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.
Accelerating Software 2.0

Kunle Olukotun
Co-Founder, Chief Technologist
SambaNova Systems
Three Computing Trends

- Multi-core processing utility is at end of life
- Convergence of training and inference
- General applicability of next-gen compute beyond ML
Software 1.0 vs Software 2.0

- Written in code (C++, ...)
- Requires domain expertise
  - Decompose the problem
  - Design algorithms
  - Compose into a system
- Programmer input: training data
- Written in the weights of a neural network model by optimization
- Reduced lines of code

Andrej Karpathy, Scaled ML 2018 talk

©2019 SambaNova Systems
Software 2.0 is Dataflow

1000x Productivity
Google shrinks language translation code from 500k imperative LoC to 500 lines of dataflow (TensorFlow)
Dataflow Graphs

Dataflow Graph Analyzer

Hierarchical Parallel Patterns
Next gen Software 2.0 systems need support for

- **Hierarchical parallel pattern Dataflow**
  Natural ML execution model

- **Terabyte sized models**
  Higher accuracy

- **Sparsity**
  Graph based neural networks

- **Flexible mapping**
  Model and data parallelism

- **Data processing**
  SQL in inner loop of ML training
Reconfigurable Dataflow Architecture (RDA)
SambaNova Systems Cardinal SN10 RDU

- First Reconfigurable Dataflow Unit (RDU)
- TSMC 7nm
- 40B transistors
- 50 Km of wire
- 100s of TFLOPS
- 100s MB on chip
- Direct interfaces to TBs off chip
Reconfigurable Dataflow for Unprecedented Flexibility

Performance balances computation & communication

Bottleneck: Yesterday’s platforms only program compute

Flexibility unlocks:
- 10x performance
- 0-to-1 applications
Rapid Dataflow Compilation to RDA
World’s First DataScale Systems Family

DataScale SN10-8
Single 8-socket DataScale system

DataScale SN10-8R Full Rack
4 x 8-socket DataScale systems,
SambaFlow Open Software for DataScale Systems

Graph Entry Points
- Write to OSS ML frameworks or user’s graph
- Push-button automation path

API Entry Point
- User programs to DSL
- Mix of manual and automatic

Optimizations
- Model parallel
- Data parallel
- Tiling
- Streaming
- Nested pipelining
- Op parallel

©2020 SambaNova Systems Confidential | Do Not Distribute
Open Standards, Disruptive Technology, Easy to Deploy
Designed to integrate into existing environments for faster time to results

Open standard rack,
Open standard form factor,
Open standard power,
Open standard cooling,
Open standard operations ...

Open Standards Connectivity
Ethernet
Open Source Frameworks
PYTORCH, TensorFlow
Open Source OS
Linux

©2020 SambaNova Systems Confidential | Do Not Distribute
Enabling New Capabilities (0 ⇒ 1)

**Trillion parameter NLP models**
Key to knowledge understanding

**High Resolution Deep Learning**
50k x 50k
Astronomy, medical imaging, X-ray imaging, ...

**Recommendation models with huge 100GB embedding tables**
Recommendation is the backbone of internet services
Trends in NLP
Today’s platforms constrain NLP

Richer contextual information

Larger Models
Richer, Contextual Information

3-wide encoders
A three-layer BERT model in production at Bing. Richer context, same space.

VS

24-slim encoders
Fewer Parameters, Better Quality on Natural Language Inference
QNLI: 3-layer 78.7 vs. Deeper 79

More than 6x faster on Deeper BERT

SambaNova enables Deeper Design Points

©2020 SambaNova Systems Confidential | Do Not Distribute
100 Billion Parameters on a Single DataScale System
Enabling Large Model Architectures With a Single System
Order of magnitude performance improvement, an order of magnitude fewer systems

64 DGX-2
1,024 V100s
32 TB HBM
16 racks
6,200 kW

8 RDU,
12 TB DRAM,
¼ rack
2,000x Less Power

1 DataScale system

“ONE Model” 1Trillion Params in a Single System: Same Programming Model
Convergence of Training and Inference

Current ML pipeline consists of multiple incompatible HW and SW components

One HW/SW Dataflow platform for training and inference

©2020 SambaNova Systems Confidential | Do Not Distribute
Low-Latency, High-Throughput Inference

Microsoft open sources breakthrough optimizations for transformer inference on GPU and CPU

January 21, 2020

EMMA NING
Senior Program Manager, Azure Machine Learning

16x Throughput
<1/2x Latency
Batch size 1 (unlike GPU)

3-Layer BERT-Large

Latency

CPU (BS=1) 9 ms
GPU (BS=4) 8 ms
RDU (BS=1) 3.5 ms
More Capabilities to Come

- HPC workloads (Material Sci, Oil & Gas, etc.)
- Medical Imaging and Surveillance
- Traditional Data Processing

These are all data flow
Three Computing Trends

1. Multi-core processing utility is at end of life
2. Convergence of training and inference
3. General applicability of next-gen compute beyond ML

©2020 SambaNova Systems Confidential | Do Not Distribute
Richer Context, In a Small Amount of Space

Microsoft open sources breakthrough optimizations for transformer inference on GPU and CPU

January 21, 2020

EMMA NING
Senior Program Manager, Azure Machine Learning

A three-layer BERT model in production at Bing.

Richer context, same space.
Extending the Data Science Pipeline
In the data center or at the edge

**Data Collection**
Identify, acquire, aggregate

**Pre-Process**
Data type transformation, data cleanup

**Train Models**
Test, analyze, select models

**Validate & Scale Models**
Promote selected models to production

**Deployment & Insight**
Predict and monitor

- **Data**
  - OLTP & DW
  - Data Lake
  - Parallel FS

- **Data Prep**
- **De Novo Training**
- **Model Repo**
- **Training**
- **Inference**

- **Optimization**
- **Incremental Re-Training**