Surpassing State-of-the-Art Accuracy in Recommendation Models

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

Recommender systems are a ubiquitous part of many common and broadly used internet services. They are utilized in retail and e-commerce applications to cross-sell and up-sell products and services. Online consumer services for ridesharing, peer reviews, and banking services rely heavily on recommendation models to deliver fast and efficient customer experiences. Everyday examples of recommender systems offering users hit or miss advice on social media, news sites, etc. are abundant. This is because 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.

Recommendation Tasks Place Huge Demands on Both Memory and Computation

The backbone that enables recommendation models to encode such massive volumes of data is the embedding. Embedding tables are large numerical tables that contain encodings of every feature in the data – every user, product, region, etc. It's well known that larger embedding tables lead to better model quality by making them more expressive and accurate. In order to fully capture all of the information in their data, SambaNova's industry partners easily utilize embeddings that are hundreds of gigabytes in size—often terabytes!

These embeddings are attached to deep neural networks which perform a large number of calculations in order to generate the final recommendation result.

The Benchmark

As a demonstration, we used the Sambanova DataScale system, which is a complete integrated software and hardware system, to train the Deep Learning Recommendation Model (DLRM) on the Criteo Terabyte Clicklogs dataset. This is the MLPerf standard benchmark for recommendation, where the performance metric is AUC on a test set.

NOTE: Despite containing ~1TB in data and ~100GB in embedding features, it's important to note that this dataset still does not represent a real large-scale production workload. Deployed systems are at least 5x more demanding in terms of both data and embedding sizes. But rest-assured—SambaNova Systems Reconfigurable Dataflow Unit (RDU) and the SambaNova DataScale system are built to scale and are well-equipped to tackle those gigantic use cases too.

Unleashing the Power of Embeddings

It's known that increasing embedding dimensions improves recommender model accuracy at the cost of model size. Many recent studies have been devoted to sharding the model or reducing the embedding dimensions to fit in GPU memory. SambaNova Systems researchers have pioneered superior methods for solving this problem via vertical engineering through our integrated software and hardware stack. We demonstrate this by exceeding state-of-the-art accuracy on the DLRM model by significantly increasing the embedding dimensionality. In an ablation study where everything else is held constant, we find that the model's accuracy strictly increases with embedding dimensions when trained on a single SambaNova Systems RDU. Meanwhile, on a single GPU, model execution attempts result in catastrophic failure.

