SambaNova®

SYSTEMS

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 Intro

PC Cluster Consortium

Jan 20, 2022

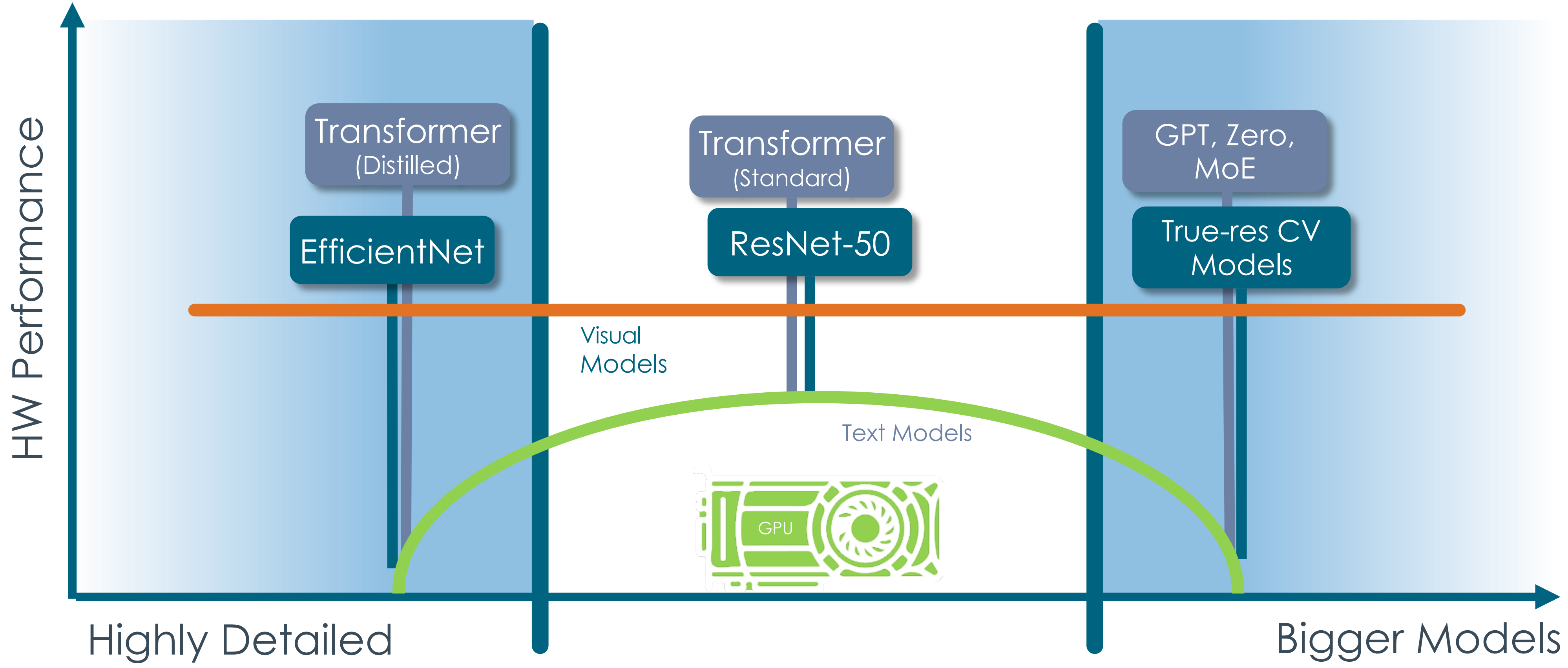
Toshinori Kujiraoka

Country Sales Director

Brief Background

Yesterday's Goldilocks Zone is Constraining Progress

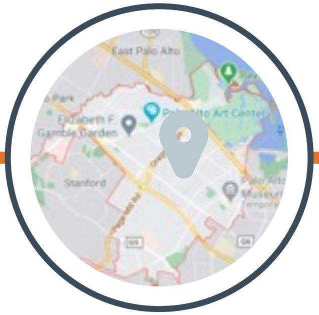
Existing architecture bottlenecks are hard to overcome



About SambaNova Systems

2017
Established
the company

Palo Alto
Austin
London



ML/AI
Software-Defined
Hardware

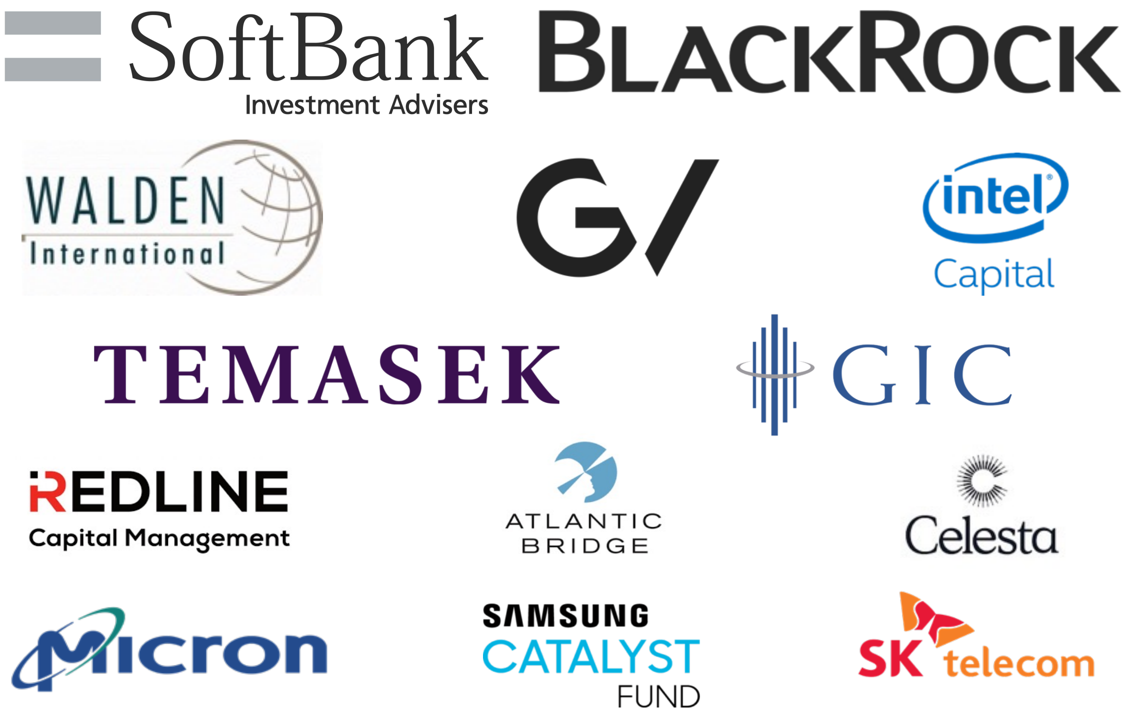
500+
HW/SW
AI Engineers

Kunle Olukotun
Professor EE/CS
Stanford University

Rodrigo Liang
CEO

Chris Ré
Professor CS
Stanford University

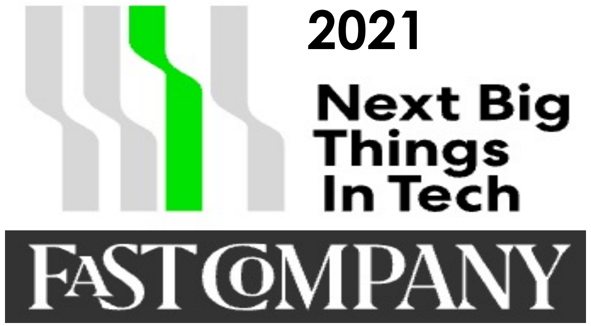
SambaNova Systems is the Best Funded AI Start-up



Over $1 Billion Raised Through Series D

Industry Awards and Recognition

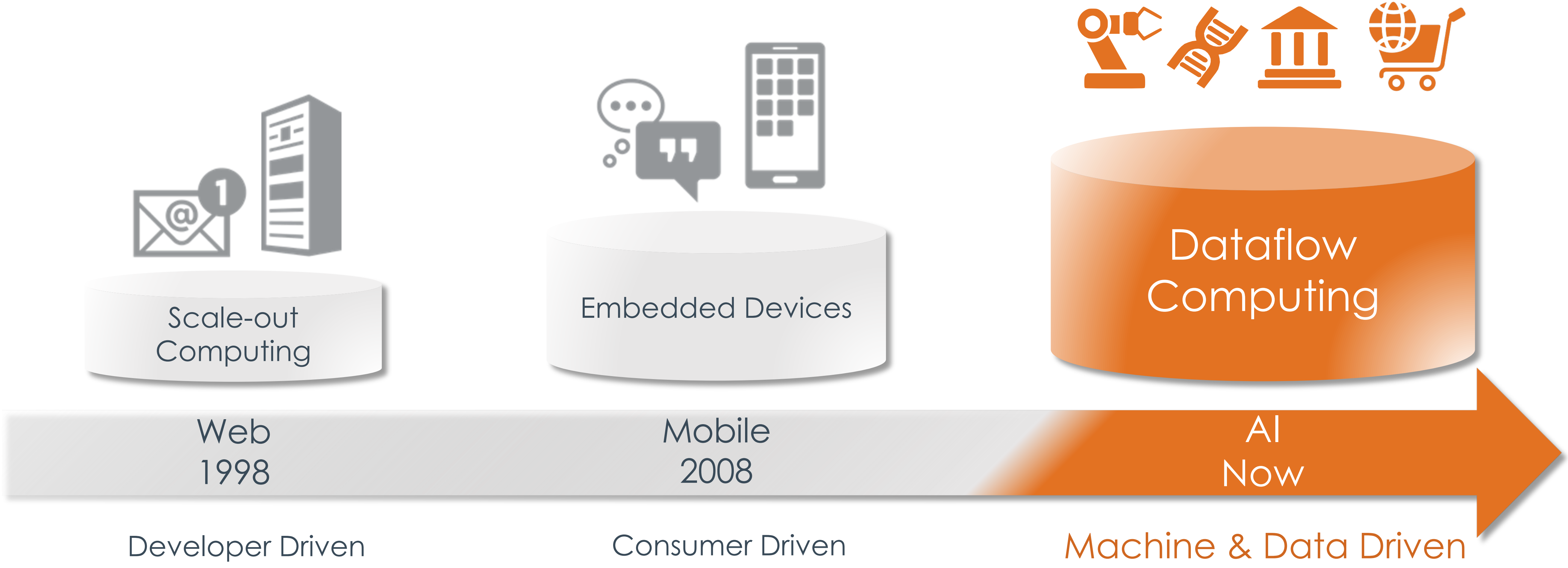
2021 GSA AWARDS CELEBRATION
WINNER
Most Respected Private Semiconductor Company
SambaNova
SYSTEMS
GSA



CogX Festival 2021
14th - 16th June 2021
Virtual & Kings Cross, London
Best AI Product in Next-Generation Infrastructure
CogX Awards 2021
Winners
SambaNova
SYSTEMS

Why AI/ML, Why Now?

