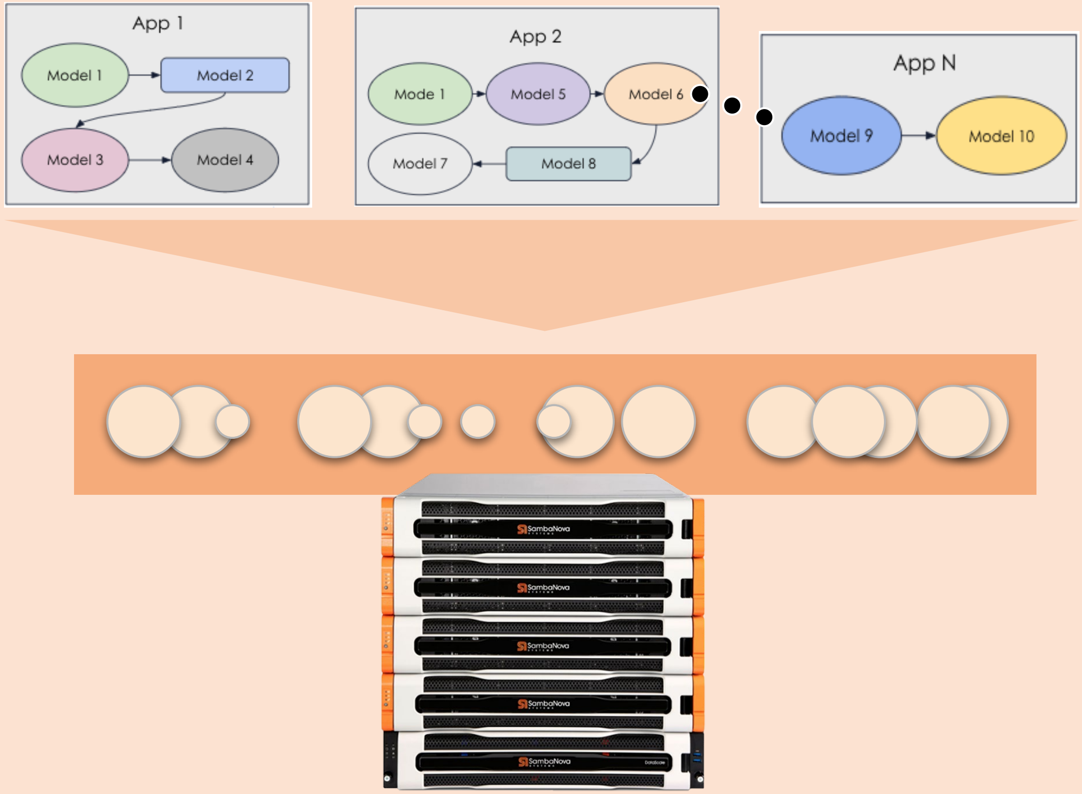


SambaNova SYSTEMS

SambaNova SYSTEMS

All models in memory

(Super low latency model switching)

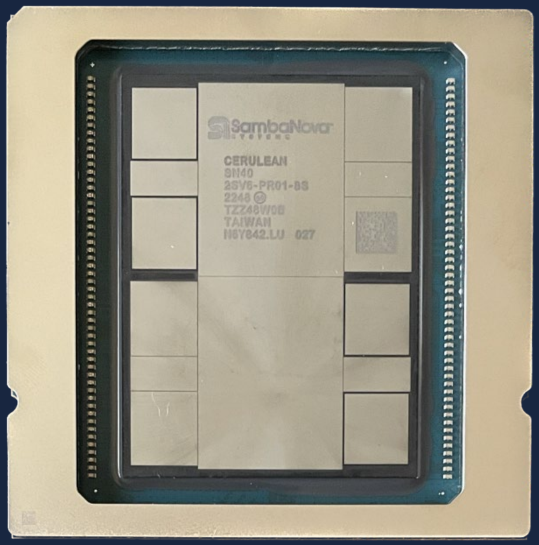


SN40L: The Best Chip Designed for AI

“Cerulean” Architecture - based Reconfigurable Dataflow Unit

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
- 1.5 TB High Capacity Memory

SN40L: SambaNova's LLM optimized RDU

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

On-Chip SRAM
[8 GB, PBs per sec]

Dataflow enabled by large On-Chip Memory

High throughput inference with caching

25.6 TB/s

RDU High Bandwidth Memory [1 TB]

