SambaNova SYSTEMS

WHITEPAPER

AI is changing everything you know about hardware and software. Here's why

How SambaNova Systems' RDA rewrites the rules for AI and HPC

SITUATION PUBLISHING

The demands of ever more sophisticated Artificial Intelligence have pushed classical compute hardware and software to its limits. But will the architectures being put in place to enable the next generation of AI also transform how we approach traditional HPC and general computing?

That is precisely what software-defined AI hardware startup SambaNova Systems is working towards with its Reconfigurable Dataflow Architecture (RDA), which reimagines how we can free AI from the constraints of traditional software and hardware.

This extended profile explains how SambaNova Systems' founders have drawn on decades of experience at some of Silicon Valley's most storied companies and institutions to pinpoint why traditional architectures are running out of road when it comes to advancing machine learning and AI, and how this led to the development of RDA.

It also explains how the application of RDA at institutions like Lawrence Livermore National Laboratory to help crack fundamental physics problems also points to how the architecture can redefine the way we approach classical compute problems.

Businesses will need to adopt AI technologies not just because they can, but because they must – AI is the technology that will help businesses to be agile, innovate, and scale. So says the tech analyst firm IDC, which forecasts global spending on AI systems will double over the next four years, from $50.1 billion this year to more than $110 billion in 2024.

- IDC Report 2020, "Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide"

Drivers for AI adoption include "delivering a better customer experience and helping employees to get better at their jobs," says IDC. "This is reflected in the leading use cases for AI, which include automated customer service agents, sales process recommendation and automation, automated threat intelligence and prevention, and IT automation." Some of the fastest growing use cases include automated human resources and pharmaceutical research and discovery, the research firm adds.

However, the benefits of this technological revolution are spread very unevenly, according to Kunle Olukotun, co-founder and chief technologist of SambaNova Systems, the software defined AI hardware startup. "If you look at the people who are able to develop these sorts of systems, they're in the hands of the few, the large companies that have the data, the computation and the talent to develop these

sorts of algorithms, and of course they've used these systems to become the most valuable companies in the world – Google, Apple, Amazon, Facebook and the like," he says.

The fundamental challenge lies with the sheer amount of compute power needed to build and train many of the more advanced models that are being developed. The models are getting larger and larger, and for some applications the volumes of data required to train them are also ballooning. This is exacerbated by the slowing of performance gains for successive generations of processor chips, a trend that some have labelled the end of Moore's Law, according to SambaNova's vice president of product, Marshall Choy.

"Multi-core has run its course, and single cores are inefficient, so obviously, just putting many of these together on a chip just increases the inefficiency," Choy says. "So, we need a much more efficient architecture, as the platform for future AI and machine learning innovations to enable that whole new class of AI applications."

Today, AI workloads are typically being processed by racks of systems using a combination of CPUs and GPUs. The latter have an architecture designed for much greater parallelism, with hundreds of relatively simple cores optimized for floating-point throughput, which has proved much better suited than CPUs to tasks such as training machine learning.

THE AI GOLDILOCKS ZONE

However, this success up until now masks the fact that GPUs were not originally designed for machine learning, and may not be suitable for every AI workload, according to Choy.

"If you look at where the GPU performs very well, it's actually a narrow band of the overall research field in machine learning algorithms and applications. The GPU fits into this AI Goldilocks zone, in that it basically runs models really well that fit into the size of the GPU memory, and within the constraints of the architecture," he claims.

Researchers are now pushing out of the 'Goldilocks zone' towards smaller, more highly detailed models using transformers and aimed at efficiency on the one hand, and on the other towards bigger models using bigger datasets and higher parameter counts, such as BERT and GPT in natural language processing, which can result in out of memory errors or might require thousands of GPUs to deliver. These limitations matter because many organizations today are investing heavily in infrastructure that may be too inflexible to allow them to adapt to a rapidly changing economic and business environment.

