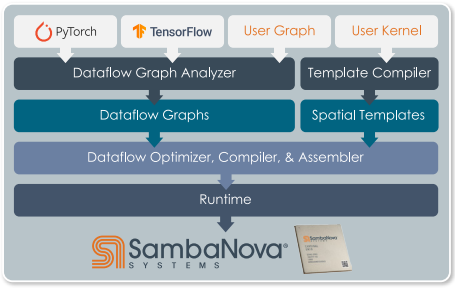
Break Through the Limits of your GPU

SambaFlow



A Software First Approach

A complete software stack for SambaNova DataScale, SambaFlow™ is designed to advance developer productivity. Built to fully integrate with popular standard frameworks such as PyTorch and TensorFlow, SambaFlow provides an open, flexible, and easy-to-use development interface for exploring the new capabilities unlocked by DataScale.

SambaFlow automatically extracts, optimizes, and executes the optimal dataflow graph of any of your models on SambaNova's Reconfigurable Dataflow Units (RDUs). This enables you to achieve out-of-the-box performance, accuracy, scale, and ease of use. With SambaFlow, you can maximize productivity by focusing on your development in the frameworks without ever again worrying about low-level tuning.

Ease of Use

SambaFlow manages the DataScale platform from frameworks to hardware.

- **Fully integrated with popular open source ML frameworks, such as PyTorch and TensorFlow.** No code required to run your existing models.
- **Automated data and model parallel mapping** simplifies scaling by using the same programming model you would use for a single device across any number of devices and configurations.
- **High-level API** provides power users with additional control without needing to understand the details of hardware implementation.
- **Native integration with developer productivity repos, such as Hugging Face.** Run thousands of models at state-of-the-art performance and accuracy in seconds with zero code change.

Out-of-the-Box Performance and Efficiency

Our vertically integrated approach combines innovations across the entire stack, eliminating the need for low-level tuning or custom kernel development.

- **Reconfigurable Dataflow Architecture** allows SambaFlow to configure the optimal dataflow graph for each model onto RDUs—effectively resulting in a custom accelerator for any model and any batch size.
- **Spatially pipelined execution** fuses operators into highly-parallelized processing pipelines arranged to meet the optimal communication patterns.
- **High-performance data transfer protocols** ensure data transfer between devices can run at full speed.

Breakthrough Accuracy, Simplified

Surpass the limits of today's AI technology with new possibilities and approaches, including:

- **Automated input image tiling,** allowing higher accuracy to be obtained directly from your high-resolution images.
- **Efficient memory allocation and optimized data movement** enable orders-of-magnitude larger embeddings or higher-fidelity datasets to be used as-is, with no compression or down-sampling necessary.
- **Sustained throughput at any batch size** provides the freedom to explore the most ideal batch size that best suits your model and dataset, including batch size 1.

Flexibility and Security

With unparalleled flexibility and built-in security features, SambaFlow is designed to serve all of your AI applications at scale.

- **Reconfiguration in microseconds** enables rapid experimentations during development and adapts seamlessly to dynamically changing deployment environments.
- **Secure multi-tenancy and concurrent multi-graph execution** provide seamless scale-up and scale-out flexibility for the best utilization of resources.
- **Virtualization and container support** ensure the secure deployment of Docker, Kubernetes, Singularity, or VM environments.

For more information visit sambanova.ai
or call +1 650 263-1153 to speak to a
SambaNova representative

sambanova.ai

Palo Alto, CA and Austin, TX (650) 263-1153