EXECUTIVE BRIEF

GPU-ACCELERATED COMPUTE MARKS A NEW ERA FOR FINANCIAL TRADING

SUMMARY

In the fast-paced world of financial trading, where milliseconds can mean millions of dollars, the stakes are high. A continuous stream of quantitative and qualitative data informs market conditions, and the speed and accuracy with which this data is analyzed directly impacts decision making and the bottom line. This is why the financial services industry is historically a quick adopter of new technologies, which has transformed the exchange of monetary assets.

With artificial intelligence, combined with high-performance computing, financial institutions can harness NVIDIA tools to learn from vast amounts of data and respond quickly to market fluctuations.

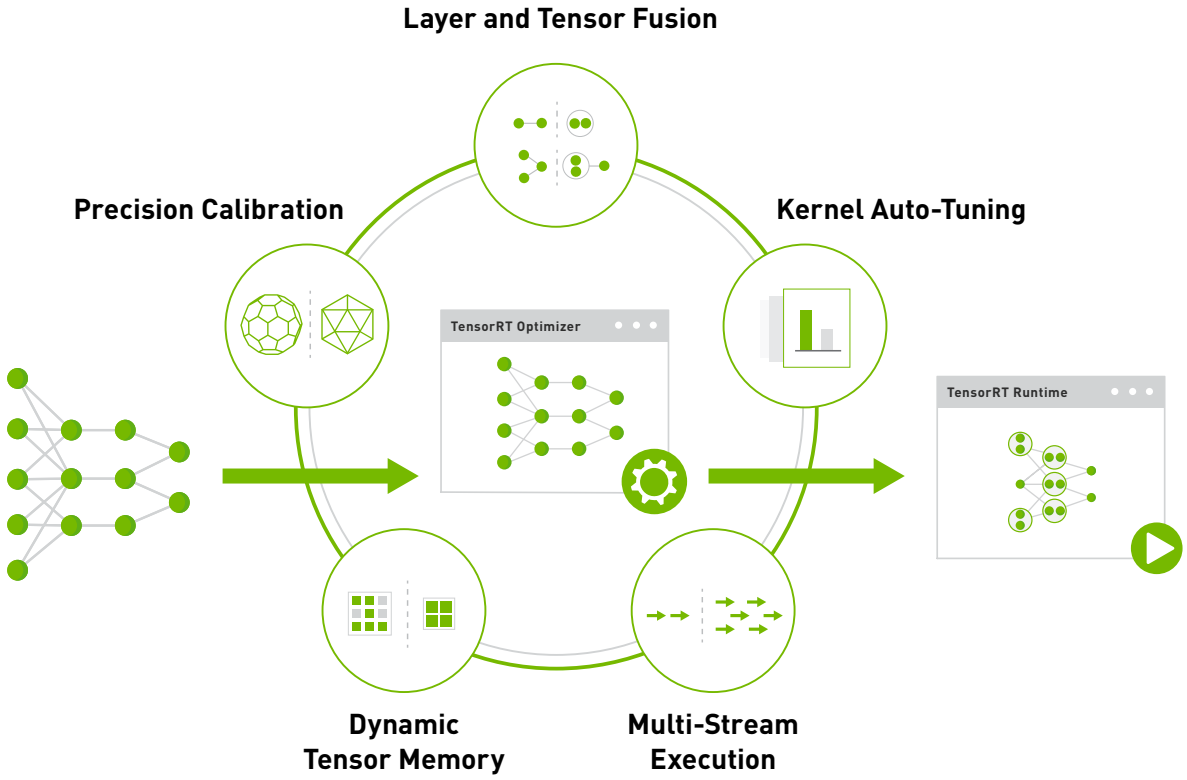
REDUCE TRADING ALGORITHM BACKTESTING TIME

Backtesting is a key step in a trading algorithm's journey from concept to revenue. Once an algorithm is developed and tested, usually on a very small subset of data, it's ready for more robust backtesting and tuning on vast volumes of historical data. Various parameters are tweaked, and the ensuing algorithm simulates its trading behavior over history. Each parameterization of the algorithm has a profit or loss calculated, and these scores are used to determine whether the algorithm will be moved to live trading. This iterative process takes days, weeks, or months, so it's crucial for it to be completed as soon as possible with the highest degree of accuracy on the widest array of information.

