SambaNova
SN10 RDU: Accelerating Software 2.0 with Dataflow
Raghu Prabhakar, Sumti Jairath
8/24/2021
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.
SambaNova Systems® Cardinal SN10 RDU

- First Reconfigurable Dataflow Unit (RDU)
- TSMC 7nm
  - Taped Out first half of 2019
  - 40B transistors, 50 Km of wire
- 640 Pattern Compute Units
  - >300 BF16 TFLOPs
  - BF16 with FP32 accumulation
  - Also supports FP32, Int32, Int16, Int8 data formats
- 640 Pattern Memory Units
  - >300 MB on-chip memory
  - 150 TB/s on-chip memory bandwidth
  - Memory transformation operations

©2021 SambaNova Systems
SambaNova DataScale® SN10-8R Systems
Scalable performance for training and inference

- DataScale SN10-8R
  - 8x RDU in Quarter Rack

- 12 TB Memory
  - 48 DDR4-2667 Channels

- Host and RDU-RDU Communication
  - 32 PCIe-Gen4 x16 Links

Up to 38x more memory than conventional systems
SambaFlow™ Software

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

ML Optimizations
- Model parallel
- Data parallel

Dataflow Optimizations
- Tiling
- Op parallel
- Streaming
- Nested pipelining
SambaFlow Produces Highly Optimized Dataflow Mappings

PyTorch

Dataflow Graph

Communication Pattern

SambaFlow Spatial Compilation

©2021 SambaNova Systems
Machine Learning Today: Software 2.0

Adapted from Jeff Dean
HotChips 2017

Data size, model complexity

Accuracy

1980s

More Compute

Today

Software 2.0

Software 1.0

©2021 SambaNova Systems
Evolving Nature of Computational Models

Software 1.0
- Programmer input: code (C++, etc.)
- Solution encoded in composed algorithms
- Deterministic computations
- Only has a single correct result

Software 2.0
- Programmer input: training data
- Solution encoded in trained weights
- Probabilistic models
- Results need only be statistically correct
Yesterday’s Goldilocks Zone is Constraining Progress
Dataflow Exploits Data Locality and Parallelism

**SOFTMAX:**
\[
\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}
\]

**LAYERNORM:**
\[
y = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}} \ast \gamma + \beta
\]

- Dataflow captures data locality and parallelism abundant in Software 2.0
- Flexible scheduling in space and time for higher utilization
- Flexible memory system and interconnect to sustain high compute throughput
- Build custom dataflow pipeline
Cardinal SN10: Chip and Architecture Overview

- TILE: Sea of programmable compute and memory components in a programmable interconnect

- 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 SN10: Tile

Virtual Memory Manager
Top-Level Interconnect

Software-Driven Architecture
Tiled architecture with reconfigurable SIMD pipelines, distributed scratchpads, and programmed switches
Cardinal SN10: PCU

- **Pattern Compute Unit**: SN10’s compute engines
- **Reconfigurable SIMD data path** for efficient dense and sparse tensor algebra in FP32, BF16, and integer formats
- **Programmable Counters** to efficiently program loop iterators
- **Tail unit** accelerates common functions like exp, sigmoid
Cardinal SN10: PMU

- **Pattern Memory Unit**: SN10’s on-chip memory system

- **Banked SRAM arrays** to write and read multiple high-bandwidth SIMD data streams concurrently

- **Address ALUs** for high throughput address calculation for arbitrarily complex accesses

- **Data Align** units for high throughput Tensor layout transformations like transpose
Cardinal SN10: Switch and On-chip Interconnect

- **Switch:** SN10’s programmable packet-switched interconnect fabric
- **Independent Data and Control** buses to suit different traffic classes inherent in real applications
- **Programmable Routing** for flexible on-chip bandwidth allocation to concurrent streams
- **Programmable Counters** to implement outer loop iterators, on-chip metric collection
Cardinal SN10: AG & CU

- **Address Generation and Coalescing Units**: SN10’s interface to IO subsystem
- **Address ALUs** for high throughput address calculation for arbitrarily complex accesses
- **Coalescing Units** to enable transparent access to memories across RDUs and host memory
- **Address space management using programmable, variable length segments**
**Programming Model**

**SOFTMAX:**

\[
\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}
\]

**LAYERNORM:**

\[
y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x] + \epsilon}} \ast \gamma + \beta
\]
Example: Softmax

\[
\text{SOFTMAX: } \quad \text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}
\]
Example: Softmax

\[
\text{SOFTMAX: } \quad \text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}
\]
Example: LayerNorm, Pipelined in Space

\[ y = \frac{x - E[x]}{\sqrt{\text{Var}[x]} + \epsilon} \ast \gamma + \beta \]
Example: LayerNorm, Pipelined in Space + Fused

LayerNorm: \[ y = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}} * \gamma + \beta \]
Example: LayerNorm, Hybrid Space + Time Execution