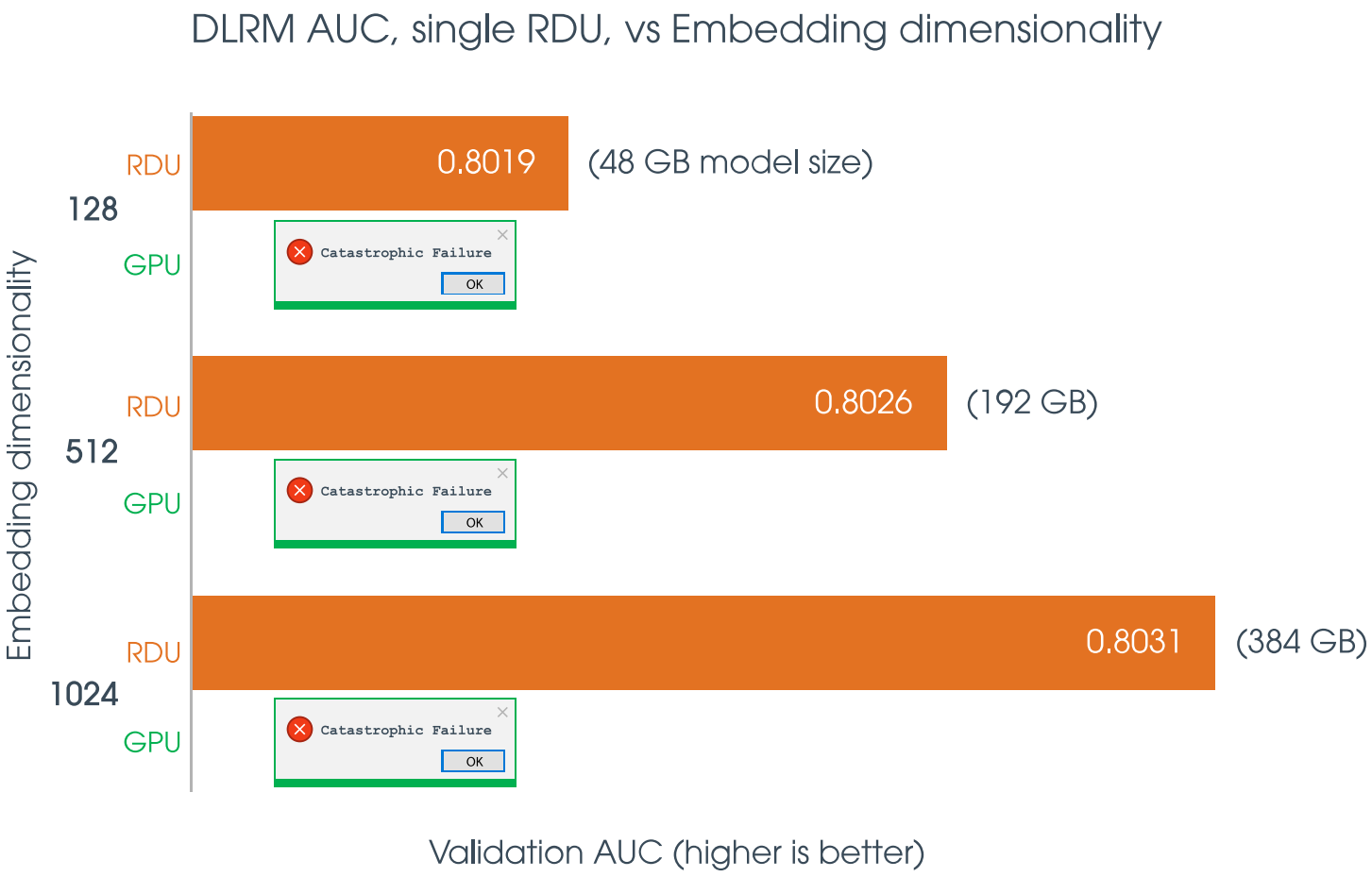

Fig 1: Effects of Embeddings dimensions on single RDU and single GPU

Exploring New Batch Sizes and Breaking the GPU Mold

Popular training techniques place a large focus on increasing mini-batch size to saturate GPU computation. For example, Nvidia's demo implementation of DLRM uses batch sizes of 32768 and higher.

From a statistical standpoint, this isn't always the preferred decision. As studied, decreasing the batch size can actually have strong benefits, helping a model avoid sharp minima so it can generalize more effectively. When training DLRM on the SambaNova Systems RDU, we observed noticeable improvements in validation performance when decreasing the batch size.

