

# SambaNova + DDN Ref Architecture

Building the easiest, fastest, most scalable AI solutions

## Challenge

Emerging AI work cases require a significantly different type of balance between compute, storage, and networking, and developers can no longer depend on traditional instruction set architectures to provide the flexibility and scale to accurately and efficiently analyze the massive quantities of data.

## Solution

Optimizing from algorithm to silicon, with a focus on scaling and continued flexibility of configuration, the SambaNova Reconfigurable Dataflow Architecture™ (RDA) is a new approach to deep learning computing architecture. Capable of being coupled with a tiered memory storage system from DDN, which allows for limitless scalability, the SambaNova RDA provides the compute power needed to enable the next generation of machine learning and high-performance computing applications.

## Key Benefits

- Complete platform for AI/ML workflows
- Security, capacity, and accessibility of data at petascale
- Maximum performance
- Limitless scaling
- Ease of integration into existing environments

Across a variety of AI and Machine Learning applications, there is a trend towards large memory models. These larger models are being developed for more sophisticated and complex applications, requiring higher compute capability and more memory at scale. The SambaNova Reconfigurable Dataflow Architecture™ (RDA), in conjunction with DDN® scalable memory systems, is a complete platform for AI/ML workflows with the flexibility and scalability to grow with applications as those workflows become more complex and sophisticated.

## The Challenge

As Artificial Intelligence and Machine Learning (AI/ML) applications become more capable, they are recognized as essential for businesses to compete in existing and future markets. The trend towards large memory models is creating a demand for new architectures that can accommodate these exceptionally compute intensive, storage (memory) intensive, and network intensive workloads.

The rate of advancement in traditional processing architectures (CPU fat-core and GPU thin-core) has drastically slowed while the amounts of data to process grow exponentially. Attempting to process large amounts of data on architectures that lack flexible compute power and sufficient memory results in lower efficiency and lower accuracy.

A new approach to compute architecture is necessary to meet the current performance requirements needed to support emerging AI/ML applications and large memory models: a platform that has the flexibility and capability to scale with those models and applications.

## SambaNova Reconfigurable Dataflow Architecture™ (RDA)

The SambaNova Reconfigurable Dataflow Architecture™ (RDA) is a computing architecture designed to enable the next generation of machine learning and high-performance computing applications. As a complete, full-stack solution that incorporates innovations at all layers, SambaNova RDA can be programmed specifically for each model, resulting in a highly optimized, application-specific accelerator.

The Reconfigurable Dataflow Architecture is composed of the following:

1. **SambaNova Dataflow-as-a-Service™** is a fully managed and supported extensible ML services platform. Dataflow-as-a-Service uses Natural Language Processing, High-resolution Computer Vision, and Recommendation services to analyze data and produce industry specific insights.
2. **SambaNova Systems DataScale®** is a complete, rack-level, data-center-ready accelerated computing system. Each DataScale system configuration consists of one or more DataScale nodes, integrated networking and management infrastructure in a standards-compliant data center rack, referred to as the SN10-8R.
3. **SambaNova Reconfigurable Dataflow Unit™** is a next-generation processor designed to provide native dataflow processing and programmable acceleration.
4. **SambaFlow™** is a complete software stack designed to take input from standard machine-learning frameworks such as PyTorch and TensorFlow. SambaFlow also provides an API for expert users and those who are interested in leveraging the RDA for workloads beyond machine learning.

## DataDirect Networks A<sup>3</sup>I solutions

DataDirect Networks is the world's leading data storage supplier to data-intensive, global organizations. While enterprise file storage architectures and protocols like NFS starve AI computing of data, DDN A<sup>3</sup>I speeds up applications by providing the highest level of bandwidth and lowest latency for data transfers. DDN's A<sup>3</sup>I AI-optimized Intelligent Infrastructure removes data management risks. The A<sup>3</sup>I solutions enable instant and accurate insight for customers processing massive amounts of data and consolidate all AI workflows for training, inference, and analytics into a unified, easy-to-deploy, easy-to-scale platform, with unparalleled performance and ROI at scale.



### DDN A<sup>3</sup>I provides:



Highest Bandwidth throughput with the lowest latency data transfer



Superior Performance and Scalability



Unsurpassed Reliability and Data Availability



Secure Multi-tenant environments



NFS/SMB/S3 Protocols Storage Access

## SambaNova RDA + DDN

As a complete solution, the SambaNova RDA and DDN storage provide the architecture to power large memory models and flexibly process data at scale, delivering the insights needed to achieve business value faster.

