# NVIDIA A100 Aces Throughput, Latency Results in Key Inference Benchmark for Financial Services Industry

A100 GPUs running on Supermicro servers deliver top throughput and leading latency in STAC-ML Markets standard.

Author: Malcolm deMayo

NVIDIA A100 Tensor Core GPUs running on Supermicro servers have captured leading results for inference in the latest STAC-ML Markets benchmark, a key technology performance gauge for the financial services industry.

The results show NVIDIA demonstrating unrivaled throughput — serving up thousands of inferences per second on the most demanding models — and top latency on the latest STAC-ML inference standard.

The results are closely followed by financial institutions, three-quarters of which rely on machine learning, deep learning or high performance computing, according to a recent survey .

The STAC-ML inference benchmark is designed to measure the latency of long short-term memory (LSTM) model inference — the time from receiving new input data until the model output is computed. LSTM is a key model approach used to discover financial time-series data like asset prices.

The benchmark includes three LSTM models of increasing complexity. NVIDIA A100 GPUs, running in a Supermicro Ultra SuperServer, demonstrated low latencies in the 99th percentile.

Considering the A100 performance on STAC-ML for inference — in addition to its record-setting performance in the STAC-A2 benchmark for option price discovery and the STAC-A3 benchmark for model backtesting — provides a glimpse at how NVIDIA AI computing can accelerate a pipeline of modern trading environments.

It also shows A100 GPUs deliver leading performance and workload versatility for financial institutions.

Predictable performance is crucial for low-latency environments in finance, as extreme outliers can cause substantial losses during fast market moves.

Notably, there were no large outliers in NVIDIA’s latency, as the maximum latency was no more than 2.3x the median latency across all LSTMs and the number of model instances, ranging up to 32 concurrent instances. 1

NVIDIA is the first to submit performance results for what’s known as the Tacana Suite of the benchmark. Tacana is for inference performed on a sliding window, where a new timestep is added and the oldest removed for each inference operation. This is helpful for high-frequency trading, where inference needs to be performed on every market data update.

A second suite, Sumaco, performs inference on an entirely new set of data, which reflects the use case where an event prompts inference based on recent history.

NVIDIA also submitted a throughput-optimized configuration on the same hardware for the Sumaco Suite in FP16 precision. 2

On the least complex LSTM in the benchmark, A100 GPUs on Supermicro servers helped serve up more than 1.7 million inferences per second. 3

For the most complex LSTM, these systems handled as many as 12,800 inferences per second. 4

NVIDIA GPUs offer multiple advantages that lower the total cost of ownership for electronic trading stacks.

For one, NVIDIA AI provides a single platform for training and inference. Whether developing, backtesting or deploying an AI model, NVIDIA AI delivers leading performance — and developers don't need to learn different programming languages and frameworks for research and trading.

Moreover, the NVIDIA CUDA programming model enables development, optimization and deployment of applications across GPU-accelerated embedded systems, desktop workstations, enterprise data centers, cloud-based platforms and HPC supercomputers.

The financial services industry stands to benefit from not only data throughput advances but also improved operational efficiencies.

Reduced energy and square footage usage for systems in data centers can make a big difference in operating expenses. That's especially pressing as IT organizations make the case for budgetary outlays to cover new high-performance systems.

On the most demanding LSTM model, NVIDIA A100 exceeded 17,700 inferences per second per kilowatt while consuming 722 watts, offering leading energy efficiency. 5

The benchmark results confirm that NVIDIA GPUs are unrivaled in terms of throughput and energy efficiency for workloads like backtesting and simulation.

Learn about NVIDIA delivering smarter, more secure financial services .

[1] SUT ID NVDA221118b , max of STAC-ML.Markets.Inf.T.LSTM_A.2.LAT.v1

[2] SUT ID NVDA221118a

[3] STAC-ML.Markets.Inf.S.LSTM_A.4.TPUT.v1

[4] STAC-ML.Markets.Inf.S.LSTM_C.[1,2,4].TPUT.v1

[5] SUT ID NVDA221118a, STAC-ML.Markets.Inf.S.LSTM_C.[1,2,4].ENERG_EFF.v1

Original URL: https://blogs.nvidia.com/blog/2023/02/02/stac-ml-inference-gpu/