Going Beyond the State-of-the-Art in Recommendation Models

Recommender systems are a ubiquitous part of many common and broadly used internet and consumer services across industries, spanning retail and e-commerce applications that cross-sell and up-sell products and services to online consumer services such as ridesharing or peer reviews. They are also found in banking, insurance, healthcare, and other industries to deliver fast and efficient customer recommendations and experiences.

Everyday examples of recommender systems offering users hit or miss advice on news feeds, social media posts, streaming media suggestions, travel and hotels, and highest engagement ads are abundant. And for good reason, small improvements in conversion percentages can equate to large dollars. Furthermore, a company's ability to provide richer, more meaningful recommendations requires many more attributes to be incorporated into a recommendation system beyond just a user's browsing or purchase history.

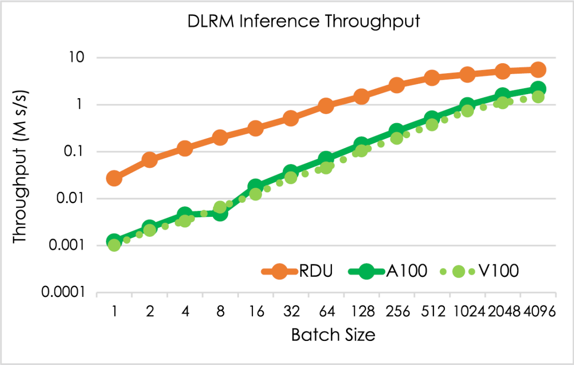
This seems simple and intuitive enough. However, real-world implementations with legacy technology components can diminish efforts to achieve state-of-the-art accuracy, which is critical to enhancing business outcomes. This is true of both key phases of implementing a recommendation system: training; and inference. However, SambaNova's DataScale architecture offers the possibility of a single platform that offers significant benefits in both stages.

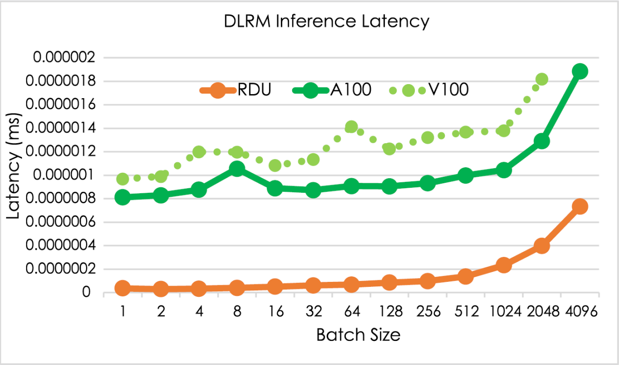
Inference for recommender systems is one of the most widespread machine learning workloads in the world. However, traditional architectures are reaching their limits.

To break through current limitations, SambaNova Systems demonstrated that by using the SambaNova DataScale system, the company can perform recommendation inference at a level that enables new capabilities and business opportunities over 20x faster than the leading GPU on an industry-standard benchmark model. This type of improvement in increased throughput while reducing latency provides significant user engagement improvement to enable faster revenue generation.

The implication of this is massive, both from technology and business perspectives. According to Facebook, 79% of AI inference cycles in their production data centers are devoted to recommendation (source). These engines serve as the primary drivers for user engagement and profit across numerous other Fortune 100 companies, with 35% of Amazon purchases and 75% of watched Netflix shows coming from recommendations (source).

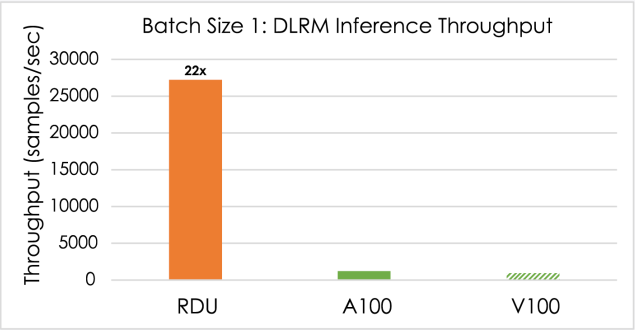
RECORD-BREAKING RECOMMENDATION SPEED

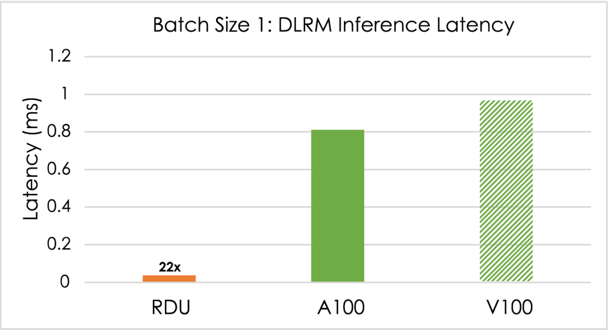




To measure the performance of the SambaNova DataScale system, SambaNova used the recommendation model from MLPerf, the authoritative benchmark for machine learning researchers and practitioners. The task for measuring recommendation performance uses the DLRM model on the Terabyte Clickthrough dataset. For A100 numbers, since Nvidia has not reported such numbers, SambaNova measured an Nvidia optimized version of this model (source) running on a single A100 deployed using Triton Server (version 20.06) using FP16 precision. SambaNova ran this at a variety of batch sizes as this simulates a range of realistic deployed inference scenarios. For V100 numbers, SambaNova used the FP16 performance results reported from Nvidia (source).

Low batch sizes are often needed in deployment scenarios where queries are streamed in real time and latency is critical. At these low batch sizes, the benefit of the dataflow architecture is clear and the SambaNova DataScale system commands 20x faster performance than a single A100 at batch size 1.





While online inference at batch size 1 is a common use case in deployed systems, customers also sometimes want to batch some of their data to improve the overall throughput of the system. To demonstrate the benefits of the SambaNova DataScale system, SambaNova also showed the same DLRM benchmark at a batch size of 4k. At this higher batch size, the DataScale achieved over 2x faster performance than an A100 for both throughput and latency.

THE COMBINED SOLUTION: TRAINING AND INFERENCE TOGETHER

While many of these measurements are geared towards MLPerf's inference task, the DataScale system excels at both inference and training. By retraining the same DLRM model from scratch, and exploring variations which aren't possible at all on GPU hardware, SambaNova Systems' Reconfigurable Dataflow Unit handily exceeds state-of-the-art. Read this article to learn more.

To meet the demands of global organizations, SambaNova Systems is now offering several Dataflow-as-a-Service (DaaS) subscriptions to support recommendation, natural language processing and high-resolution vision workloads. These 'quick start' op-ex subscriptions enable organizations to rapidly build AI solutions and scale on-demand, all within an easy to deploy framework that is managed and paid for on a cloud consumption model. Additional benefits of these 'as-a-service' offerings include the ability to augment in-house ML expertise with more power to perform, a complete solution with added vendor expertise that can accelerate your transition to AI, while seamlessly staying current with the latest R&D directions as SambaNova continuously updates DaaS with the latest models and algorithmic techniques.

BEYOND THE BENCHMARK: RECOMMENDATION MODELS IN PRODUCTION

The MLPerf DLRM benchmark simulates a realistic recommendation task, but it cannot capture the scale of a real deployed workload. In an analysis of recommendation systems, Facebook writes that "production-scale recommendation models have orders of magnitude more embeddings" compared to benchmarks (source). As models grow, CPUs and GPUs start to falter. Yet the SambaNova DataScale system has no problem handling these larger compute and memory requirements, and continues to be a long-term solution that's built to scale.

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.