

TECHNICAL OVERVIEW

CREATING A NEW ERA OF INTELLIGENT TRADING WITH GPU-ACCELERATED COMPUTE



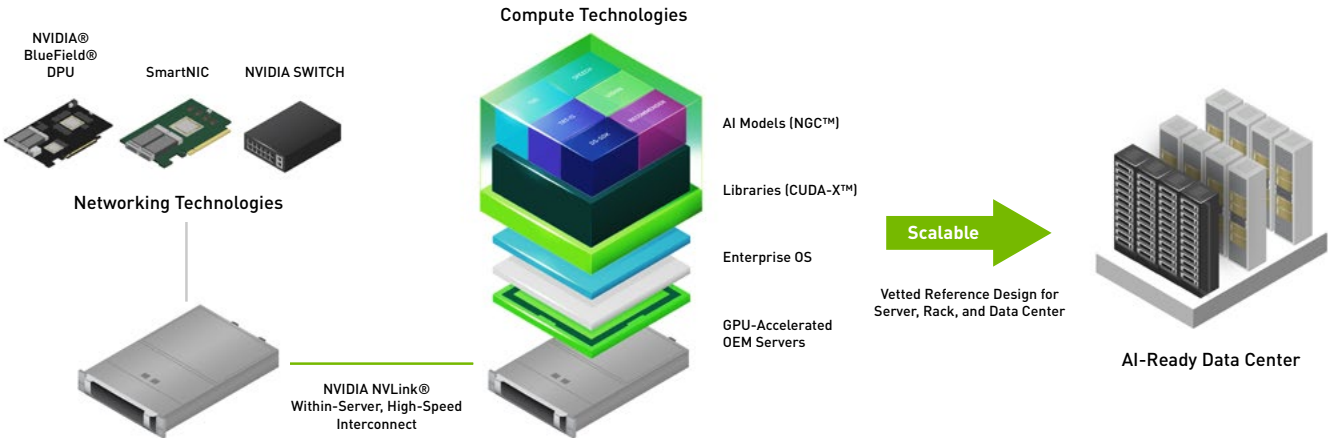
Algorithmic trading—responsible for the majority of U.S. equities trading volume—is expected to be an **\$18 billion industry** by 2024 according to **MarketsandMarkets**. As previous sources of competitive advantage such as alternative data and pure speed become more commoditized, technologies like AI are playing a critical role in determining the next generation of trading leaders.

The public cloud helped squeeze the final few drops out of Moore’s law, but the pace of legacy computing is no longer sufficient to compete with the cutting-edge trading techniques of today. NVIDIA has set its sights on a 1,000,000X speedup in the decade to come—with innovation spanning the full computing stack.

By using NVIDIA tools to harness the power of AI, combined with high-performance computing (HPC), institutional investors can generate better signals from rapidly increasing data volumes, optimize portfolio adjustments, automate execution, improve risk management, price derivatives with more accuracy, and respond quickly to market fluctuations.

NVIDIA AI End-to-End Platform

The Fastest AI Solution with Easy Deployment into Production

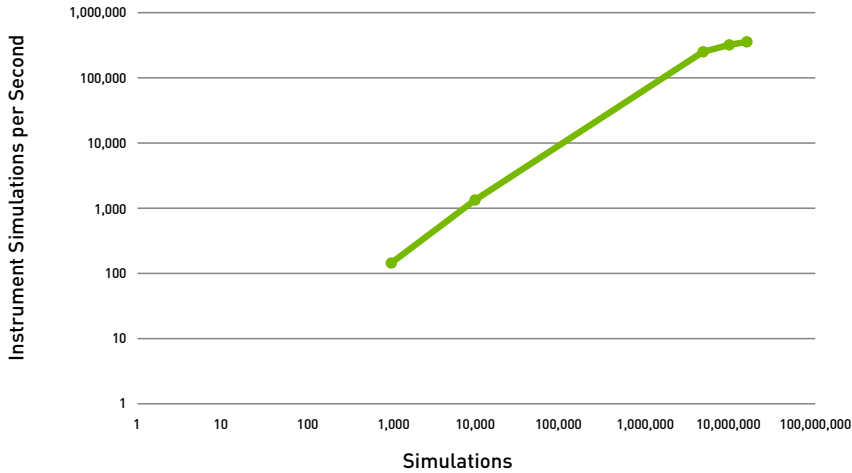


Reduce Algorithm Backtesting Time

The NVIDIA AI platform delivers a 1,000X 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 3.2 million simulations, versus the previous STAC-A3 record of 3,200 simulations, in 60 minutes.