The Biggest Transformation in Business and Tech is Here



Scale-out Computing

Web
1998

Developer Driven

Embedded Devices

Mobile
2008

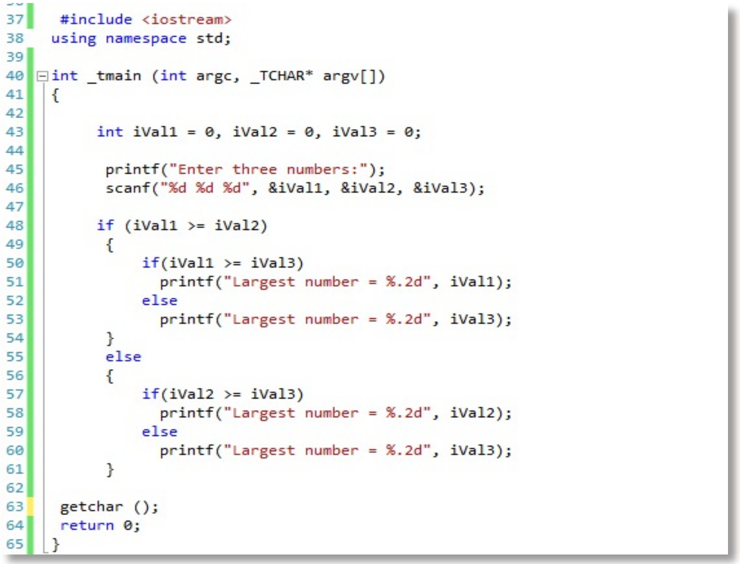
Consumer Driven

Dataflow Computing

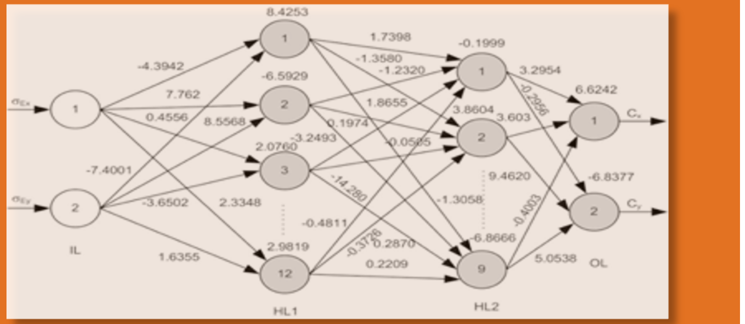
AI
Now

Machine & Data Driven

AI is a Software Problem - Models Are The New Code



```
 #include <iostream>
using namespace std;

int _tmain (int argc, _TCHAR* argv[])
{

     int iVal1 = 0, iVal2 = 0, iVal3 = 0;

      printf("Enter three numbers:");
      scanf("%d %d %d", &iVal1, &iVal2, &iVal3);

     if (iVal1 >= iVal2)
      {
           if(iVal1 >= iVal3)
             printf("Largest number = %.2d", iVal1);
           else
             printf("Largest number = %.2d", iVal3);
      }
      else
      {
           if(iVal2 >= iVal3)
             printf("Largest number = %.2d", iVal2);
           else
             printf("Largest number = %.2d", iVal3);
      }

 getchar ();
 return 0;
}
```



Traditional (Software 1.0)

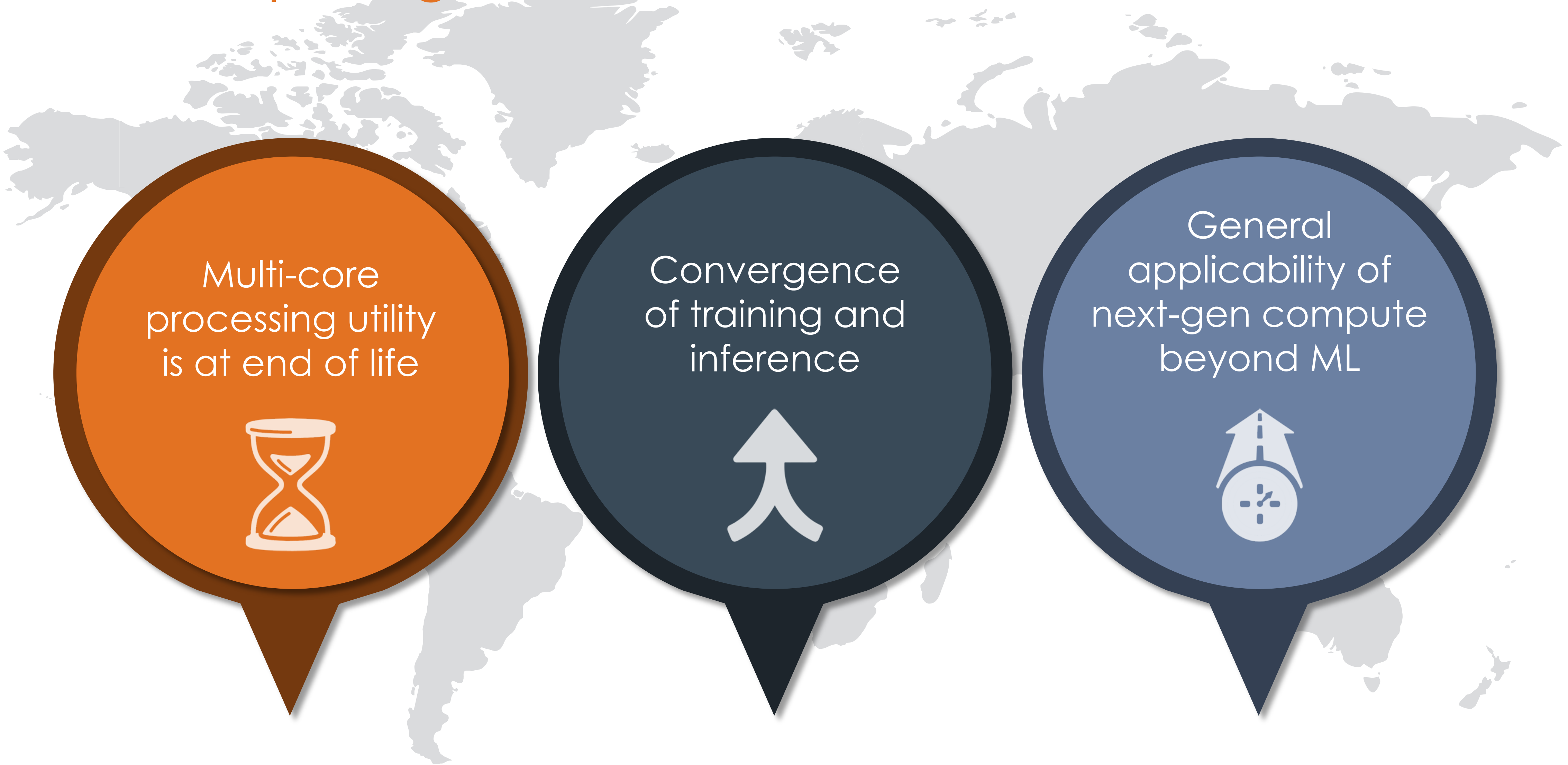
- Written in code (C++, ...)
- Requires domain expertise
  - Decompose the problem
  - Design algorithms
  - Compose into a system

AI is all about DataFlow (Software 2.0)

- Data -not code- trains the models
- Written in the weights of a Neural Network
- Fewer lines of code, greater productivity, improved accuracy

Andrej Karpathy. Scaled ML 2018 talk

Three Computing Trends for ML

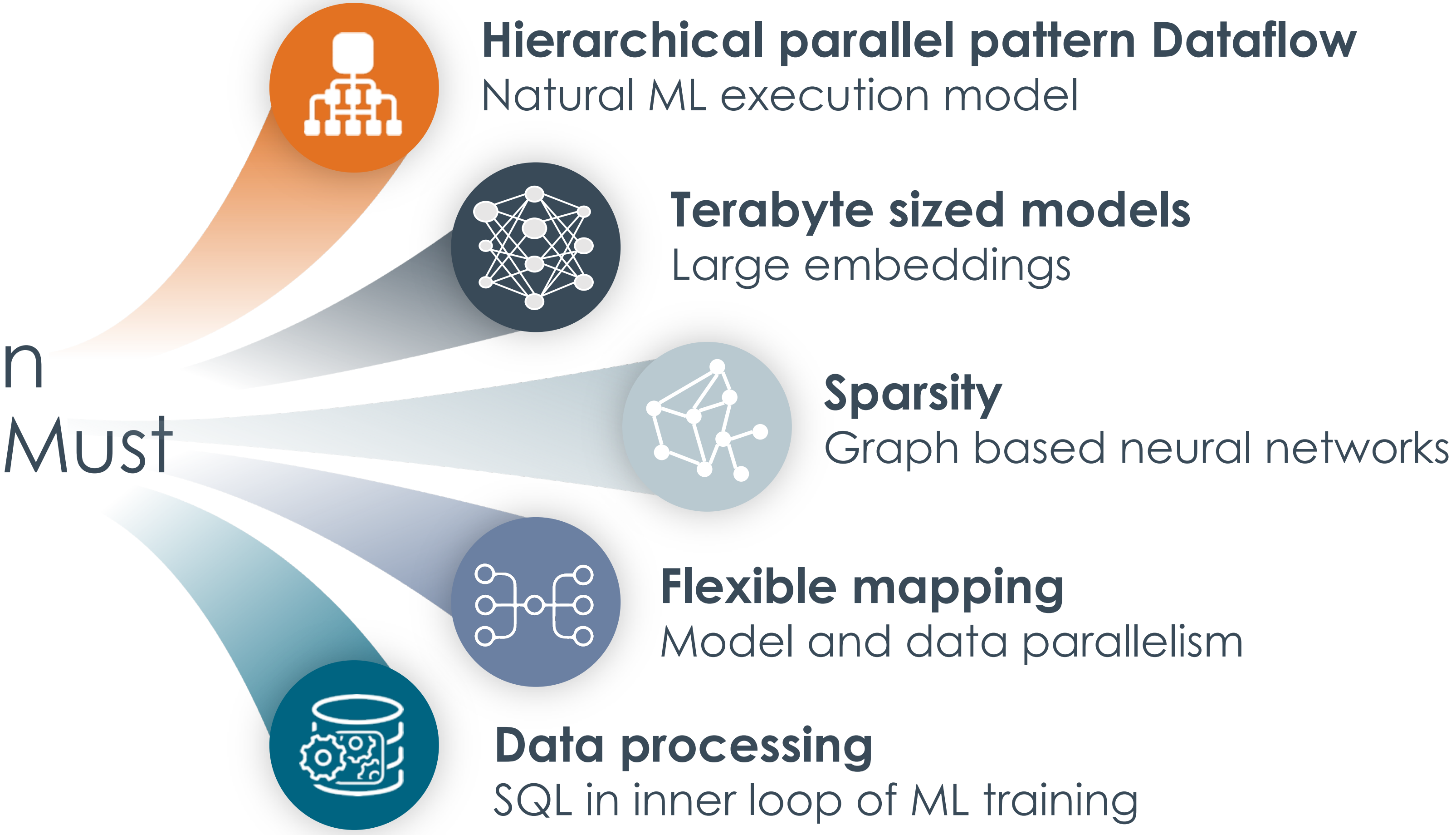


Multi-core processing utility is at end of life

Convergence of training and inference

General applicability of next-gen compute beyond ML

Next Generation Compute Must Support...