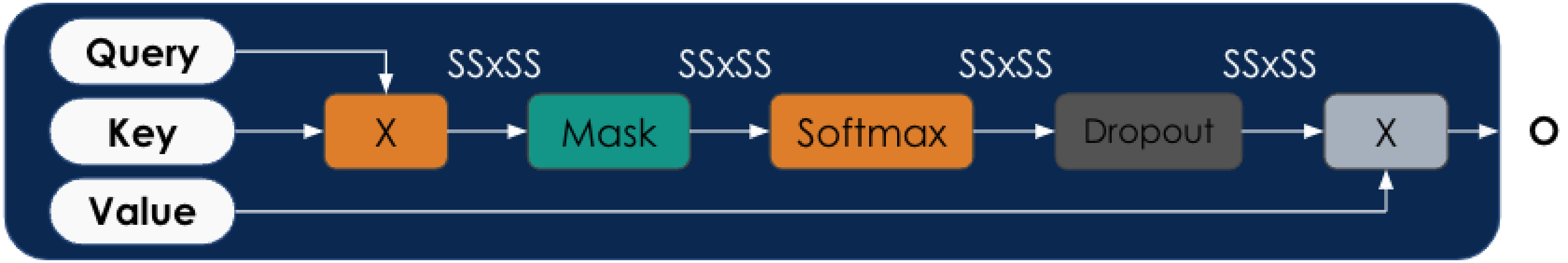
Low Latency Model Switching
(E.g., < 0.01s for llama3.1 8B)

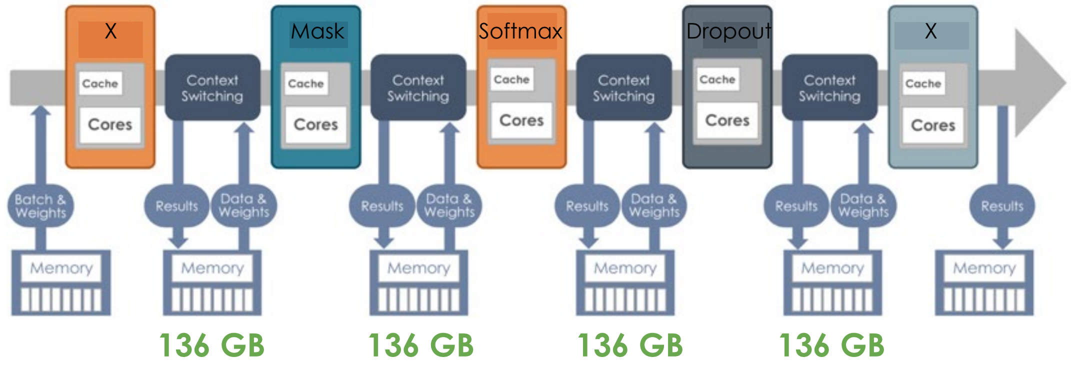
1600 GB/s

RDU High Capacity DDR Memory [24 TB]

The Spatial Dataflow Advantages on RDUs

Automatic kernel fusion eliminates unnecessary I/O read and write

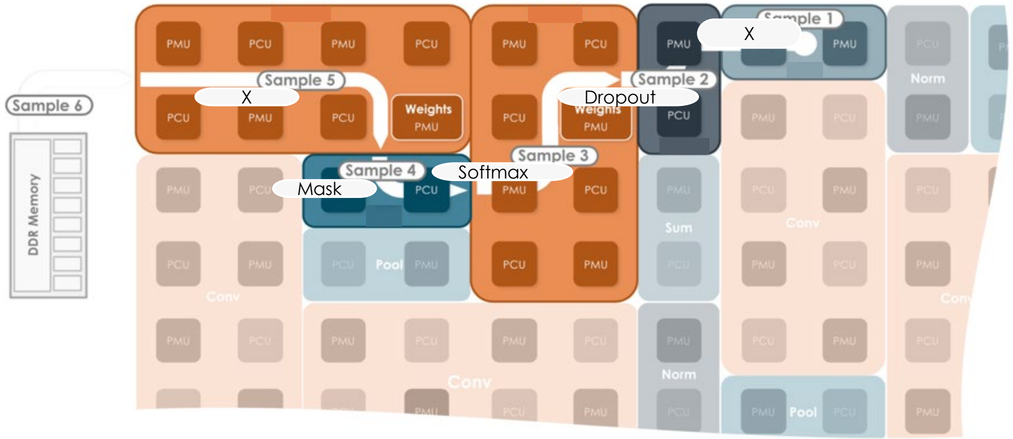




GPU: kernel - by - kernel execution

Bottlenecked by memory bandwidth

(544 GB size of intermediate representations have to be written out and read back in)