Apart from the potential lack of widely available processing power to drive newer machine learning models, there are other trends that highlight the need for new approaches and new architectures in AI. The first, according to Choy, is that the processes of training and inference have traditionally been kept separate. Typically, the training of a model is performed using the brute force power of GPUs, while the inference, using the trained machine learning algorithm to make a prediction, is more often performed using an ASIC or CPU.

"What we see now in terms of real world applications is the need to converge both training and inference. Because you want to be able to do things like model fine tuning to specific use cases and enable continuous learning on small models, to enable things like transfer learning and incremental retraining on the inference node," he explains.

Doing this with different systems, moving the results back and forth from one to the other, can be expensive and incur high latency in datacenters, so moving to an architecture that can do both tasks on the same system makes more sense.

SOFTWARE 2.0

This concept of dataflow is central to SambaNova's notion of how next generation computer architectures will operate, and it is such a step change that the company believes it will usher in a new era of computing.

Dataflow computing is the end result of approaching applications the "software 2.0" way, a term coined by Andrej Karpathy that refers to the way machine learning algorithms are developed.

"Before machine learning, we had what we're now calling software 1.0, and here code is written in C++ or some other high-level language, and it requires domain expertise to decompose the problem and design algorithms for the different components and then compose them back together," says Olukotun.

"Contrast that with software 2.0, where the idea is that you train neural networks using training data, and the program is written in the weights of the neural network. This has a number of advantages, and the key one is that you have a reduced number of lines of code that have to be explicitly developed by the programmer," Olukotun explains.

As an example, Olukotun cites the Google Translate service, which Google reduced from 500,000 lines of C code to just 500 lines of dataflow code in TensorFlow, a domain-specific framework for machine learning developed by Google but widely used elsewhere.

"What we see is that if we look at the development of machine learning applications, they are done using these high-level frameworks like TensorFlow and PyTorch. And these frameworks generate a dataflow graph of machine learning operators like convolution, matrix multiply, Batch Norm and the like," Olukotun says.

These domain-specific machine learning operators can then be converted into 'parallel patterns', which express the parallelism and locality in the application, and can be optimized for higher performance.

"And what we see is that not only can these parallel patterns represent machine learning applications, they can also be used to represent the operators in SQL that are used for data processing. And these can be represented efficiently using parallel patterns," Olukotun adds.

This then is SambaNova's prescription for the new era of computing: support for hierarchical parallel pattern dataflow as the natural machine learning execution model; support for very large, terabyte-sized models that will provide much higher accuracy; support for flexible mapping of those machine learning graphs onto the underlying hardware; and the need to support data processing, specifically SQL operations, as these form a key part of machine learning training.

To make this a reality, SambaNova Systems has developed a new computing architecture called 'Reconfigurable Dataflow Architecture' (RDA), built to support Software 2.0 and bring machine learning to all types of dataflow computation problems.

Reconfigurable Dataflow derives its name from the way it has been designed around the flow of data and dataflow graphs rather than using traditional computing approaches, and because it is reconfigurable to execute any dataflow graph. SambaNova recently announced the availability of DataScale, a platform built around RDA.

HOUSTON, WE HAVE A DATAFLOW PROBLEM

The dataflow process is familiar to anyone who knows how AI and machine learning operate, but

SambaNova argues that the architecture it has developed for accelerating these dataflow graphs is applicable to problems beyond machine learning, including many of the applications seen in HPC. This is because many of these HPC applications are also reducible to dataflow problems, and also happen to involve very large datasets, as machine learning does.

However, traditional compute architectures are not optimized for these use cases. Both CPUs and GPUs read the data and weights from memory, calculations are performed, and the output results are written back to memory. The process has to be repeated again for each stage in the dataflow graph, meaning that huge amounts of memory bandwidth are needed to keep moving the data back and forth.

So, if you have a dataflow problem, you need a system designed around dataflow to solve it, a system built from the ground up as an integrated software and hardware platform. This is where SambaNova's Reconfigurable Dataflow Architecture comes in.