Hierarchical parallel pattern Dataflow
Natural ML execution model

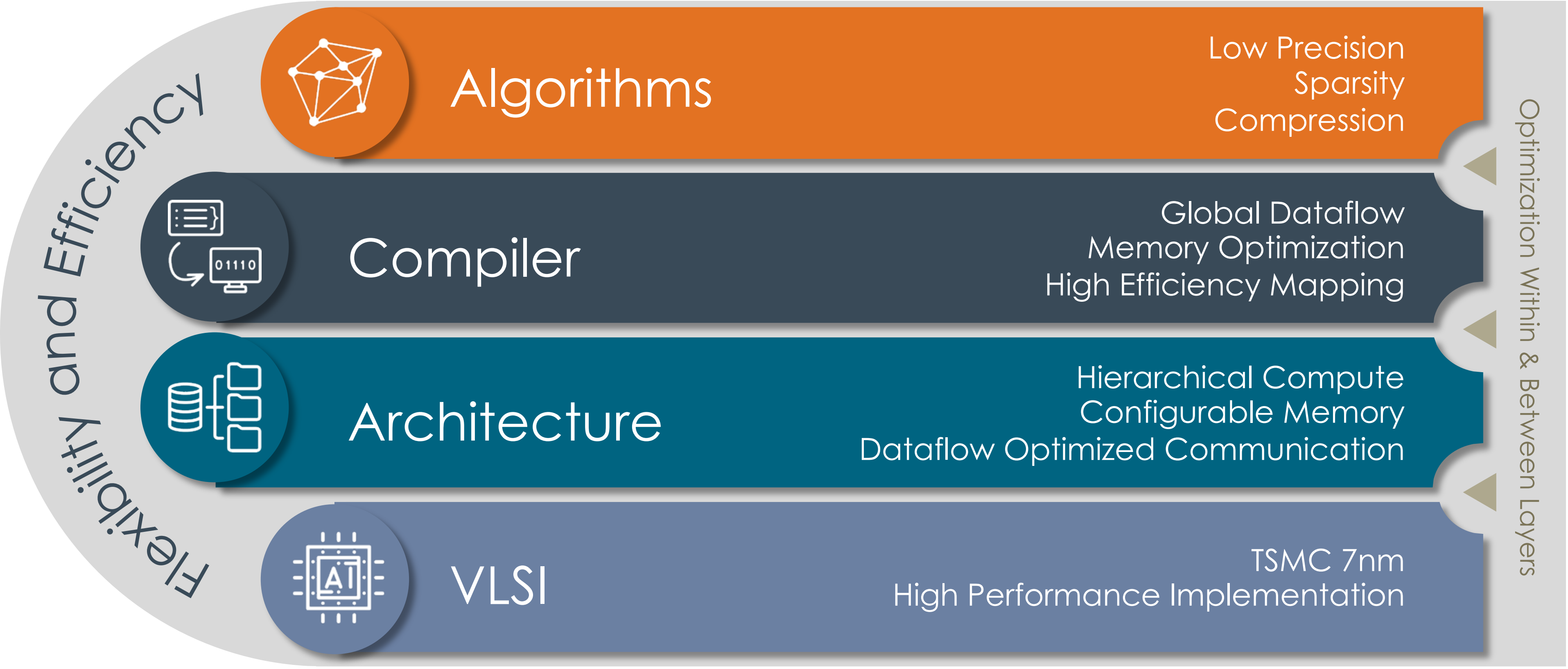
Terabyte sized models
Large embeddings

Sparsity
Graph based neural networks

Flexible mapping
Model and data parallelism

Data processing
SQL in inner loop of ML training

The SambaNova Systems Advantage: Vertically Integrated Approach



Flexibility and Efficiency

Algorithms

Low Precision
Sparsity
Compression

Compiler

Global Dataflow
Memory Optimization
High Efficiency Mapping

Architecture

Hierarchical Compute
Configurable Memory
Dataflow Optimized Communication

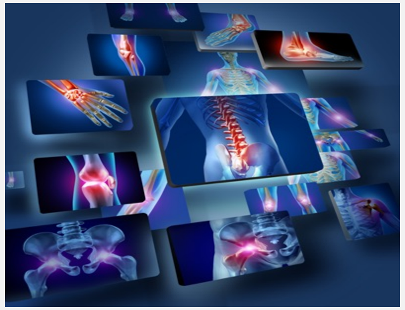
VLSI

TSMC 7nm
High Performance Implementation

Optimization Within & Between Layers

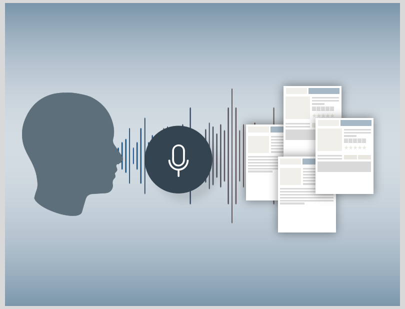
Four Classes of Models Covering Broad Industries

True Resolution Computer Vision



Every imaging device

Natural Language Processing



Speech and text is ubiquitous

Recommendation

80% of online retail AI investment

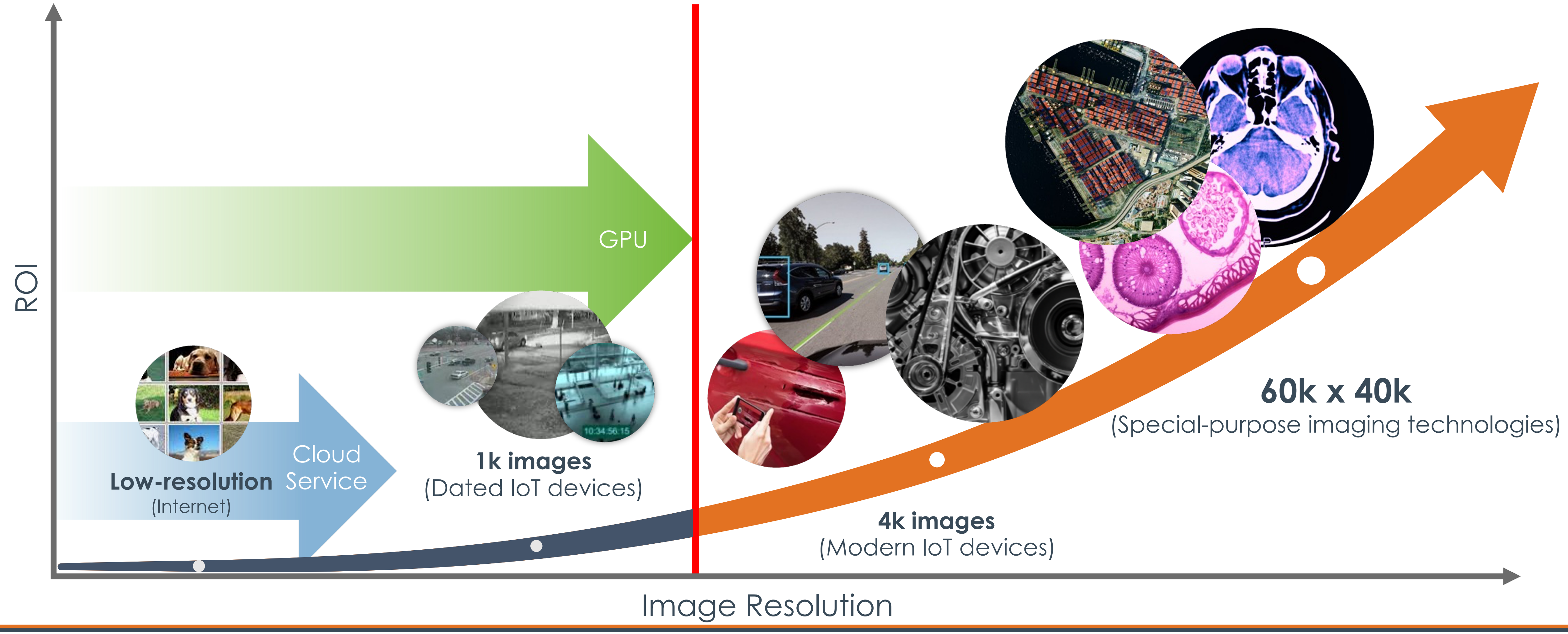
AI for Science

Exascale data processing

True Resolution Vision

Limited Possibilities with Conventional Solutions

Today's computer vision capabilities are falling behind Imaging technology



Case Study: True Resolution Computer Vision

Break through the accuracy limit from conventional solutions

Conventional Solution

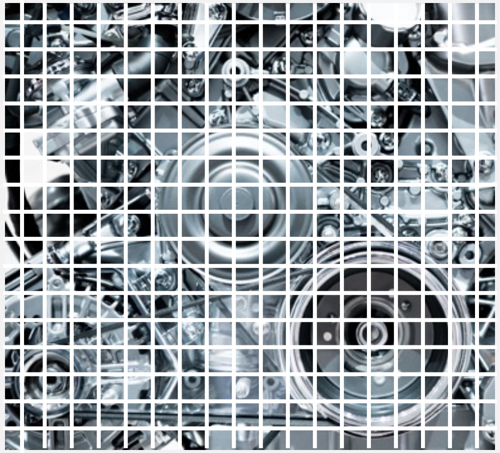
MANDATED WORKAROUNDS TO TRAIN

1. Down Sampling

- Large field of view with blurry large features
- Loss of details
- Lowered resolution

2. Patching

- Small field of view
- Loss of large features in each tile
- Loss of information at boundaries
- Requires more compute resources



Failed to identify the defect

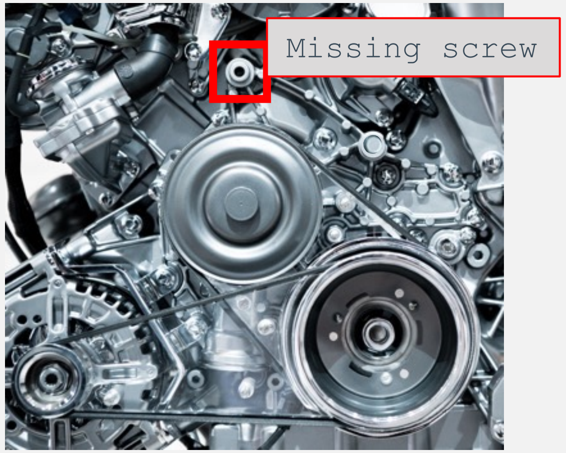
SambaNova SYSTEMS

TRAIN IN NATURAL FORM AS-IS

Train model at true resolution of the original high-res images

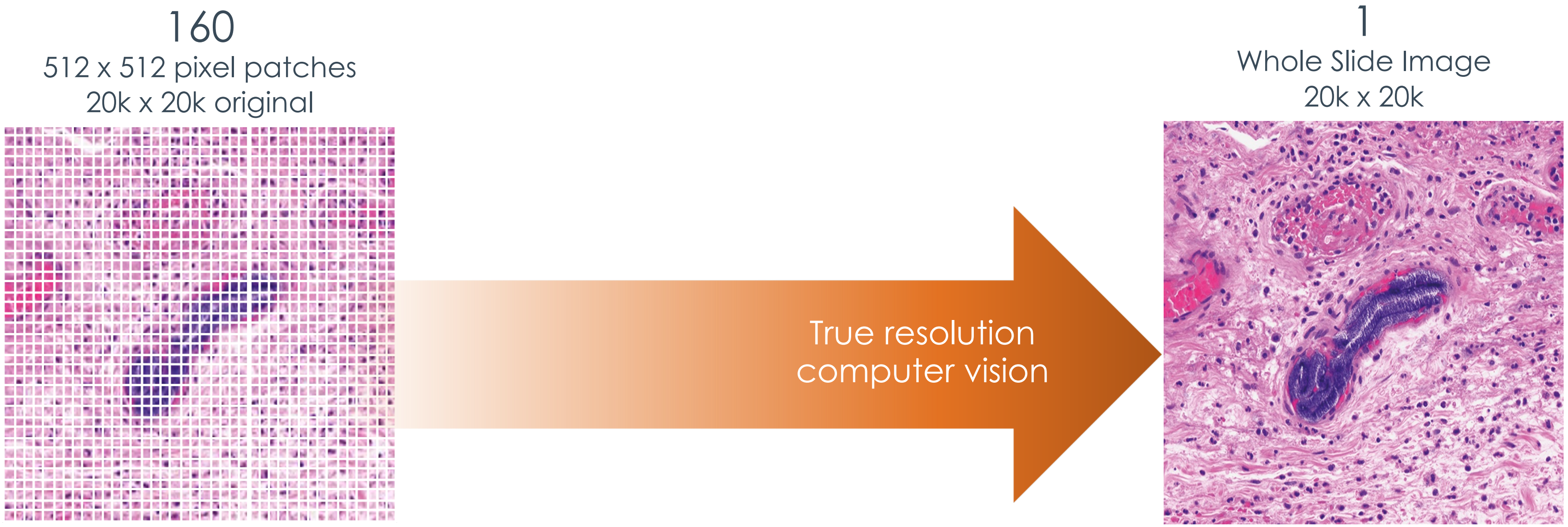
- Large field of view and high-resolution details in the same pipeline
- No compromise on accuracy

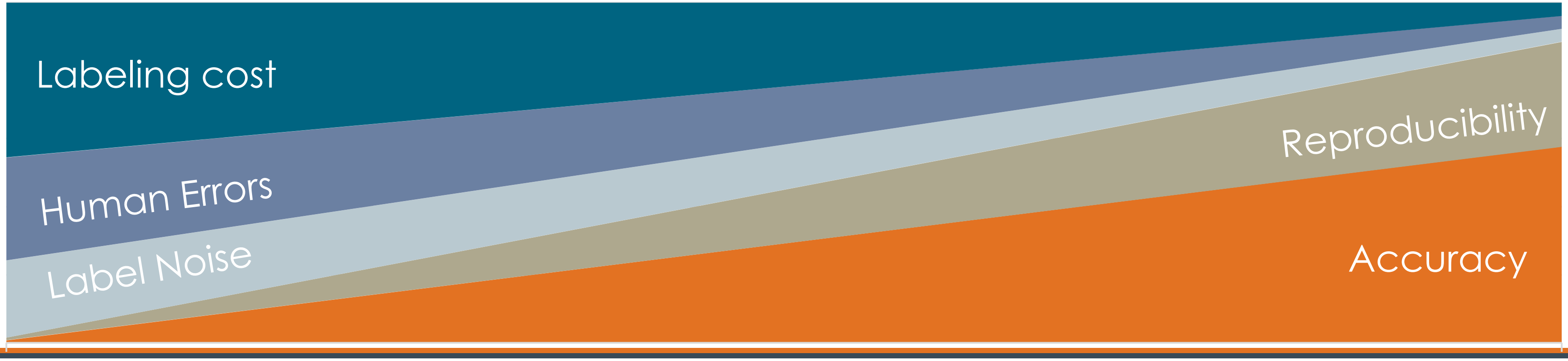
State-of-the-art Accuracy



Dataflow-as-a-Service Efficiently Trains with Your Images

Eliminate the costly workarounds from conventional solutions





Natural Language

DataFlow-as-a-Service™ GPT: for OTP Bank

Deployment of next generation language as-a-service capabilities

What is being announced?