RDU: Spatial Execution

Eliminates memory traffic and overhead

(**Zero** Intermediate representation needs to be written out)

Example

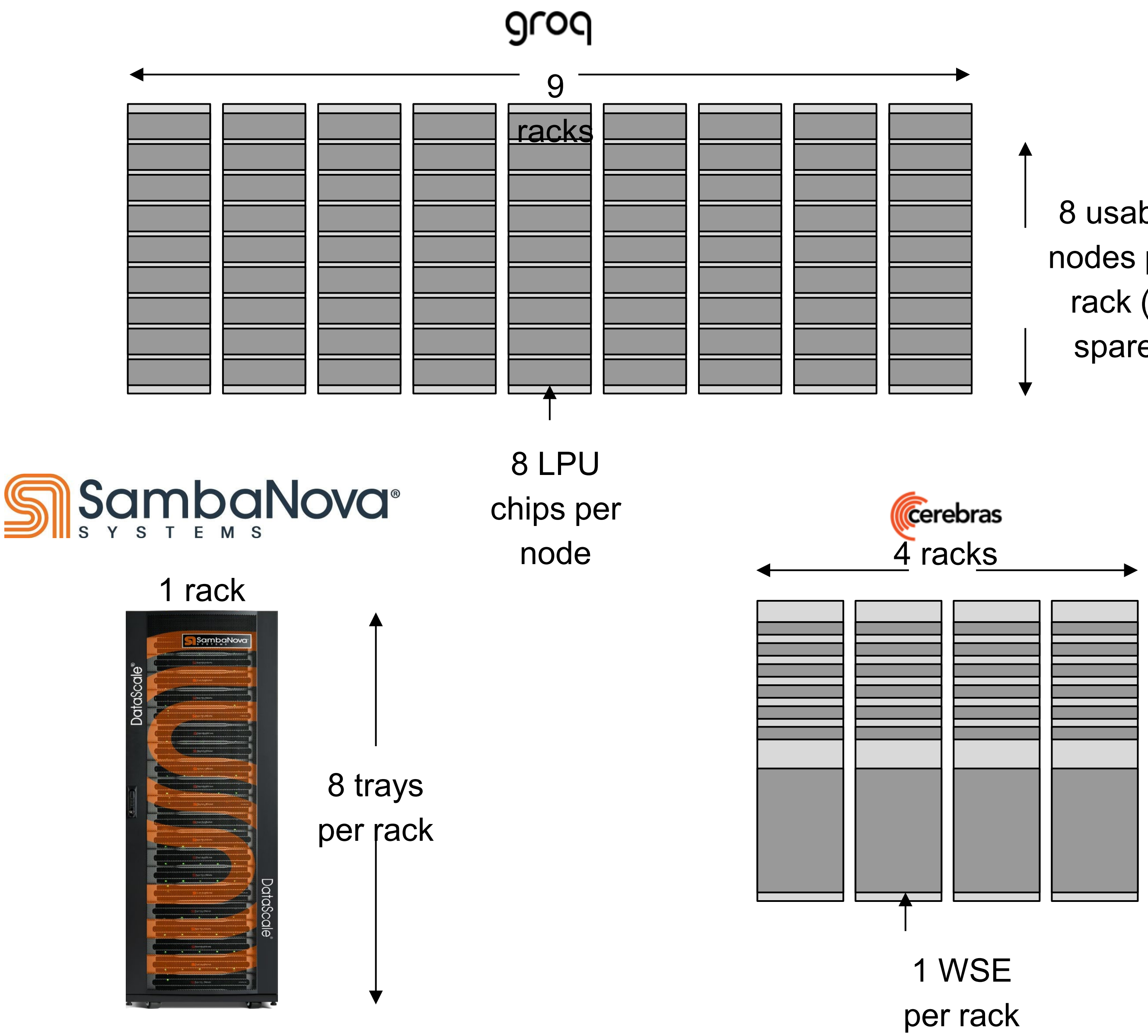
- 13B GPT at **32k sequence length** (SS)
- Number of layers = 40, numbers of heads = 40, embedding dimensions=5120
- Size of input/output tensors = BS x 5120 x 32768
- Between two matrix-multiplications size expanded to: BS x 40 x 32768 x 32768
- For **BS=2**, tensor size between two matmuls is **136 GB** (for 16-bit representation)