NVIDIA's AI platform delivers a 6,000X speedup over the previously set benchmark for backtesting in algorithmic trading. The breakthrough result has been validated by the Securities Technology Analysis Center (STAC), which defined the parameters for the benchmark testing. Using an NVIDIA® DGX-2™ system running accelerated Python libraries, NVIDIA ran 20 million simulations, versus the previous STAC-A3 record of 3,200 simulations, in 60 minutes.

Whether a traditional C++ model or a machine learning model needs backtesting, NVIDIA-accelerated infrastructure will likely pay for itself in the first year just through electricity costs. Also, the faster the model can get through backtesting, the faster it can get to market, which may allow the algorithm to successfully trade for a longer window before becoming stale.

For deep learning trading models developed in Tensorflow or PyTorch, NVIDIA TensorRT™ software optimizes trained deep learning networks. TensorRT takes the carefully trained network, once all the parameters and weights are known, and effectively compiles the model into an equivalent but more efficient version.



TENSORRT OPTIMIZES NEURAL NETWORK MODELS TRAINED IN ALL MAJOR FRAMEWORKS

Depending on the model and the data domain, data scientists can also choose to have TensorRT automatically optimize the model for reduced-precision computing using the Tensor Cores built into NVIDIA V100 and T4 GPUs. It allows even greater acceleration with minimal impact on network accuracy: Speedups by 10X are possible, depending on the data size.

ACCELERATE MODEL DEVELOPMENT

Financial modeling involves a considerable amount of expertise and time. The speed of NVIDIA-accelerated systems enables new design choices for a variety of models.

In the case of custom C++ models, a quant may write CUDA® C++ code—standard C++ with some additional decorators—and leverage optimized libraries for matrix- or signal-processing functions.

For more traditional machine learning models, a data scientist or quant may use Python on the Rapids suite of open-source libraries to unlock GPU acceleration. With deep learning models, TensorFlow, Keras, or PyTorch models can be used to configure the framework for GPUs.

FEATURES OF RAPIDS OPEN-SOURCE SOFTWARE

HASSLE-FREE INTEGRATION

Accelerate Python data science toolchain with minimal code changes and no new tools to learn.

TOP MODEL ACCURACY

Increase machine learning model accuracy by iterating on models faster and deploying them more frequently.

REDUCED TRAINING TIME

Drastically improve productivity with near-interactive data science.

OPEN SOURCE

Customizable, extensible, interoperable—the open-source software is supported by NVIDIA and built on Apache Arrow.

NVIDIA-accelerated systems allow a variety of speed-enabled design choices. Assuming only a 10X speedup on a model that used to take a week to run:

- Take it all as speed, and develop the same number and complexity of models in four hours.
- Explore the model parameter space, and test 5X the parameters to deliver a better model in half the time.
- Build a smarter model. Double the complexity of the model and test 2X the parameters in half the time.

PROPEL TRADING SIGNALS WITH SMARTER MARKET ANALYTICS

The end goal is for models to initiate timely and accurate trading signals. Intelligent market analytics boost risk management and reduce the total cost of ownership and infrastructure.

In 2011, J.P. Morgan Chase, the largest investment bank entity, used NVIDIA Tesla® GPUs to deliver a 40X increase in the end-to-end speed of its risk calculations while reducing the cost of ownership by 75 percent. Risk calculations now run in minutes instead of hours. The integration of GPUs into the shared global computational infrastructure has resulted in GPU-utilization rates approaching 70 percent, 24 hours a day.

A leading provider of real-time risk analytics on global derivatives markets, Hanweck Associates uses CUDA C++ programming language on Tesla GPUs to calculate implied volatilities for the entire Options Price Reporting Authority feed in real time.

Learn more about NVIDIA's array of solutions for financial services on its **industry webpage.**

NVIDIA