- Multiple versions of GPT pre-trained with English language
- Domain specific pre-training for finance
- Service Delivery on/before Feb 2022
- Training and Inference
- Multi-year subscription
- Multi-rack installation

otpbank

*"This is a unique collaboration between **OTP Group, ITM, and SambaNova Systems to provide an incredible resource to the country and the Central and Eastern European region**. We are pleased to announce that **this groundbreaking supercomputer will represent a unique AI capability to build GPT-3 level language models** for languages across CEE."*

Péter Csányi, Deputy CEO, Head of Digital Divison, OTP Group

Why is it significant?

- OTP has leapfrogged regional competitors with significantly more advanced language capabilities at a fraction of the time and cost

- OTP's language as-a Service deployment will enable a next generation of model driven application products and services to accelerate digital customer acquisition

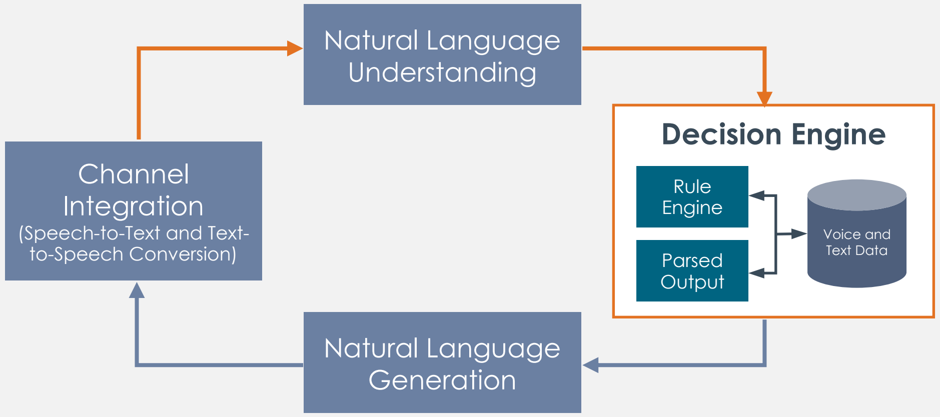
DataFlow-as-a-Service^{TM} GPT: for OTP Bank

European AI supercomputing for large language models on SambaNova

Textual Data

Voice Activated Devices

Mobile/ Web App Chatbot



Natural Language Understanding
Channel Integration (Speech-to-Text and Text-to-Speech Conversion)
Decision Engine
Rule Engine
Parsed Output
Voice and Text Data
Natural Language Generation

Initial Use Cases

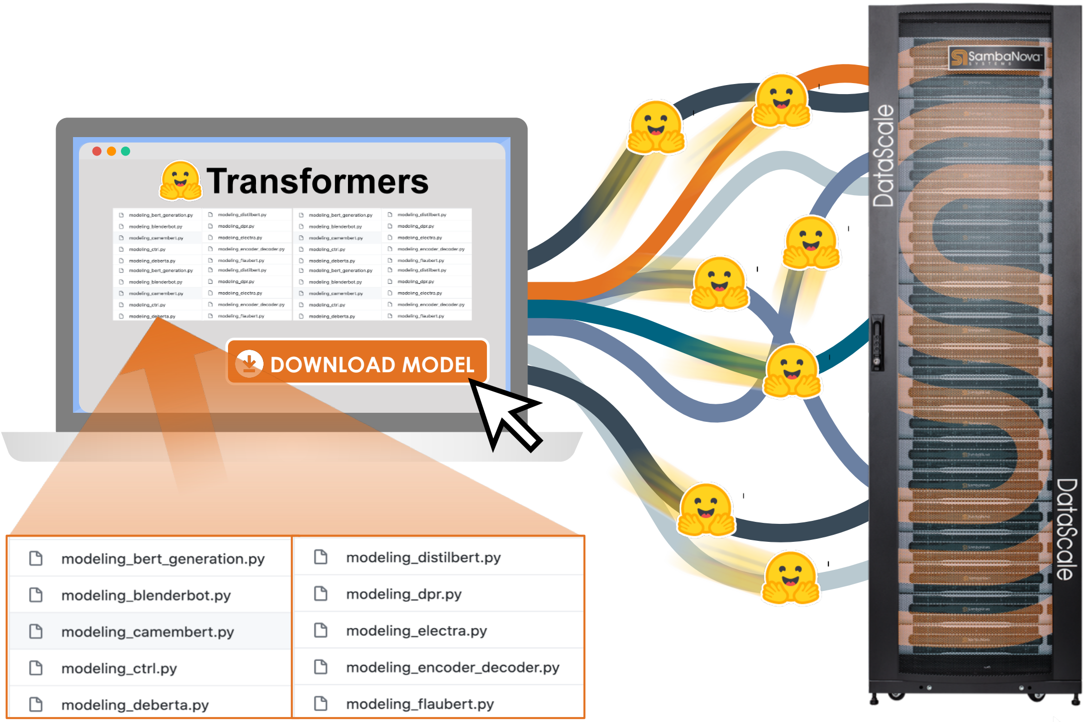
- Complex linguistics
- Sentiment Analysis
- Document Classification
- Named-Entity Recognition
- Text/Document Generation

Line of Business Applications

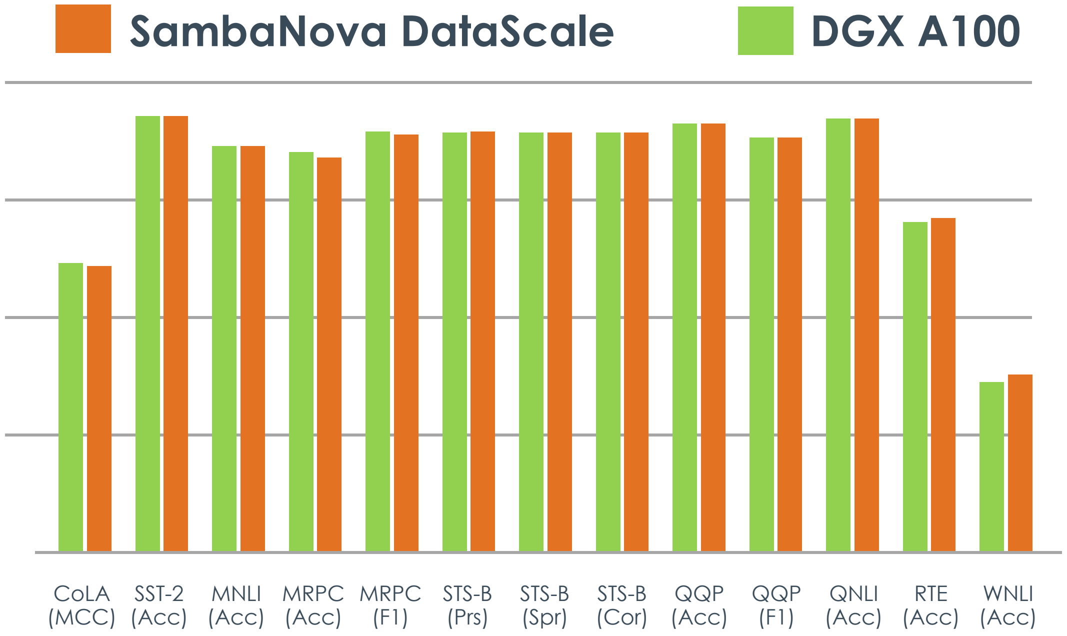
- CRM (contact center)
- Customer Engagement
- Omni Channel Marketing
- Loan Origination & Servicing
- Conduct Surveillance
- Digital Voice Innovations
- Risk & Compliance

Run State of the Art Accuracy Transformers in Seconds

Instantly run thousands of Hugging Face models with zero code change



- modeling_bert_generation.py
- modeling_blenderbot.py
- modeling_camembert.py
- modeling_ctrl.py
- modeling_deberta.py
- modeling_distilbert.py
- modeling_dpr.py
- modeling_electra.py
- modeling_encoder_decoder.py
- modeling_flaubert.py



Downloadable Models

Results in Seconds

Same Accuracy as on GPU

Recommendation

World Record Recommender Training and Inference

Highest throughput DIN recommender results

AI Matrix.

AI Matrix.

AI Matrix: A Deep Learning Benchmark for Alibaba Data Centers

Wei Zhang, Wei Wei, Lingjie Xu, Lingling Jin
Alibaba Group

Cheng Li
University of Illinois Urbana-Champaign

Introduction

Overview of AI Matrix

Results

Roofline Model

Low-Precision Inference Results

Website and GitHub

Future Development

5x Faster

Training Than GPU

Training Performance on NVIDIA V100 and T4

	V100			T4		
	batch 256	batch 512	batch 1024	batch 256	batch 512	batch 1024
DIN	12493.394	14781.57	15198.929	10423.465	10841.937	10085.002

8x Faster

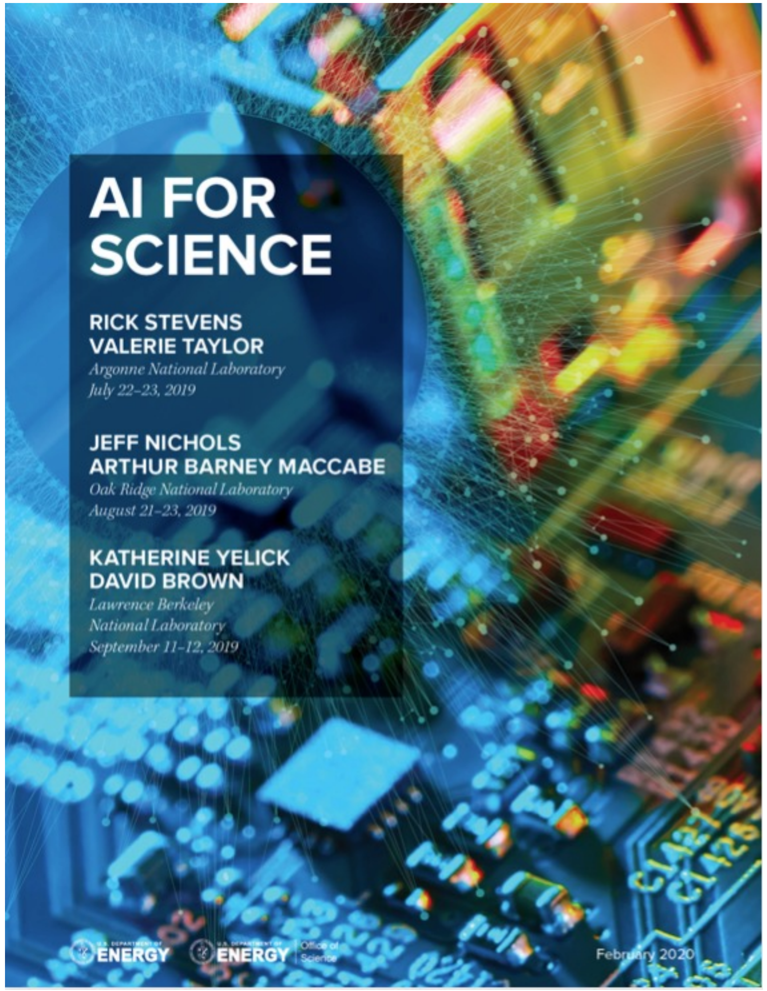
Inference Than GPU

Inference Performance on NVIDIA V100 and T4

	V100			T4		
	batch 256	batch 512	batch 1024	batch 256	batch 512	batch 1024
DIN	95296.656	130629.08	150438.253	66888.048	73245.628	67220.958