**LAYERNORM:**

\[ y = \frac{x - E[x]}{\sqrt{\text{Var}[x]} + \epsilon} \times \gamma + \beta \]
Spatial Dataflow Within an RDU

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

The Dataflow way: Spatial
Eliminates memory traffic and overhead

CONVOLUTION GRAPH

Weights
Sample → Conv1 → Pool → Conv 2 → Norm → Sum

Weights
Accurate modeling of intra-molecular interactions using SpMM + NN

Representation of atom interactions: Sparse CSR-like format
**Sparse Matrix Multiply on RDU**

**Logical**

\[
\begin{bmatrix}
0 & -1 & 0 \\
0 & 0 & -2 \\
0 & -3 & -4
\end{bmatrix}
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
5 & 6
\end{bmatrix}
= 
\begin{bmatrix}
-3 & -4 \\
-10 & -12 \\
-29 & -36
\end{bmatrix}
\]

CSR-like Representation

\[
col\_indices = [1\ 2\ 1\ 2]
\]

\[
values = [-1\ -2\ -3\ -4]
\]

\[
nzns\_per\_row = [1\ 1\ 2]
\]

**Spatial**

\[
[1\ 2\ 1\ 2]
\]

\[
\begin{bmatrix}
3 & 4 \\
5 & 6
\end{bmatrix}
\]

\[
\begin{bmatrix}
-3 & -4 \\
-10 & -12 \\
-9 & -12 \\
-20 & -24
\end{bmatrix}
\]

Dynamic Accumulation

\[
[1\ 1\ 2]
\]

\[
\begin{bmatrix}
-3 & -4 \\
-10 & -12 \\
-29 & -36
\end{bmatrix}
\]

©2021 SambaNova Systems
Spatial Dataflow Architecture

Hierarchical parallel pattern Dataflow
Natural ML execution model, Communication-efficiency, Ease-of-use

Terabyte sized models
Large embeddings, True-resolution, Flexible batch-size

Sparsity
Graph based neural networks

Flexible mapping
Model and data parallelism

Data processing
SQL in inner loop of ML training
Dataflow Architecture for Terabyte Sized Models

Dataflow Efficiency +

Compute Capability +

Large Memory Capacity

DataScale SN10-8R ¼ Rack System

©2021 SambaNova Systems
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
DataScale Systems Scale-Out
DataScale SN10-8R: Scalable performance for training and inference

©2021 SambaNova Systems
Train Large Language Models, Without Code Changes

1T parameter NLP training with a small footprint and programming ease

**Out-of-Box Models**
- Huggingface Models
- Write yours in Pytorch

**Developer Efficiency**
- Focus on ML problems instead of System Engineering

**High Accuracy Models**
- No Compromise on Model Architecture required to hide System Deficiencies

Large Kernels
- To Hide Communication Costs
- Statistically Nonoptimal

Complex System Engineering
To Enable Model Architecture Exploration

Transformer layer #1
Tensor MP partition #1
Pipeline MP partition #1

Tensor MP partition #2

416 GPUs, 32 TB HBM, 8 racks


©2021 SambaNova Systems
Train with 4k to 50k Convolutions, Without Code Changes

SambaNova trains true-resolution Computer Vision models effortlessly

<table>
<thead>
<tr>
<th>Image Resolution</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low-resolution</strong> (e.g. cats)</td>
<td></td>
</tr>
<tr>
<td>4k images (e.g. Autonomous driving)</td>
<td></td>
</tr>
<tr>
<td>50k x 50k (e.g. oil &amp; gas, medical imaging, anti-viral research, astronomy …)</td>
<td></td>
</tr>
</tbody>
</table>

SambaNova DataScale can run these workloads out of the box
Enabling High-Resolution Full-Image Pathology

Direct analysis with 100X higher resolution
World Record Accuracy High-Res Convolution Training
Out of the box world record accuracy

90.23%
Accuracy

World Record CosmicTagger Training Accuracy

Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture
Murali Erramilli, Venkatram Vishwanath, Corey Adams, Michael E. Papka, and Rick Stevens, Argonne National Laboratory, Lemont, IL 60439, USA
The base algorithm for HPC
Public sector, energy, BFSI, manufacturing, automotive, pharma, biomed

Differential equation approximation
Heat transfer, stress/strain mechanics, fluid dynamics

Molecular energies modeling
Materials science, chemistry, molecular biology, drug design

These Problems Are All Dataflow
Surpassing State-of-the-Art Accuracy and Performance with Dataflow

- > 10x: Transformer (Distilled), EfficientNet
- 2-5x: Transformer (Standard), ResNet-50
- 0 → 1: GPT, Zero, MoE, True-res CV Models

©2021 SambaNova Systems
Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!

Visit the product pages

Learn More About Dataflow!