"Reconfigurable Dataflow is a ground-up re-imagining of how to do compute, focusing on what machine learning algorithms need," Olukotun says. "It is a sea of computation and memory closely coupled together by a programmable network. And the key to that network is that you can program the dataflow between the different compute and memory elements in a way that matches the needs of the application that you're trying to run."

CARDINAL RULES

This can be seen in the Cardinal SN10, the processor chip designed by the company, which calls it a Reconfigurable Dataflow Unit or RDU, to distinguish it from a CPU or GPU. The Cardinal SN10 consists of a grid of configurable elements, pattern compute units (PCUs) and pattern memory units (PMUs) that is tiled across the chip and linked together by a flexible on-chip communication fabric.

To implement an algorithm or workload, the functional components of the algorithm are mapped onto the compute and memory units, and the dataflow between them is implemented by configuring the communication fabric. This ensures that the flow of data seen in that algorithm or neural network is elegantly mirrored in the configuration of the chip elements.

SambaNova's on-chip elements should not be compared with CPU or GPU cores, said Choy. "You should think of this chip as a tiled architecture, with reconfigurable SIMD pipelines, but I think the more interesting piece is actually the memory units on the chip, where we've got an array of SRAM banks and

address interleaving and partitioning, so a lot of where the magic in the chip happens is really on the memory side."

There are several hundred of each type of element (PCUs and PMUs), with the combined memory adding up to "hundreds of megabytes" of on-chip memory, which Choy compares favorably with GPUs.

The result is a big reduction in the off-chip bandwidth requirements, resulting in very high utilization of the compute and allowing it to hit teraflops of machine learning performance, instead of wasting compute power in shuffling data back and forth as seen in traditional architectures, Choy said.

The smallest DataScale system (SN10-8) takes up a quarter rack of space and comprises eight of the Cardinal SN10 RDUs and 3TB, 6TB, or 12TB memory. The tiled architecture of RDU means that you can securely run multi-tenant, multiple high-performance mixed workloads or you can simply run one large application across all RDUs on the entire DataScale system.

SambaFlow is equally important in the DataScale system. The software integrates with popular machine learning frameworks such as PyTorch and TensorFlow and optimizes the dataflow graph for each model onto the RDUs for execution.

"Taking the example of a PyTorch graph we have a graph analyzer, that then pulls out the common parallel patterns out of that graph, and then through a series of intermediate representation layers, our data flow optimizer, compiler and assembler build up the runtime, and we execute that onto the chip," Choy explains.

The preparation work is handled by an x86 subsystem running Linux, but it can be bypassed using SambaNova's RDU Direct technology, which enables a direct connection to the RDU modules.

LAWRENCE LIVERMORE

This is how the SambaNova DataScale platform has been deployed at one early customer, the Lawrence Livermore National Laboratory (LLNL). Here, the Corona supercomputing cluster, which boasts in excess of 11 petaflops of peak performance, has been integrated with a SambaNova DataScale SN10-8R system.

LLNL researchers will use the platform in cognitive simulation approaches that involve the combination

of high performance computing and AI. The SambaNova DataScale's ability to run dozens of inference models at once while performing scientific calculations on Corona will help in the quest of using machine learning to improve nuclear fusion research efforts, LLNL said in a press statement. Researchers have already reported that the DataScale system demonstrates a five times improvement over a comparable GPU running the same models. SambaNova said a single DataScale SN10-8R can train terabyte-sized models, which would otherwise need eight racks worth of Nvidia DGX A100 systems based on GPUs.

This convergence of HPC and AI supports SambaNova's assertion that its architecture based on dataflow does not merely accelerate AI functions but represents the next generation of computing. (That said, the company does not expect this new breed of computing to replace CPU-based systems for more transaction-oriented applications.)