AI For Science

HPC + AI



AI FOR SCIENCE

RICK STEVENS
VALERIE TAYLOR
Argonne National Laboratory
July 22–23, 2019

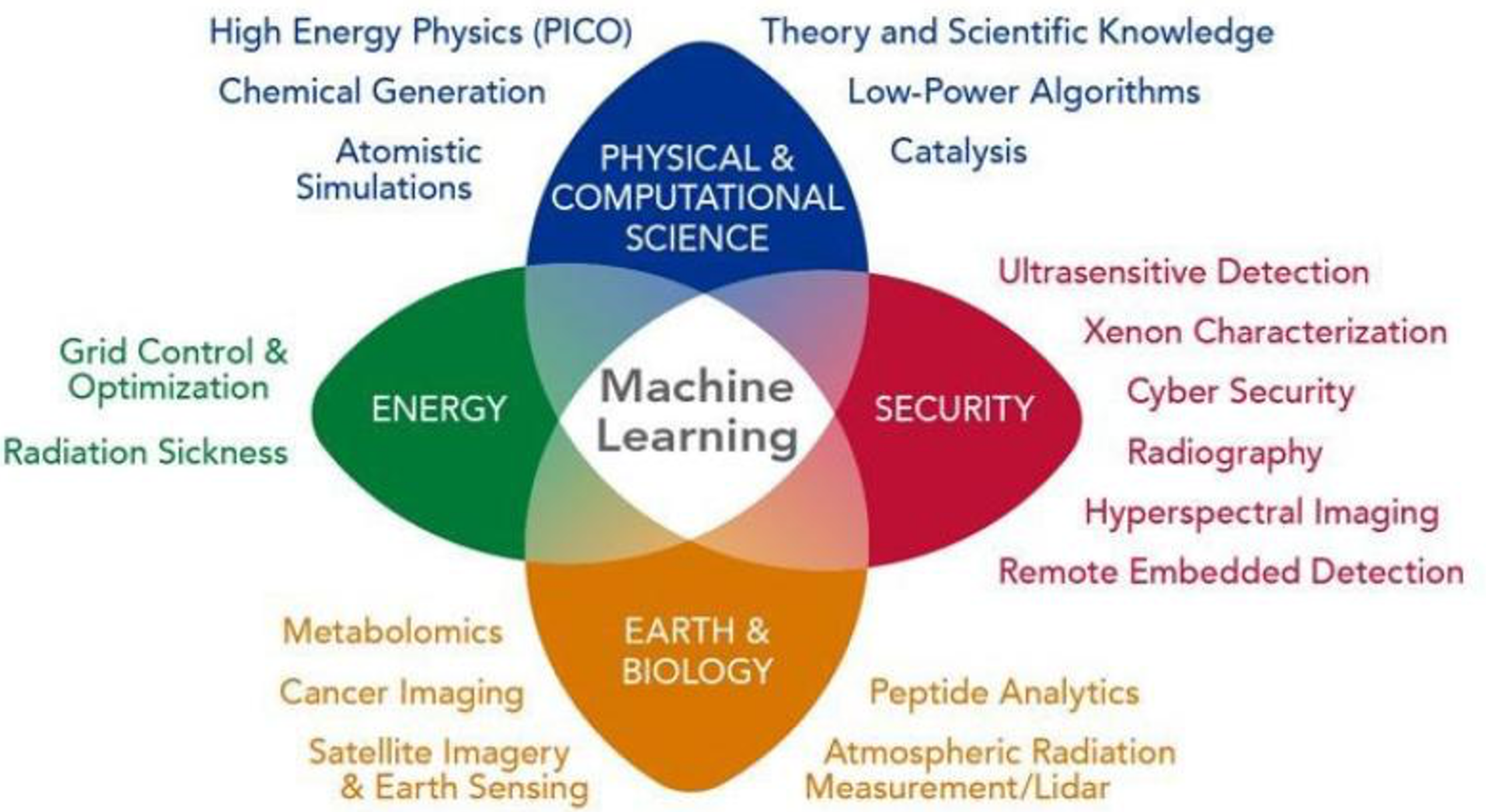
JEFF NICHOLS
ARTHUR BARNEY MACCABE
Oak Ridge National Laboratory
August 21–23, 2019

KATHERINE YELICK
DAVID BROWN
Lawrence Berkeley
National Laboratory
September 11–12, 2019

U.S. DEPARTMENT OF ENERGY
U.S. DEPARTMENT OF ENERGY | Office of Science

February 2020

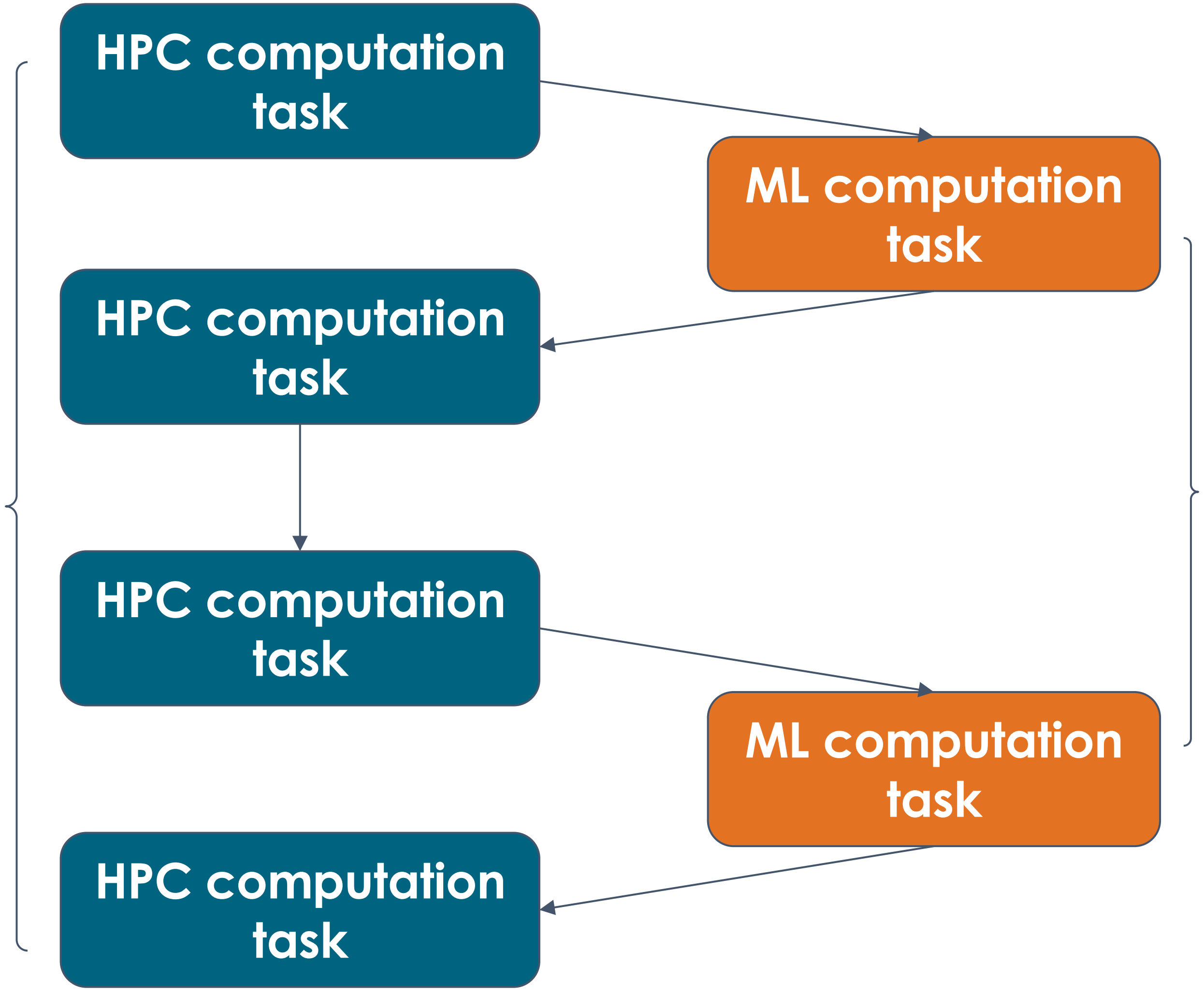
Machine Learning and AI Intersects and Helps with HPC



* Artificial Intelligence and Machine Learning to Accelerate Translational Research: Proceedings of a Workshop—in Brief (2018)

Hybrid HPC Pipeline with SambaNova Systems

Existing supercomputer for HPC

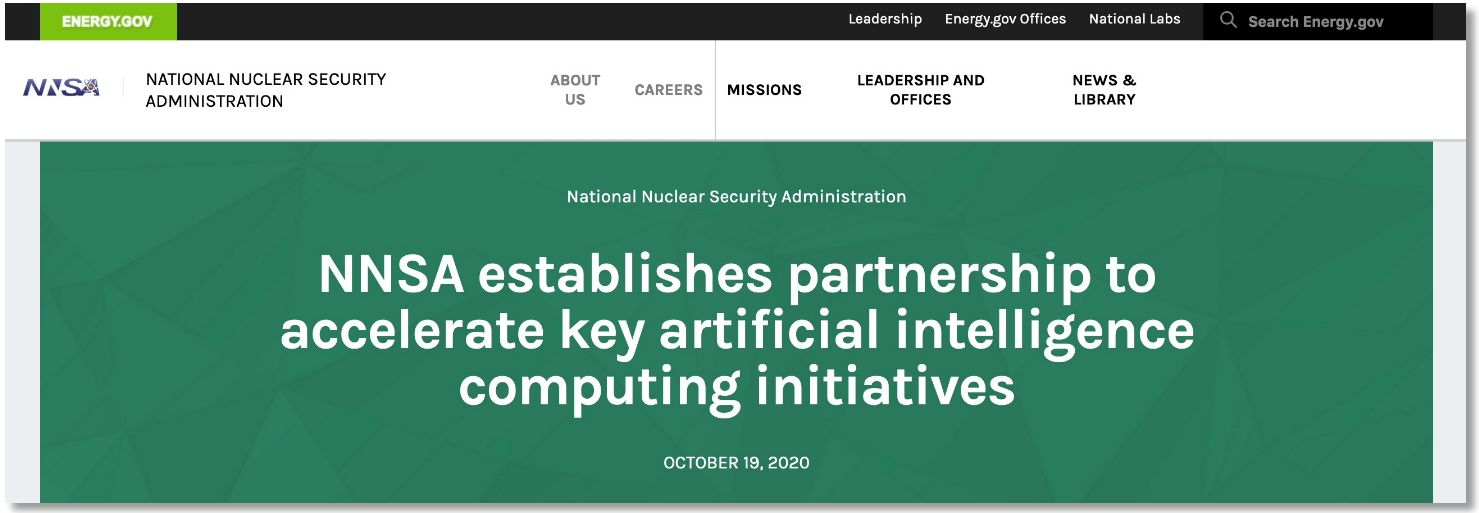


AI acceleration system like **Sambanova SN-10 with Reconfigurable Dataflow Units (RDUs)**

Strategic Partnership and a National Imperative

U.S. DEPARTMENT OF ENERGY

SambaNova SYSTEMS



ENERGY.GOV Leadership Energy.gov Offices National Labs Search Energy.gov

NATIONAL NUCLEAR SECURITY ADMINISTRATION ABOUT US CAREERS MISSIONS LEADERSHIP AND OFFICES NEWS & LIBRARY

National Nuclear Security Administration

NNSA establishes partnership to accelerate key artificial intelligence computing initiatives

OCTOBER 19, 2020

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY

"This strategic partnership agreement with SambaNova Systems will highlight extraordinary efforts that will contribute to the continuous advancing of AI and machine learning initiatives within DOE and the NNSA"

-David Etim, a federal program manager for the NNSA Office of Advanced Simulation and Computing and leader of the Advanced Machine Learning Initiative