### DDN x SambaNova

Architected to work synergistically, expand to capacity




X



=

Achieve Business Value Faster



- Simplify, Accelerate AI lifecycles
- Start small, grow to scale
- Peak Performance without the complexity
- Faster Solution Development
- Streamline, Centralize Workflows and data governance

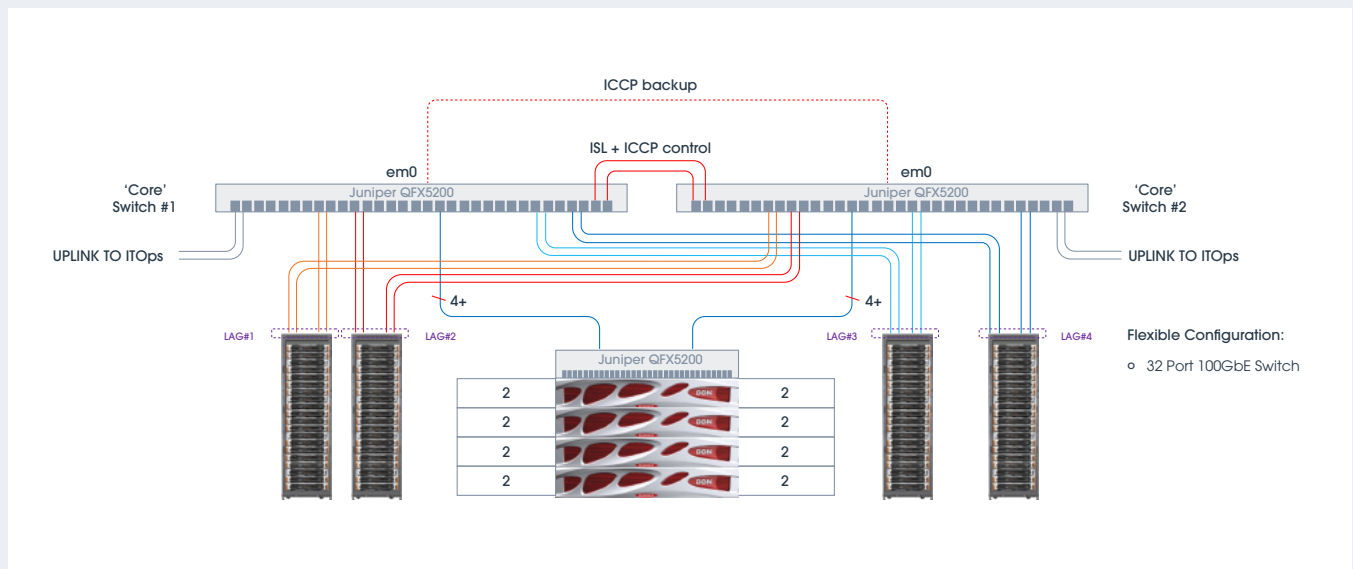
- CapEx or OpEx Models
- Single source: Software, System, Models, Service
- Models: Pre-built, fully Optimized, Pre-trained, SOTA, Converged
- Continuously updated on the latest algorithms and techniques

- Accelerate time to solution
- Reduce cost
- Increase revenue from new products and services
- Increase Operational efficiency

The complete platform architecture is designed for limitless scale and contains redundancies to ensure seamless operation.

### SambaNova + DDN Reference Network Architecture

Configure MC-LAG Active-Active with redundant 'Core' Switches





## SambaNova is Redefining AI Boundaries

To learn more about how the SambaNova Reconfigurable Dataflow Architecture, partnered with DDN, can accelerate and transform your organization with AI, [schedule a meeting](#).

Learn more at [SambaNova.AI](https://SambaNova.AI)



SambaNova Systems is an AI innovation company that empowers organizations to deploy best-in-class solutions for computer vision, natural language processing, recommendation, and AI for science with confidence. SambaNova's flagship offering, Dataflow-as-a-Service™, helps organizations rapidly deploy AI in days, unlocking new revenue and boosting operational efficiency. SambaNova's DataScale® is an integrated software and hardware system using Reconfigurable Dataflow Architecture™, along with open standards and user interfaces. Headquartered in Palo Alto, California, SambaNova Systems was founded in 2017 by industry luminaries, and hardware and software design experts from Sun/Oracle and Stanford University. Investors include SoftBank Vision Fund 2, funds and accounts managed by BlackRock, Intel Capital, GV, Walden International, Temasek, GIC, Redline Capital, Atlantic Bridge Ventures, Celesta, and several others. For more information, please visit us at [sambanova.ai](https://sambanova.ai) or contact us at [info@sambanova.ai](mailto:info@sambanova.ai). Follow SambaNova Systems on LinkedIn.