A New State of the Art in NLP: Beyond GPUs

Marshall Choy is the Vice President of Product at SambaNova Systems, responsible for product management and go-to-market.

As Natural Language Processing (NLP) models increasingly evolve into bigger models, GPU performance and capability degrades at an exponential rate. We've been talking to a number of organizations in a range of industries that need higher quality language processing but are constrained by today's solutions.

Groundbreaking Results, Validated in Our Research Labs

SambaNova has been working closely with many organizations the past few months and has established a new state of the art in NLP. This advancement in NLP deep learning is illustrated by a GPU-crushing, world record performance result achieved on SambaNova Systems' Dataflow-optimized system. We used a new method to train multi-billion parameter models that we call ONE (Optimized Neural network Execution). This result highlights orders-of-magnitude performance and efficiency improvements, achieved by using significantly fewer, more powerful systems compared to existing solutions.

Break Free of GPU Handcuffs

SambaNova Systems' Reconfigurable Dataflow Architecture™ (RDA) enables massive models that previously required 1,000+ GPUs to run on a single system, while utilizing the same programming model as on a single SambaNova Systems Reconfigurable Dataflow Unit™ (RDU). See Figure 1.

SambaNova RDA is designed to efficiently execute a broad range of applications. RDA eliminates the deficiencies caused by the instruction sets that bottleneck conventional hardware today.

Run Large Model Architectures With a Single SambaNova DataScale System

With GPU-based systems, developers have been forced to do complicated cluster programming for multiple racks of systems and to manually program data parallelism and workload orchestration.

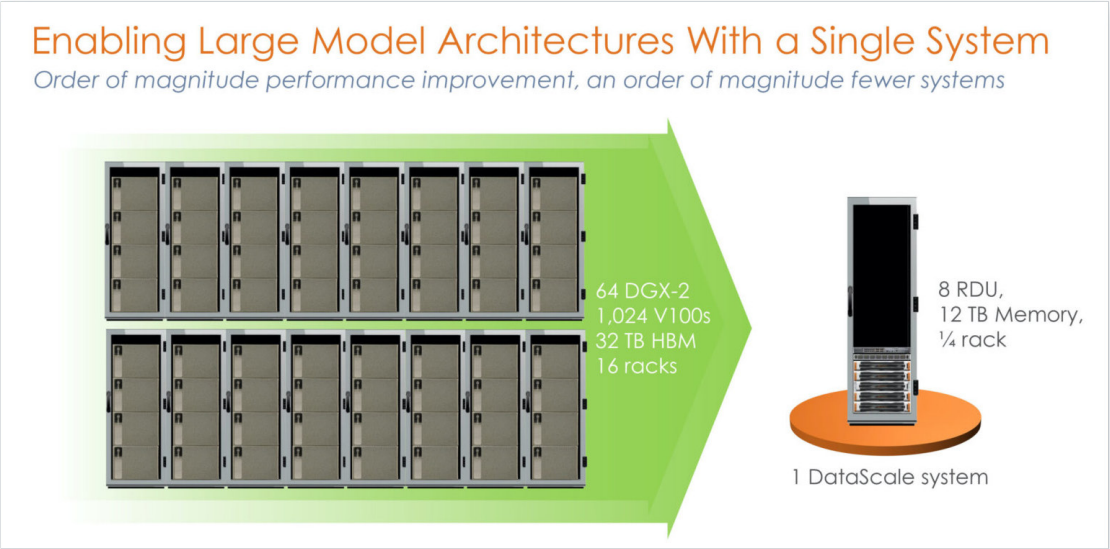

Figure 1

A single SambaNova DataScale™ System with petaflops of performance and terabytes of memory ran the 100-billion parameter ONE model with ease and efficiency, and with plenty of usable headroom. Based on our preliminary work and the results we achieved, we believe running a trillion-parameter model is quite conceivable.

The proliferation of Transformer-based NLP models continues to stress the boundaries of GPU utility. Researchers are continuing to develop bigger models, and as a result the stress fractures on GPU-based deployments are also getting bigger. By maintaining the same simple programming model from one to many RDUs, organizations of all sizes can now run big models with ease and simplicity.

The sophistication of SambaNova Systems' SambaFlow™ software stack paired with our Dataflow-optimized hardware eliminates overhead and maximizes performance to yield unprecedented results and new capabilities.

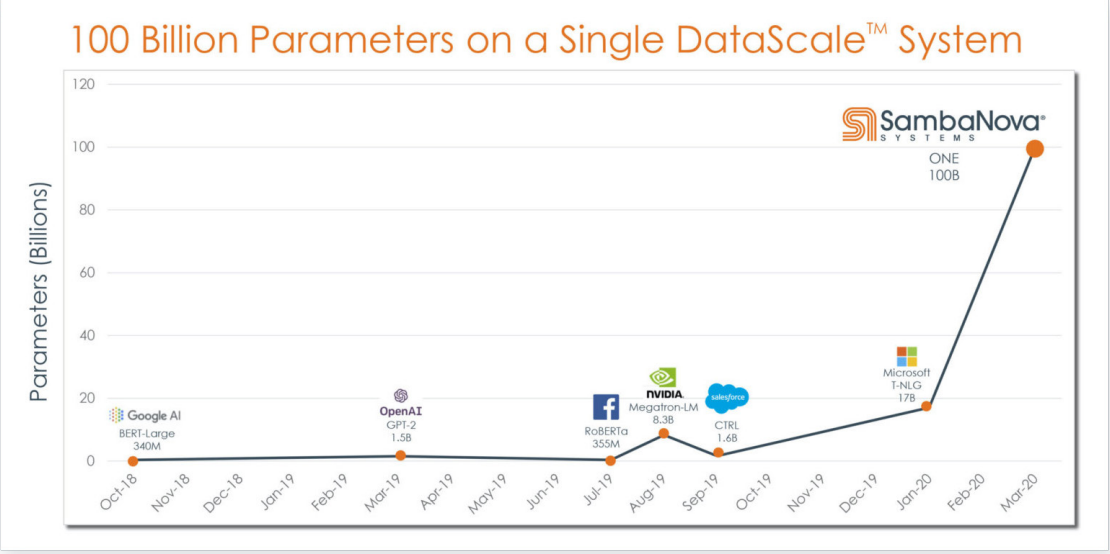

Figure 2

No Boundaries, Only New Possibilities for NLP

Three trends have emerged in NLP that are pushing infrastructure requirements far beyond the capabilities of current GPU architecture. These trends, below, highlight attributes that enhance SambaNova Systems DataScale's ability to deliver world record throughput performance and unlock capabilities that were previously unattainable.

Trends	SambaNova Systems Advantage
More tasks with richer Dataflow	Unconstrained by memory, superior model parallelism, longer sentence sizes, larger batch size 512 or higher
Richer contextual information	Elimination of high overhead kernel-by-kernel execution with spatial layout for Dataflow, highly detailed models
Large embeddings and activations	Highest internode utilization, superior model parallelism, unconstrained by memory capacity

Kunle Olukotun, one of SambaNova Systems' esteemed co-founders and the company's chief technologist, describes our systems best:

"SambaNova engineered a purpose-built Reconfigurable Dataflow Architecture that expands the horizons of capability for the future of machine learning. Users, developers, and applications are now liberated from the constraints of legacy architectures."

SambaNova provides state-of-the-art technologies to support NLP, high-resolution computer vision, and recommender models. To learn more, request a meeting.

Request a Meeting

sambanova.ai