Solutions

Deployed at Select Revenue Customers Since H1 CY2020

Real customers, real workloads, real results

Software:
No lock-in, CUDA-free computing

Hardware:
SW-defined, Dataflow optimized

System:
Ease of use and integration

Results:
New capabilities, performance

Software, Hardware, Services, Support

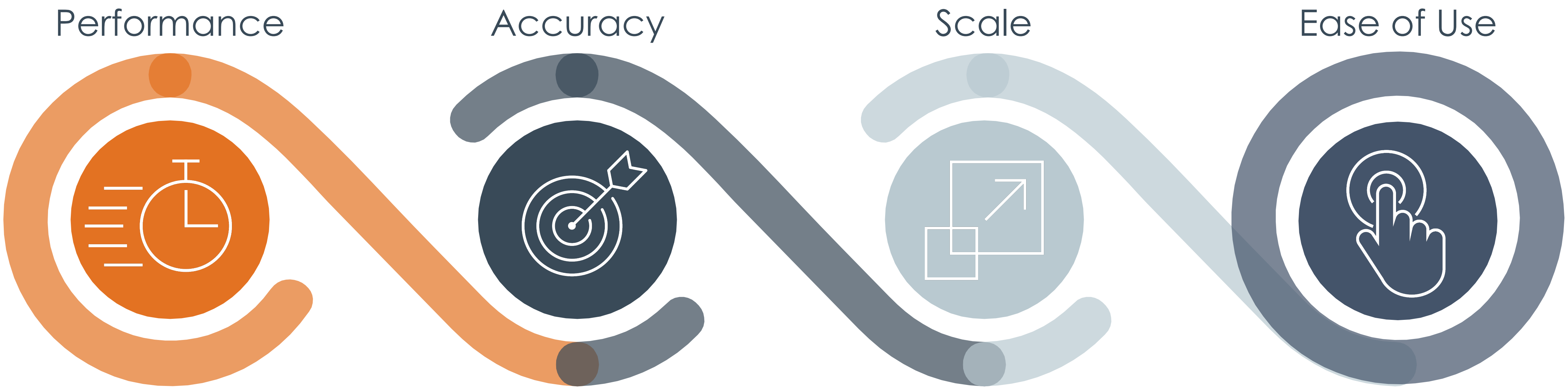
Multi-System Deployments

100% Remote Installations in 45 Mins

Customer Models Running on Day 1

AI Computing is Multi-Faceted and Complex

It’s not just about performance



Performance

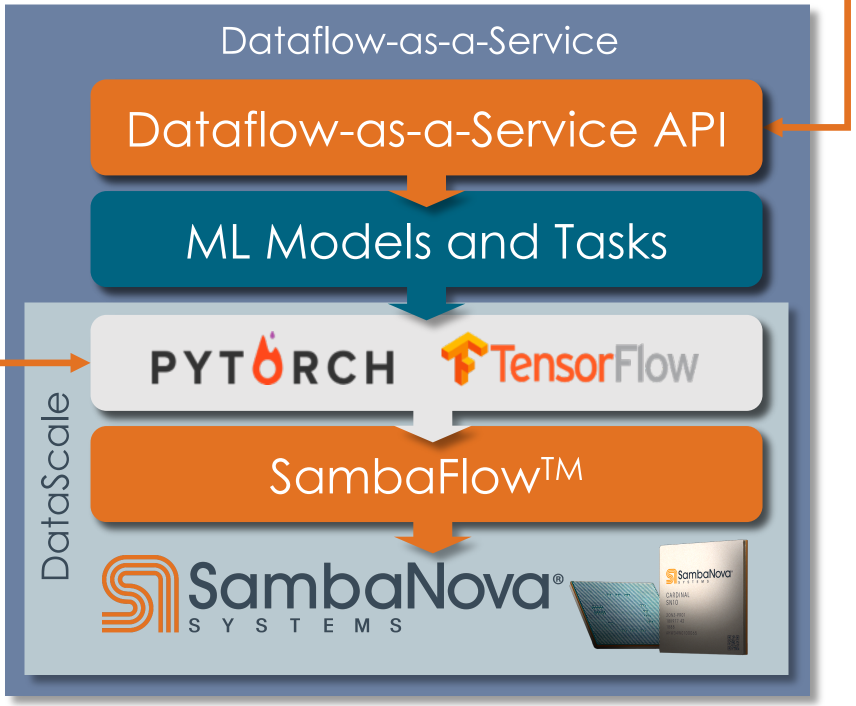
Accuracy

Scale

Ease of Use

One Platform With All the *Right* Interfaces You Need

Choose DataScale or Dataflow-as-a-Service , on-premises or in the cloud



On-Premises, Hybrid, Public Cloud

Dataflow-as-a-Service

Extensible Platform Services

- Data-centric computing
- Low Code Interfaces
  - API/SDK
  - GUI
  - CLI
- Application developers

DataScale

The Platform for Innovation

- Model-centric computing
- Common open interfaces
  - PyTorch, TensorFlow
  - SambaFlowTM
- ML engineers and data scientists

World’s First Dataflow-as-a-Service Offerings

Three Dataflow-as-a-Service subscriptions: From zero to AI, fast and simple

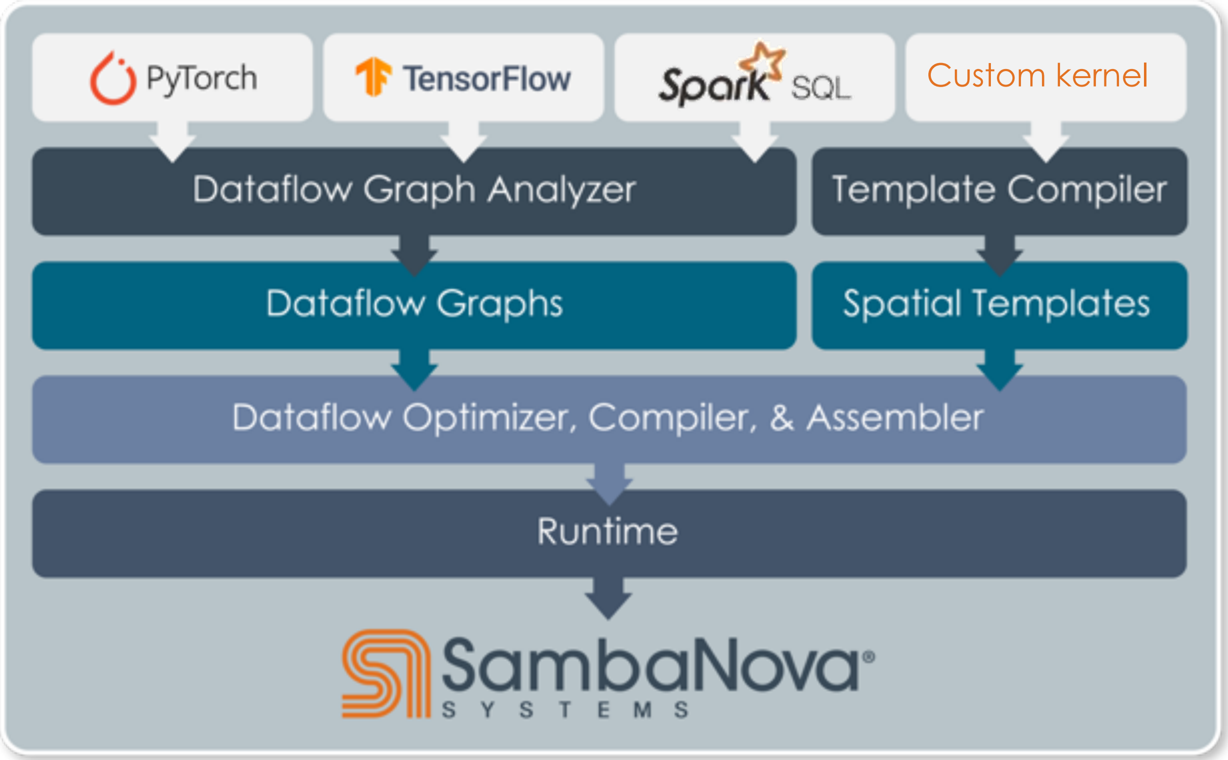
Extensible AI Services Platform

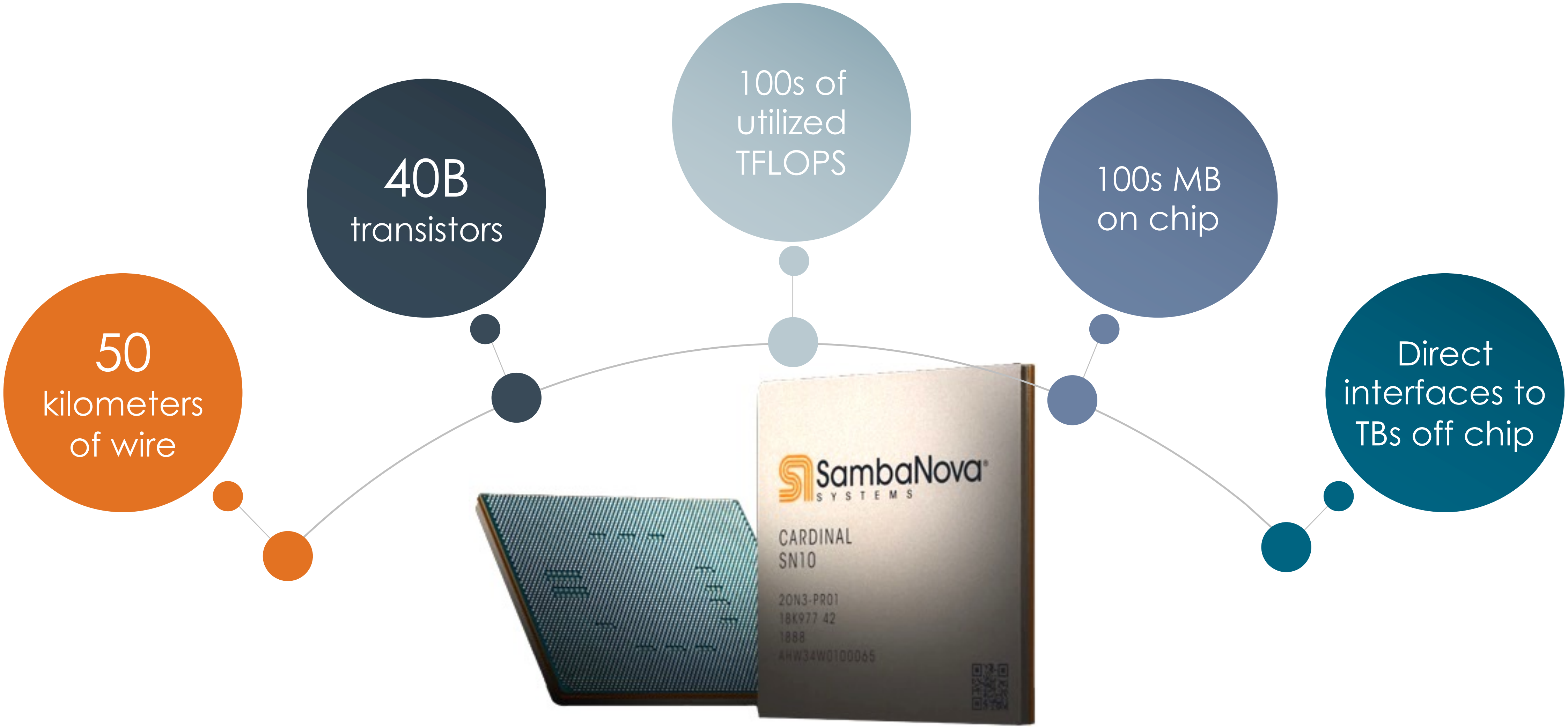
- AI cloud services @ customer site
- Scale on-demand
- State of the art accuracy
- Ease of use at scale
- Managed by SambaNova
- Cloud consumption OPEX model

