Generative AI using Foundation Models

Step function improvements over legacy AI technologies
Foundation Models deliver accuracy and flexibility

DATA
- Emails
- Voice/Transcripts
- Contracts
- Reports
- Images
- Social Media
- Structured Data
- 3D Data

TRAINING

ADAPTATION

TASKS
- Translation
- Question Answering
- Sentiment Analysis
- Information Extraction
- Object Recognition
- Summarization
- Classification
- Generation
- Semantic Search

Foundation Models:

- **Huge scale**: Huge model training with self-supervised learning
- **In-context learning**: Powerful emergent capabilities with zero and few shot prompts
- **Highly versatile**: Reduce 1000s of legacy models to a single foundation model
Foundation Model Challenges

2018
Elmo (93M)

2019
BERT (350M)

2020
GPT-2 (1.5B)
T5 (11B)

2021 - 2023
GPT-3 (175B)
Megatron-Turing NLG (530B)
PaLM (540B)

Challenges:
1. Model size
2. Model complexity
3. Training time

Foundation Model Era
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
The SambaNova Approach to AI
SambaNova is innovating at every layer of the AI stack

- **SILICON**
  - Reconfigurable Dataflow Architecture

- **SOFTWARE**
  - Dataflow Compilers

- **SYSTEMS**
  - Integrated systems, optimized out-of-the-box

- **as-a-SERVICE**
  - Pre-trained state-of-the-art models
SambaNova Cardinal SN30 RDU
The world’s most advanced Dataflow processor

- 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™
Reconfigurable Dataflow Unit™
SambaFlow Software Stack

- Ease of use
- Out-of-the-box performance and efficiency
- Breakthrough accuracy, simplified
- Native Kubernetes Support
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

Copyright © 2023 SambaNova Systems
SambaNova Dataflow-as-a-Service Foundation Model Platform

Unlimited usage, infinite scale, 22x faster time-to-value

- Pre-trained Foundation Models
  - GPT 175B, 13B, 1.5B
- Deployed in Customer Environment
  - On-premises or in the cloud
- Unlimited fine-tuning and inference
- Subscription pricing
- Fully managed by SambaNova
Vertically-integrated, full-stack state-of-the-art AI deployable instantly

No/low code access to latest models on scalable infrastructure

Dataflow-as-a-Service Platform

- BFSI Solution
- PubSec/Govt solution
- Healthcare solution
- Oil and Gas Solution
- Manufacturing Solution

Vertical Solutions use common Software platform components

Platform includes enterprise AI functions like Mlops, manageability, Observability, BI, labelling mgmt.

Base Models, compiler and compute functions

Hardware

GPT3, Conv3d, Rescale, ...

On-prem, Hosted, Cloud
The SambaNova Foundation Model Platform
Innovation at every layer of the stack

Dataflow-as-a-Service™

as-a-SERVICE
Pre-trained Foundation Models

SYSTEMS
DataScale

SOFTWARE
SambaFlow™

SILICON
RDU

DataScale®
The SambaNova S-Curve for Enterprise Generative AI

A complete, vertically integrated stack

Full-Stack Generative AI Solution Tailored To Operate Seamlessly Where Customer Data Resides
Thank You!