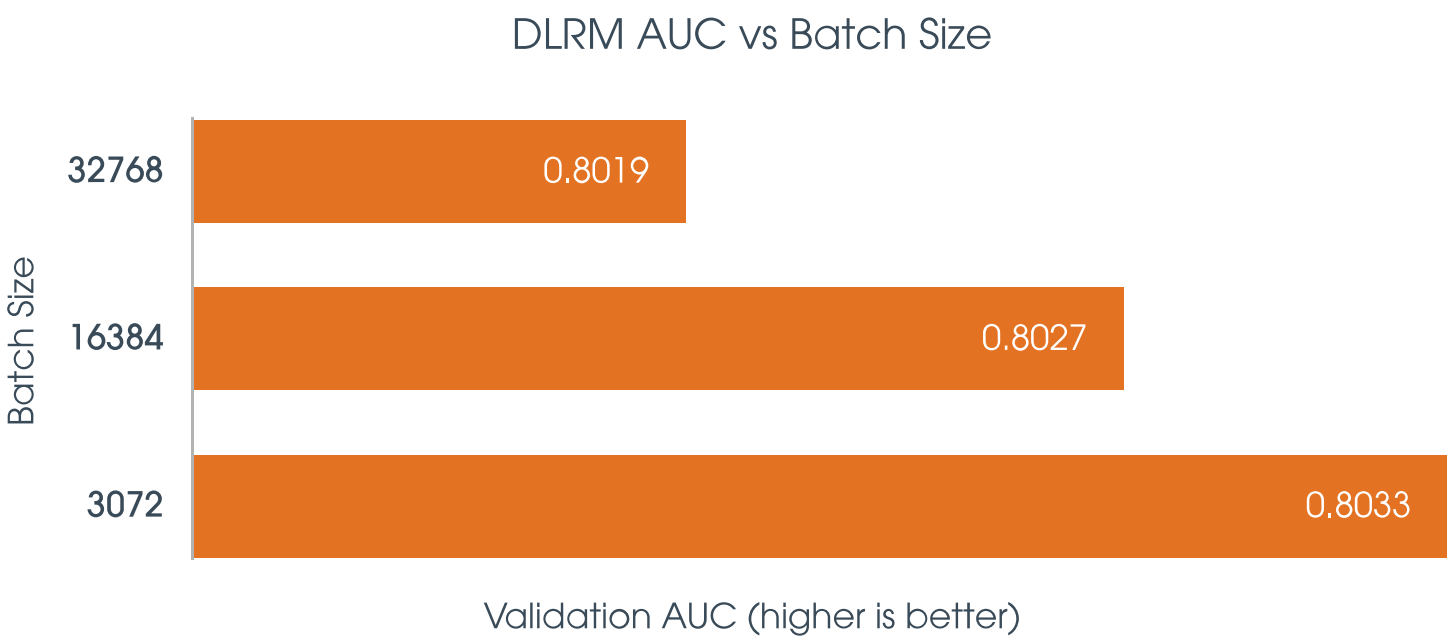

Fig 2: Enhanced RDU performance with batch size reduction

In reality, machine learning researchers and engineers choose these giant, suboptimal batch sizes because their current infrastructure leaves them no alternative. The GPU's kernel-oriented execution suffers significantly when batch size decreases. On the other hand, with the SambaNova Systems RDU's Dataflow architecture and intelligent software stack, system resources can still be fully utilized and achieve strong throughput regardless of batch size.

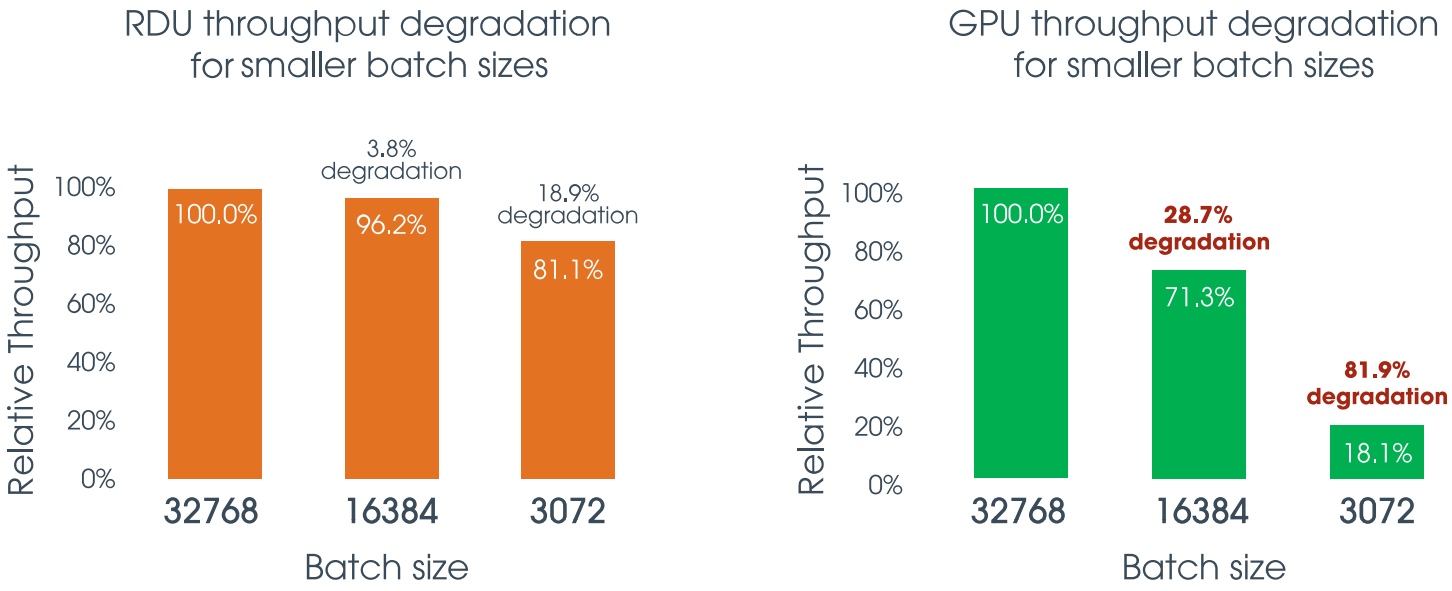

Fig 3: Negligible throughput degradation on RDU compared to GPU with smaller batch size

A New State of the Art

By combining our findings from above, we can use the SambaNova Systems RDU to train a new variant of DLRM that achieves a validation AUC of **0.8046** on the Criteo Terabyte dataset. In comparison, the best AUC reported by NVIDIA in their MLPerf submission is 0.8027. This unique large-embedding, small-batch model would be impossible to run on a GPU, and impractical to run on a CPU.