SambaFlow Open Software

SambaFlow: No lock-in, ease of use, developer productivity

Complete system solutions require an open and easy to use software stack





50 kilometers of wire

40B transistors

100s of utilized TFLOPS

100s MB on chip

Direct interfaces to TBs off chip

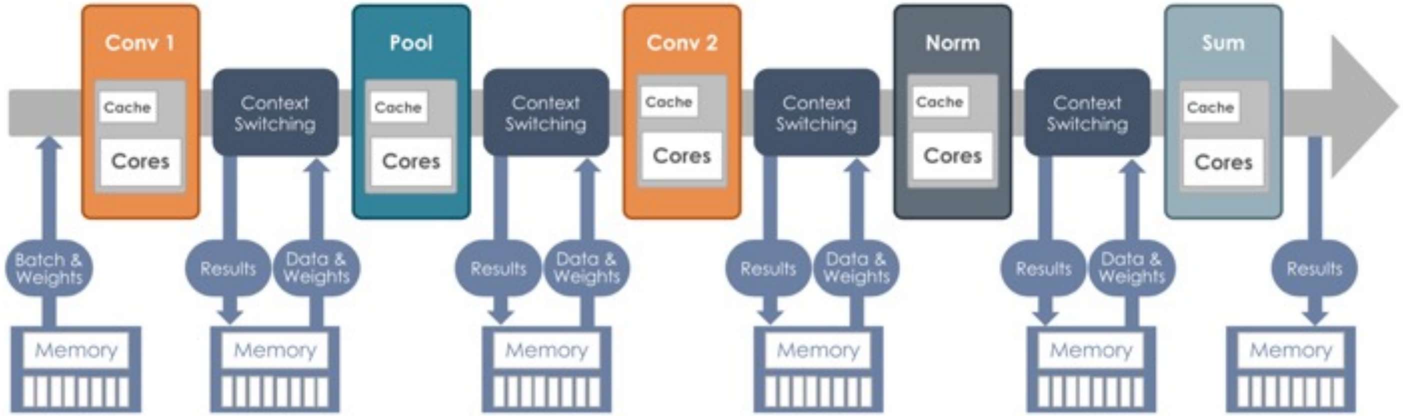
SambaNova Systems Cardinal SN10 RDU

World’s First Reconfigurable Dataflow Unit

Existing Solutions Are Built For Software 1.0

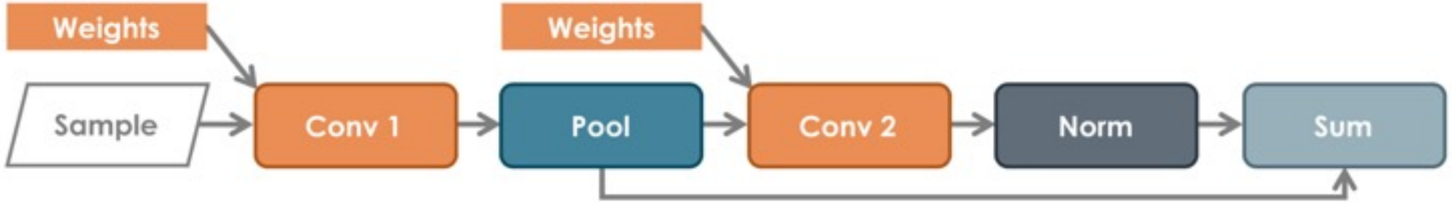
```
t1 = conv(in)
t2 = pool(t1)
t3 = conv(t2)
t4 = norm(t3)
t5 = sum(t4)
```

Sequence of instructions executed in time

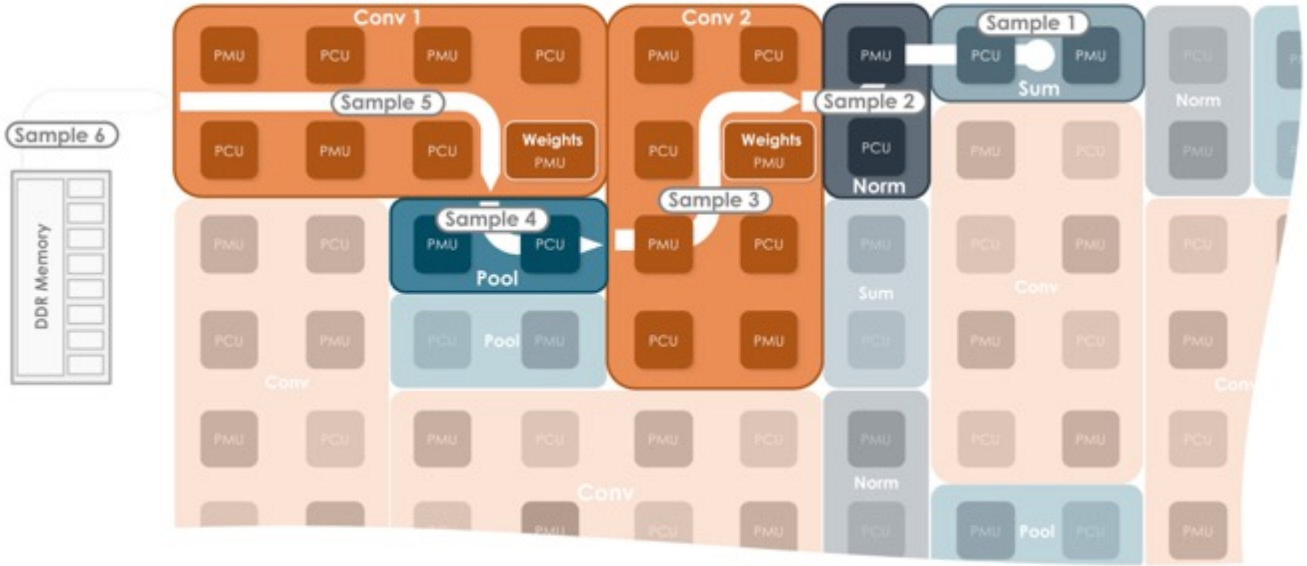


The old way: Kernel-by-kernel
Bottlenecked by memory bandwidth and host overhead

Software 2.0 is Dataflow



Convolution Graph

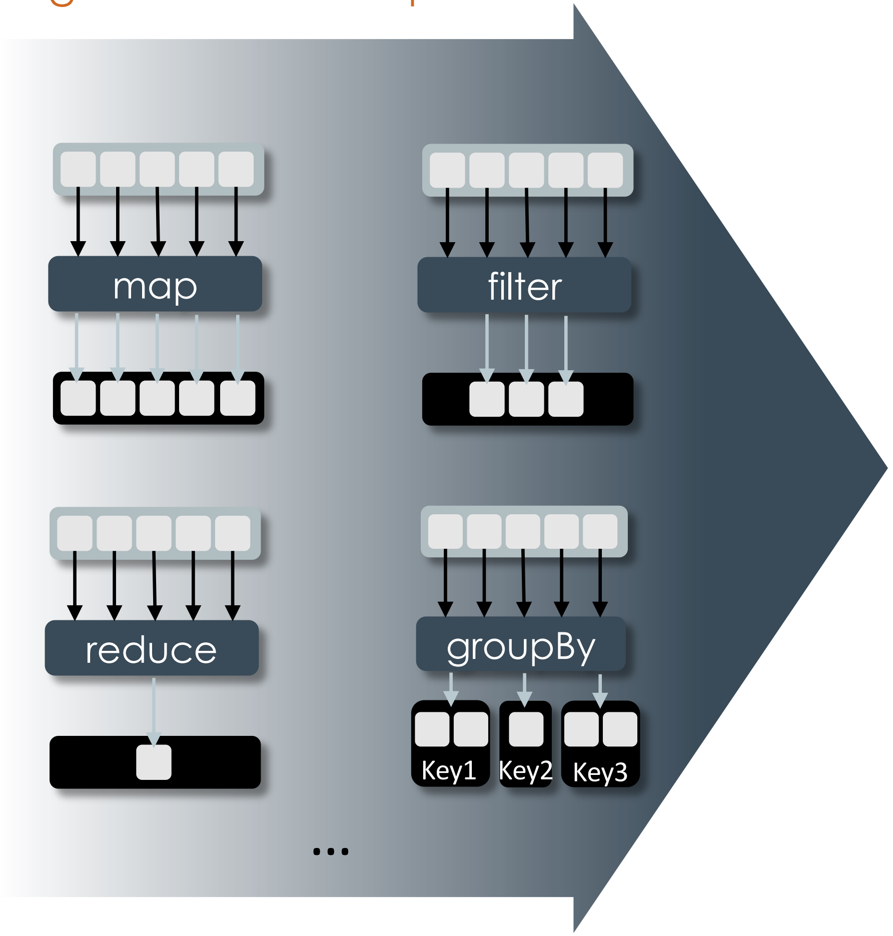


The Dataflow way: Spatial programming

Eliminates memory traffic and overhead, increases parallelism

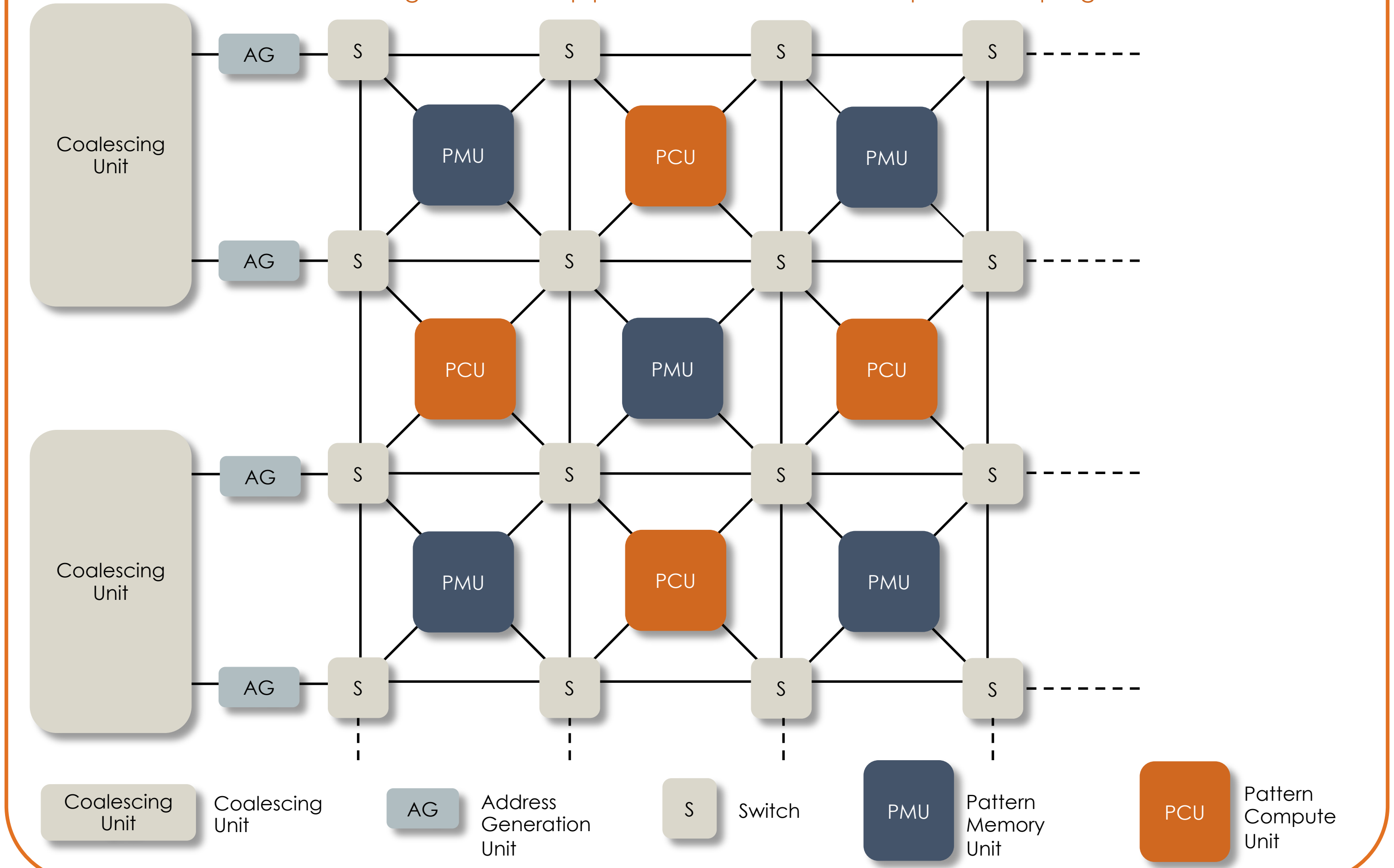
Software-Defined Hardware Architecture

Programmable Compute **and Communication**



Software-Driven Architecture

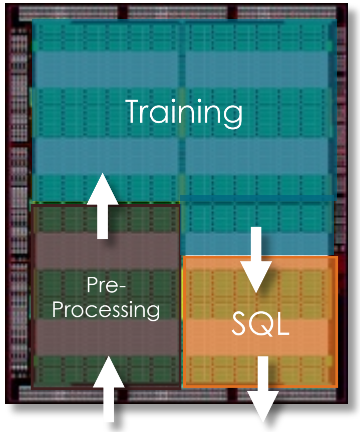
Tiled architecture with reconfigurable SIMD pipelines, distributed scratchpads, and programmed switches