"It's not just for the high-end problems of language translation and image recognition that this software 2.0 approach works, it also works for classical problems," said Olukotun.

"It turns out that many classical problems from data cleaning to networking to databases have a bunch of heuristics in them, and what you find is you are better off replacing those heuristics with models based on data because you get both higher accuracy and better performance."

He cites the example of the parallel patterns that represent machine learning applications and said these can equally well be used to represent the operators in SQL that are used for data processing.

ADVANCING AI FOR EVERYONE

Meanwhile, SambaNova is keen to show that its technology is available to customers of all sizes and not just for high-end research labs or for huge companies.

For example, the company has introduced a Dataflow-as-a-Service offering that makes DataScale systems available via a monthly subscription service. This is managed by SambaNova and provides a risk-free way for organizations to jump start their AI projects using an Opex model.

AI may seem like an arcane topic to many, but machine learning is a significant and growing part of computing across a broad spectrum of everyday applications. Anything that can make AI more accessible and more efficient is vitally needed to drive future advances, and SambaNova argues its

system-wide approach of a complete platform of software and hardware focused on dataflow is the answer.

In many ways, this broader application of AI is what SambaNova Systems' founders have been working towards through their entire careers. The start-up was founded in 2017 by a group of far-sighted engineers and data scientists who saw that the current approaches to AI and machine learning were beginning to run out of steam, and that an entire new architecture would be necessary in order to make AI accessible for everyone as well as deliver the scale, performance, accuracy and ease of use needed for future applications.

AI and machine learning in particular have grown over the past decade to become key tools for processing and making sense of large and complex data sets.

This trend is set to continue, with IDC forecasting that the overall AI software market will approach $240 billion in revenue in 2024, up from $156 billion in 2020.

- IDC Report 2020, "Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide"

But AI is no longer just for supercomputing. The volumes of data collected by organizations have become large and complex data sets, leading to machine learning being incorporated into all manner of applications from natural language processing (NLP), high-resolution computer vision, recommendations and high performance computing (HPC) to everyday business processes.

As we've seen, the success of this approach has led to machine learning models becoming larger, partly in order to increase accuracy. This has required ever more compute power. If this trend were to continue, it could hamper the development of more advanced machine learning models and mean that advanced AI would soon be out of the reach of all but the largest organizations.

The limitations of existing architectures when it comes to handling AI is something that SambaNova's founders were likely more aware of than many others in the IT industry. Kunle Olukotun, the company's co-founder and Chief Technologist, carried out some pioneering work on multi-core processor architectures as a Professor at Stanford University, and helped found a company called Afara Websystems to bring the technology to market.

Afara was acquired by Sun Microsystems, where its technology laid the groundwork for Sun's 'Niagara' Sparc T1 family of processors, and Olukotun returned to Stanford where he started looking at how to develop software to make full use of the capabilities of multi-core processors.

STEPPING BACK TO LEAP FORWARD

This led to the concept of domain-specific languages, an idea now widely adopted for performing machine learning tasks, and this eventually fed into the concepts that SambaNova was formed to commercialize – ways of developing hardware and software that will make AI technology much more accessible.

SambaNova's co-founder and CEO Rodrigo Liang also worked for Afara and stayed on after the Sun acquisition until 2017 to oversee SPARC processor development. His combination of business and technical experience made him the logical choice for CEO when SambaNova was formed to develop a platform designed from the ground up for machine learning and analytics.

The third co-founder Christopher Ré, a 2015 MacArthur "genius grant" winner, also works at Stanford University, where he is an associate professor in the Department of Computer Science affiliated with the Statistical Machine Learning Group and Stanford AI Lab. Based on his research into machine learning systems, Ré co-founded Lattice.io, a data mining and machine learning startup that was acquired by Apple in 2017. Subsequently, he helped to found SambaNova Systems, based in part on his work on accelerating machine learning.

