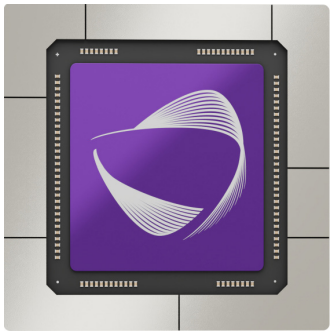
SambaNova SN40L Reconfigurable Dataflow Unit

The innovative chip that powers the fastest and most efficient inference platform

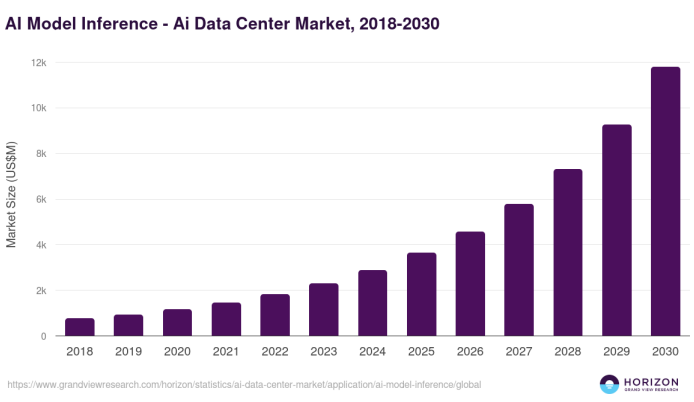


The fourth generation SambaNova Reconfigurable Dataflow Unit (RDU), the SN40L, is the heart of the SambaRack system. Purpose built to power the largest and most accurate generative and agentic AI workloads, the SN40L RDU enables the SambaNova platform to deliver extraordinary inference performance on the best and latest open-source AI models, in a very compact footprint, with ultra-low power consumption, and no need for costly and exotic cooling infrastructure.

Complementing GPUs - The need for high speed inference

AI workloads are experiencing a transition from model training to inference. As organizations move from experimenting with AI workloads to putting them in production, the need for high speed inference has become more crucial than ever.

GPUs have been the driving force of AI innovation for years. Their ability to perform parallel processing with extreme performance is ideal for AI model training and they will continue to be used for this. This has allowed GPUs to drive the development of ever larger AI models and GPU providers have built an extensive network of developers who will continue to train and fine-tune ever larger models with increasing accuracy.



The global AI inference market is predicted to grow by a CAGR of over 26% through 2030

However, organizations are transitioning from predominantly focusing on model training (teaching a model about topics) to inference (using the model to perform functions). GPUs are extraordinarily well suited to model training, which is a data processing challenge. AI inference however, is a data movement challenge, which requires a different approach.

As the industry continues to evolve, GPUs will likely remain the technology of choice for the training of new models and fine-tuning of existing ones. This will enable organizations to continue to take advantage of their existing GPU infrastructure. RDUs will then be deployed to serve those models in production, delivering the performance that users require.

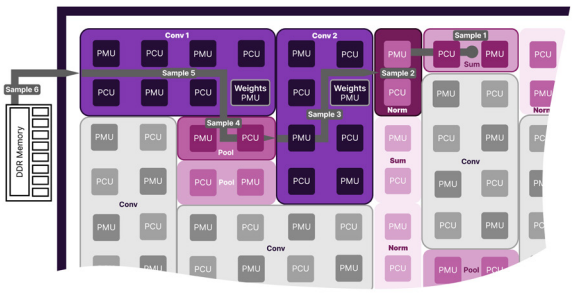
The SN40L - Designed for Inference

The SambaNova RDU was built specifically to deliver fast and efficient inference performance on both large individual models and large numbers of models, with very low power consumption and a compact footprint. Two primary innovations that enable this are its dataflow architecture, which the RDU is based upon, and a large, three-tier memory design.

Dataflow Architecture

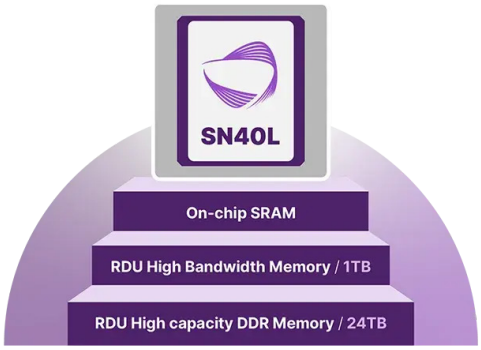
Making it ideal for inference workloads, the SN40L is built with a dataflow architecture. GPUs utilize an architecture that was not designed for AI inference and requires multiple kernel calls to run AI models. Essentially what this means is that when running AI models, GPUs have to make multiple, redundant calls back to memory. This needless overhead adds latency and slows the process down. In comparison, the RDU utilizes a dataflow architecture that combines multiple operations into a single kernel call which can handle all of the compute operations. In effect, data flows from one step of the process to the next, without calling back to memory. The result is the elimination of the additional overhead incurred from launching multiple kernels, significantly accelerating AI model processing.

Further, GPUs have low data locality, meaning that data is not always stored near where it is processed. This leads to inefficiencies as the system spends time moving data around. RDUs, with a three tier memory design, have high data locality which reduces the time spent moving data and increases efficiency.



Three-Tier Memory

The SN40L has a three-tier memory design that includes very fast memory (SRAM), high bandwidth memory (HBM), and very large memory (DRAM). The way AI models are processed is that when a user enters a prompt, the entire model is loaded onto active memory, then every possible outcome is calculated. The most likely correct response, based on its training data, is then served up to the user. The model is then removed from active memory and upon the next user prompt the process starts over again. With other systems, the model is stored in off-chip memory and then loaded into active memory, incurring needless latency. The SN40L has DRAM to hold large numbers of different models. The HBM enables the movement of a given model from DRAM to SRAM very quickly, so the time to load and run a model is drastically reduced. Since models are held in memory, switching between them can be done almost instantly.



The inherent efficiency of the SN40L architecture means that less power is required to perform inference than GPUs. The SambaRack system, which is the base system of all SambaNova solutions, consumes an average of only 10kW of power and is air cooled. This is in contrast to the latest GPU-based systems which require as much as 140kW and expensive liquid cooling infrastructure.

Conclusion

GPUs will continue to be the technology of choice for AI model training and fine-tuning as they are well suited to that workload. As AI transitions from training to inference, RDUs will become the processor of choice for those workloads. The SambaNova SN40L RDU is the right choice for fast, efficient AI inference.

SambaNova is the leading purpose-built AI system for generative and agentic AI implementations, from chips to models that gives enterprises full control over their model and private data. We take the best models, optimize them for fast tokens and higher batch sizes, the largest inputs and enable customizations to deliver value with simplicity.

090725-01