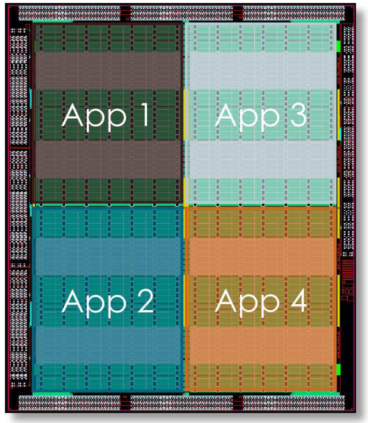
SambaNova Systems Flexibility to Support Key Scenarios

4 RDU deployment examples

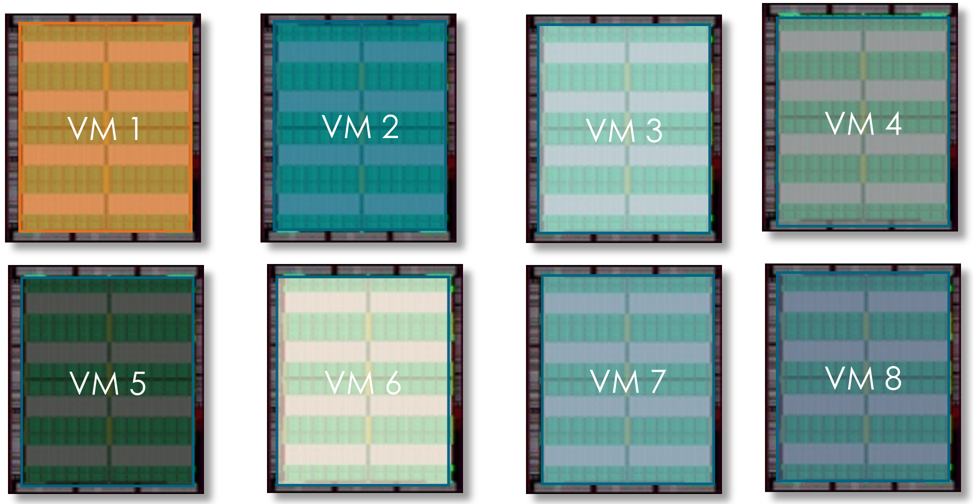
1) High Performance Mixed Workloads



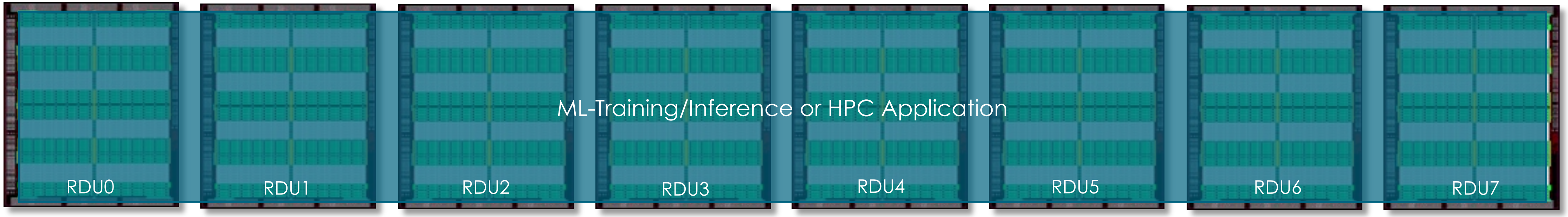
2) Efficient Concurrent Applications



3) Secure Multi-Tenancy

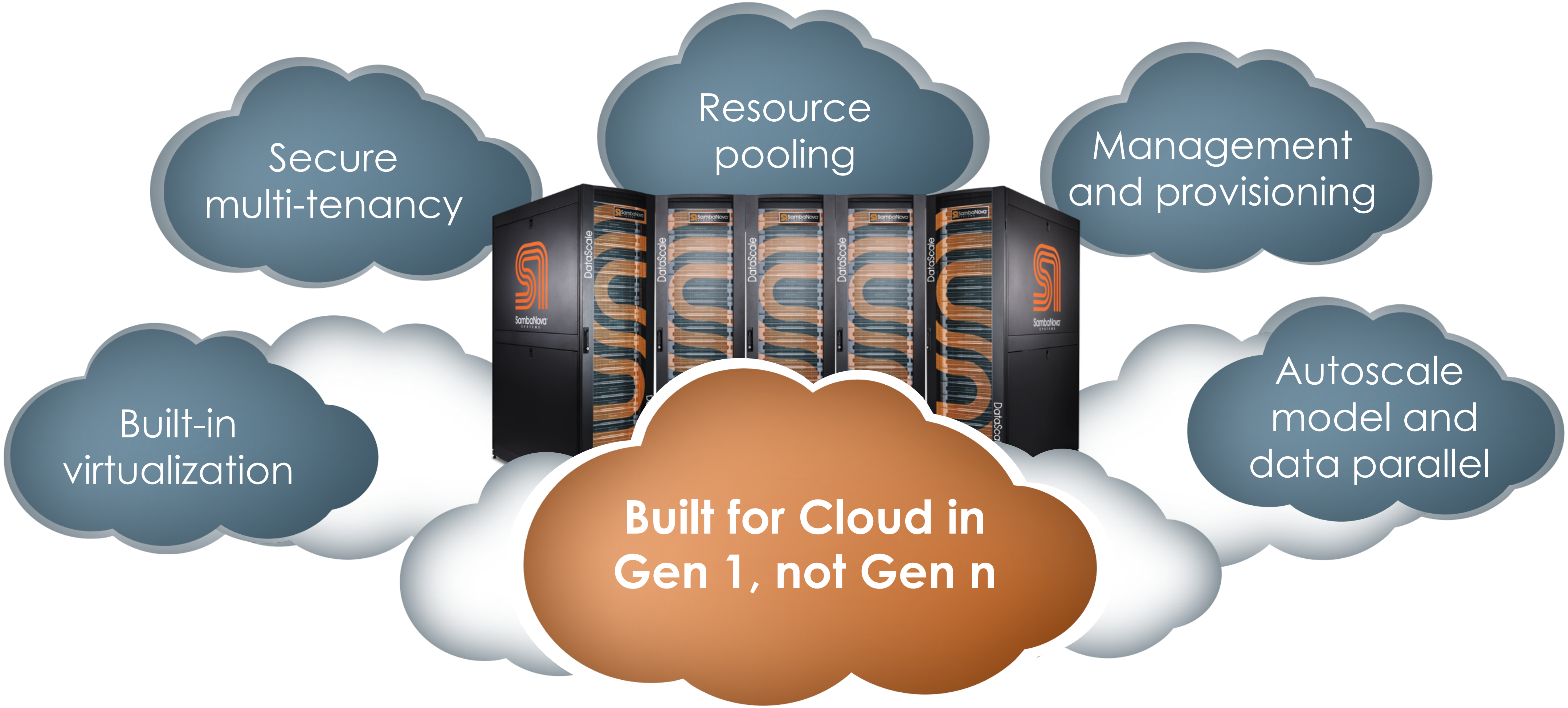


4) Compiler Driven Application Scale-Up



Delivering Cloud-Native Features From Inception

Enterprise, service provider, supercomputing



Secure multi-tenancy

Resource pooling

Management and provisioning

Built-in virtualization

Autoscale model and data parallel

Built for Cloud in Gen 1, not Gen n

Open Standards, Disruptive Technology, Easy to Deploy

AI infrastructure deployed and running customer workloads in 45 minutes

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

The New Standard

Open Source Frameworks

PYTORCH TensorFlow

Open Source Orchestration

kubernetes slurm workload manager

Open Source Containers

docker Singularity

Open Source OS

Red Hat ubuntu

Open Standards Connectivity

Ethernet PCI EXPRESS

Dataflow-as-a-Service Unlocks New Capabilities

From zero to AI with one subscription

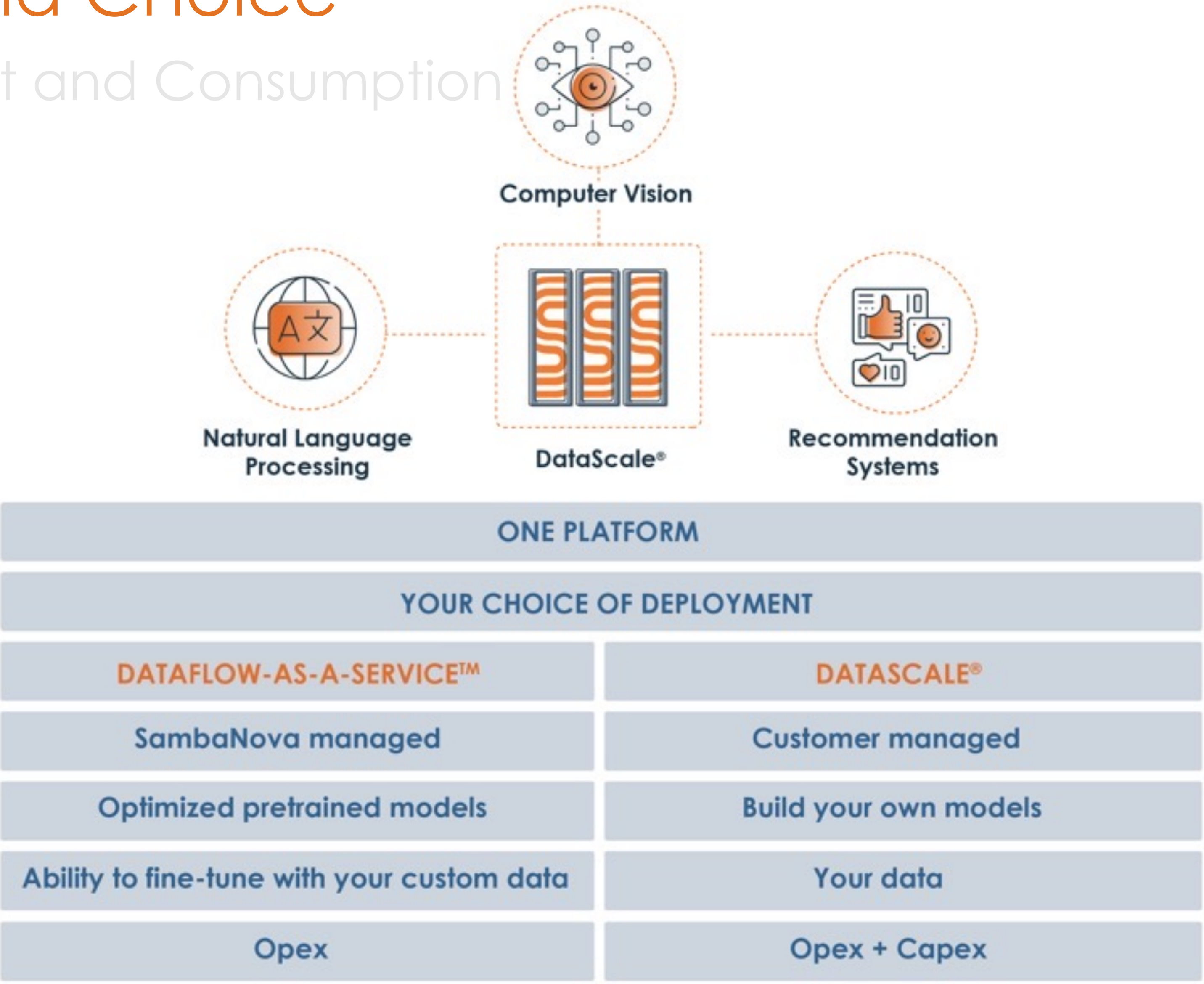
SambaNova Advantages:

- Intelligent software stack
- Dataflow-driven
- High compute efficiency
- Big memory architecture

High-Resolution and Large Parameter Models

Flexibility and Choice

For Deployment and Consumption



ONE PLATFORM

YOUR CHOICE OF DEPLOYMENT

DATAFLOW-AS-A-SERVICE™	DATASCALE®
SambaNova managed	Customer managed
Optimized pretrained models	Build your own models
Ability to fine-tune with your custom data	Your data
Opex	Opex + Capex

SambaNova®

SYSTEMS