SambaNova SYSTEMS

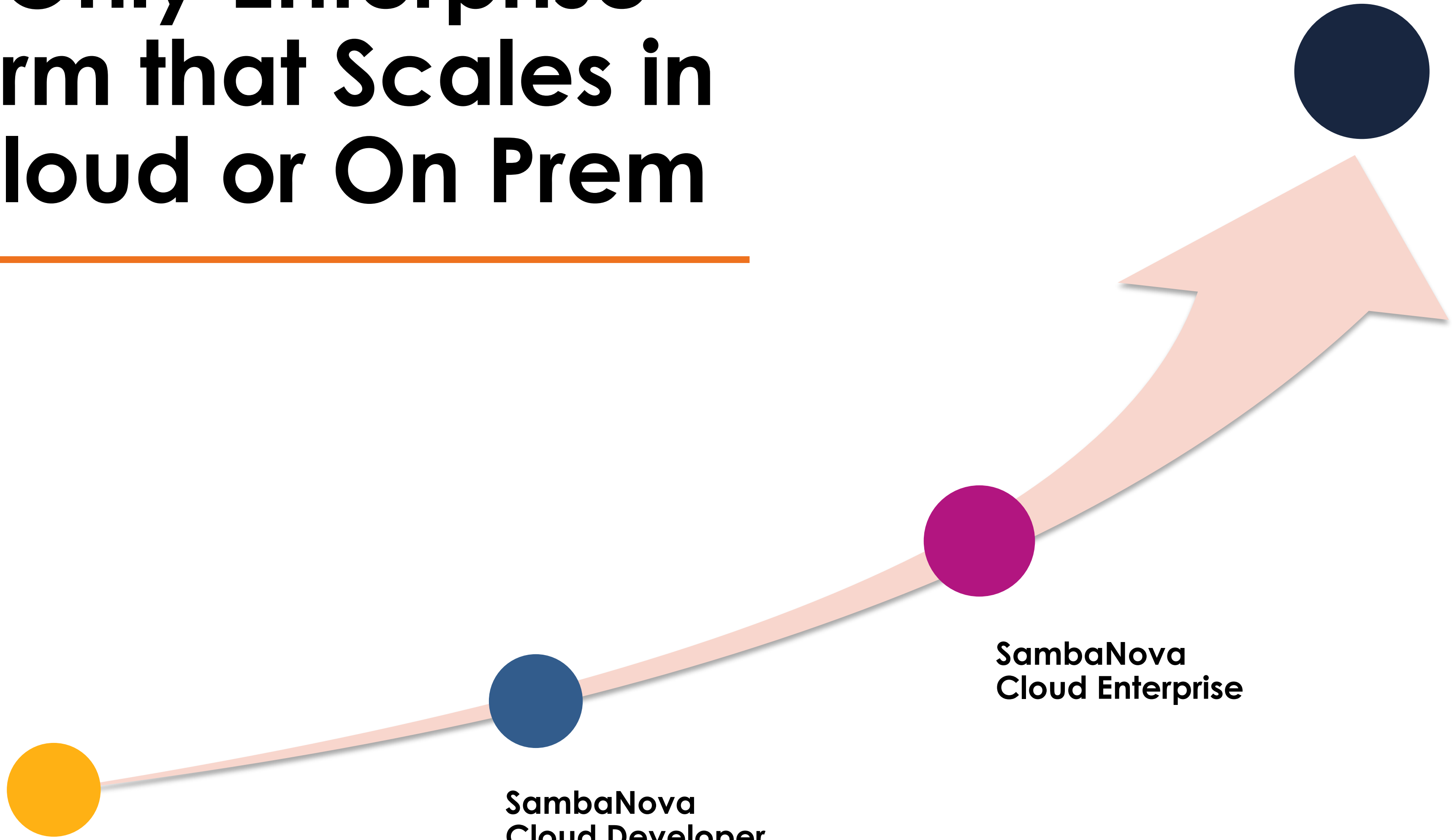
12

Hardware Configuration to run 70B

Unrivaled Chip Efficiency



The Only Enterprise Platform that Scales in the Cloud or On Prem



Start Building Today

cloud.sambanova.ai

SambaNova® SYSTEMS

Workshop: Get Started Developing on the Fastest AI Platform

Dive Deeper into SambaNova with our 1-hour workshop September 11th

- Learn about how speed enables Agentic AI applications
- Get setup on the SambaNova Cloud
- Build your first Hello World application
- Quicky setup Enterprise RAG search with our AI Starter Kits

Vasanth Mohan
Director Technical Product Marketing & Developer Relations

Varun Krishna
Sr Principal AI Engineer

QUESTIONS?