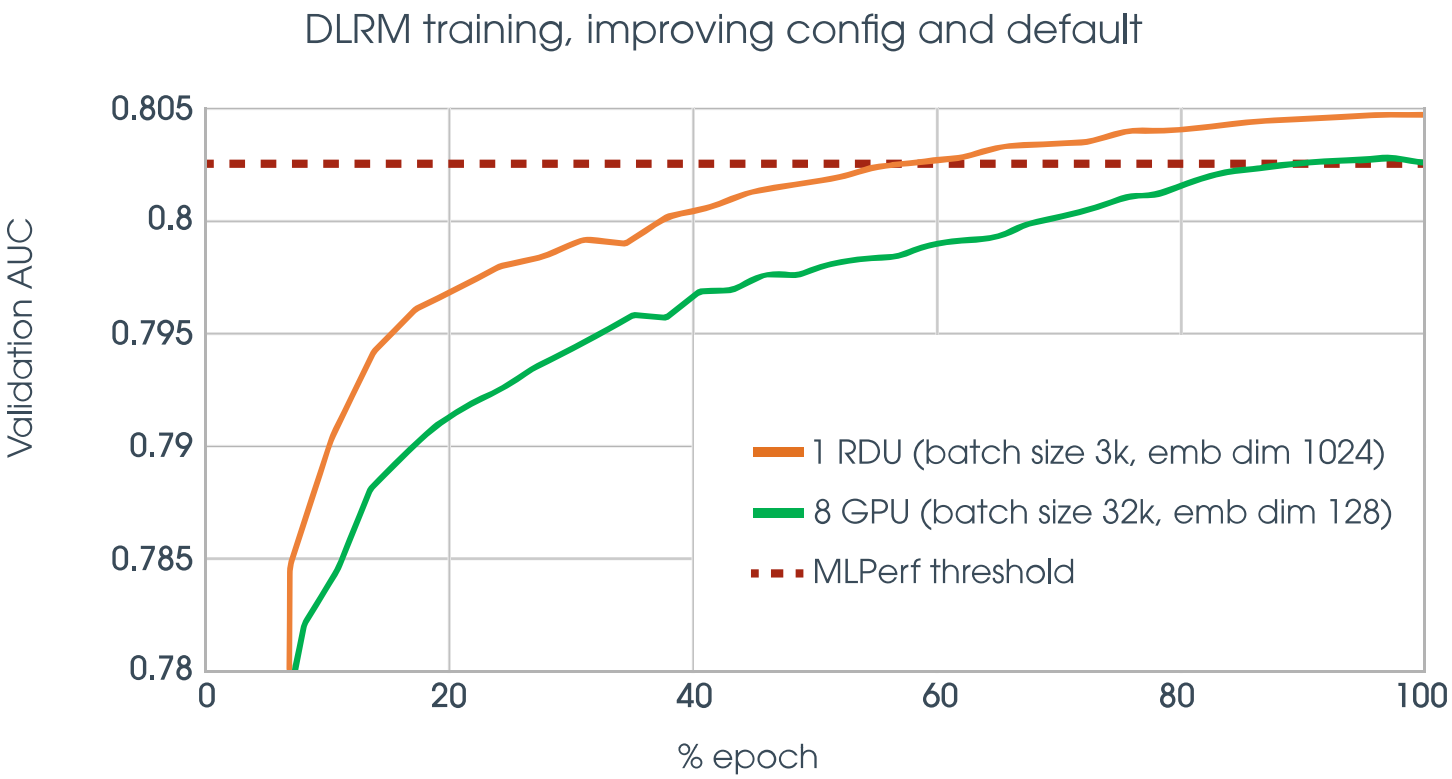

Fig 4: RDU exceeds MLPerf and GPU thresholds when training a new DLRM variant

In addition to having a noticeably higher peak AUC, the new and improved DLRM also converges much faster.

Powering Next Generation of Recommender Models

The SambaNova Systems robust yet performant RDU technology enables machine learning engineers to explore an entirely new world of models, unlocking results that surpass current state of the art. When applied to business-critical recommender models, this leads to significant enhancements in business outcomes and huge boosts in revenue. In Tencent's words, "The reason we care about small amount AUC increase is that in several real-world applications we run internally, even 0.1% increase in AUC will have a 5x amplification (0.5% increase) when transferred to final CTR".