The founders' backgrounds and expertise in hardware, software and chip design, scale-out architecture and machine learning enabled them to take a step back from familiar existing compute architectures. Starting from scratch, they have built an integrated system of software and hardware focused on the data processing needs of current and emerging applications.

Their proposition certainly seems to have impressed investors. By 2018, SambaNova had $56 million in Series A funding led by Walden International and Google Ventures, with participation from Redline Capital and Atlantic Bridge Ventures.

This was followed in 2019 with a Series B funding round of $150 million, this time led by Intel Capital, with additional participation from Google Ventures, Walden International, Atlantic Bridge Ventures, and Redline Capital. A Series C funding round followed in 2020, providing the firm with $250 million led by funds and accounts managed by BlackRock with participation from existing investors.

SambaNova's technology is largely based on research undertaken by Olukotun and Ré, which focused on workflow, and specifically the flow of data, rather than on the iterative instructions seen in traditional processors.

We've already shown how this dataflow-led approach informed SambaNova's development of the Reconfigurable Dataflow Unit (RDU), a processor which features a tiled pattern of reconfigurable memory and compute units, linked by a programmable communication fabric, which can be programmed to represent the dataflow of parallel patterns.

SCALING OUT FOR THE FUTURE

But with the complete SambaNova DataScale platform, which is offered both 'as-a-service' and as an on-premises solution, the software is an equally important piece of the puzzle. SambaFlow is a complete software stack that takes input from standard machine learning frameworks such as PyTorch and TensorFlow, and largely automates the compilation, optimization and execution of the models onto all the RDUs in the system.

This approach is already showing promise for efficiently processing complex machine learning problems, including 100 billion parameter models, and scales easily to handle terabytes of training data or process multiple models simultaneously, while still utilizing the same programming model as would run on a single RDU.

There is reason to believe that language models are growing by a factor of 10 every year, and SambaNova even claims that its preliminary work and the results so far achieved demonstrate that running a trillion-parameter model is quite conceivable. Such headroom is needed, according to the firm, as trends for richer context and larger embeddings in natural language processing in particular are set to push infrastructure requirements beyond current limits.

This dataflow approach to processing workloads also has broader general applicability beyond machine learning, according to SambaNova, since parallel patterns can be used to represent the operators in SQL that are used for tasks such as data preparation and data analytics.

According to Bronis de Supinski, Chief Technology Officer at LLNL, the DataScale platform it has integrated with its Corona supercomputing cluster is being used to explore a technique the scientists call cognitive simulation, whereby machine learning is used to accelerate processing of portions of the simulations.

This work being pioneered at LLNL is likely to benefit numerous industries in future that also run physics simulations as part of their operations, such as oil and gas exploration, aircraft manufacturing and engineering.

In fact, machine learning looks set to play a greater part in almost all aspects of the computer industry in future. Arti Garg, Head of Advanced AI Solutions & Technology at HPE, which counts SambaNova as a strategic partner, states that we are on the precipice of seeing much broader adoption, and this evolution is going to mean that AI impacts a lot more people than it currently does, and it will change expectations around for what AI technologies are able to do.

As SambaNova's Liang states, "We are at the cusp of a fairly large shift in the computer industry. It's been driven by AI, but at a macro level, over the next 20-30 years, the change is going to be bigger than AI and machine learning."

ABOUT SAMBANOVA SYSTEMS

SambaNova Systems is building the industry's most advanced systems platform to run AI applications from the data center to the cloud and to the edge. Founded in November 2017 by industry luminaries, hardware and software design experts, world-class innovators from Sun/Oracle, and Stanford University, SambaNova Systems' mission is to bring AI innovations that have been developed in advanced research to organizations around the world, helping to create AI for everyone, everywhere. Headquartered in Palo Alto, Calif., the company's investors include funds and accounts managed by BlackRock, Walden International, GV, Intel Capital, Redline Capital, Atlantic Bridge Ventures, WRVI Capital and several others. For more information please visit us at sambanova.ai or contact us directly.