3.2 MILLION SIMULATIONS,
 versus the previous STAC-A3 record of 3,200 simulations, in
60 MINUTES.

Scalability

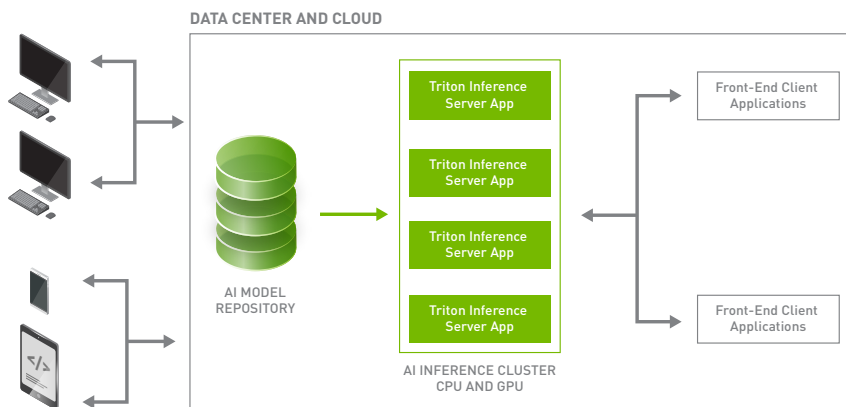


Results from STAC-A3.B1.SWEEP.SCALING.TIME | DGX-2 system is one node [16x NVIDIA V100 Tensor Core GPUs] – SUT ID: NVDA190425

Enhance Research Signal Detection

NVIDIA Triton™ Inference Server software optimizes trained deep learning models developed in Tensorflow or PyTorch. Once all the parameters and weights are known, Triton takes the carefully trained model and effectively compiles the model into an equivalent but more efficient version.

Depending on the model and data domain, data scientists can also choose to have Triton Inference Server automatically optimize the model for reduced-precision computing using the Tensor Cores built into NVIDIA A100 and T4 Tensor Core GPUs. It allows even greater acceleration with minimal impact on network accuracy: Speedups by 10X are possible, depending on the data size.



NVIDIA Triton optimizes neural network models trained in all major frameworks.

Accelerate Algorithmic Model Development

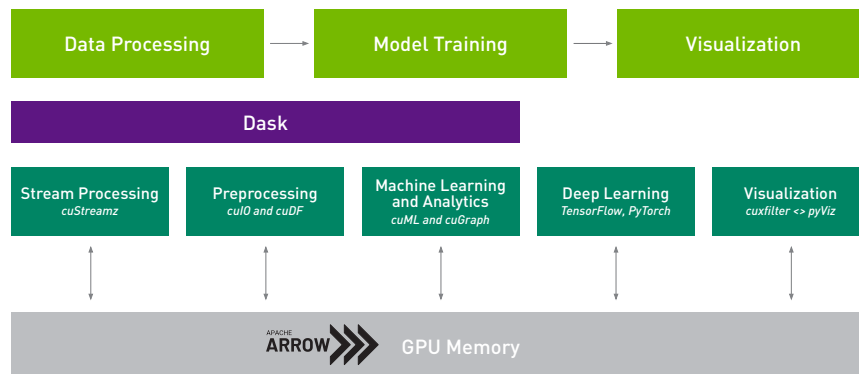
For traditional machine learning models, a data scientist or quantitative analyst (quant) may use Python on the NVIDIA RAPIDS™ suite of open-source libraries to unlock GPU acceleration. With deep learning models, TensorFlow, Keras, or PyTorch can be used to configure the framework for GPUs.

RAPIDS also focuses on common data preparation tasks for analytics and data science. This includes a familiar DataFrame API that integrates with a variety of machine learning algorithms for end-to-end pipeline accelerations without paying typical serialization costs. Conducting extract, transform, and load (ETL) functions on a 1.3 terabyte (TB) dataset using NVIDIA A100 GPUs, NVTabular took under two minutes. A CPU-equivalent took three hours.

And with Dask, RAPIDS can scale out to multi-node, multi-GPU clusters to power through big data processes.

RAPIDS Accelerates the Entire Data Science Process

Creating Enterprise-Grade Data Science from Pure Python



Boost HPC Performance, Portability, and Productivity

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

The NVIDIA HPC SDK's C, C++, and Fortran compilers support GPU acceleration of HPC modeling and simulation applications with standard C++ and Fortran, OpenACC® directives, and CUDA. GPU-accelerated math libraries maximize performance, and optimized communications libraries enable standards-based, multi-GPU, and scalable systems programming. Performance profiling and debugging tools simplify porting and optimization of HPC applications, and containerization tools make it easy to deploy on premises or in the cloud. With support for NVIDIA GPUs and Arm®, OpenPOWER, or x86-64 CPUs running Linux, the HPC SDK provides the essential tools for building NVIDIA GPU-accelerated HPC applications.

Highlights of the NVIDIA HPC SDK

| cuBlas | cuNumeric | cuQuantum |
|---|---|---|
| <ul style="list-style-type: none"> > Implement GPU-optimized basic linear algebra subroutines (BLAS). > Conduct vector-vector, matrix-vector, and matrix-matrix multiplication. > Execute mixed-precision, multi-GPU, and batched operations. | <ul style="list-style-type: none"> > Transparently accelerate and scale existing NumPy workloads. > Program from the edge to the supercomputer by changing one import line. > Pass data between legate libraries without worrying about distribution or synchronization requirements. | <ul style="list-style-type: none"> > Use optimized libraries and tools to accelerate quantum computing workflows. > Speed up quantum circuit simulations based on state vector and tensor network methods by orders of magnitude. |

Enhance Risk Calculations and Derivatives Pricing

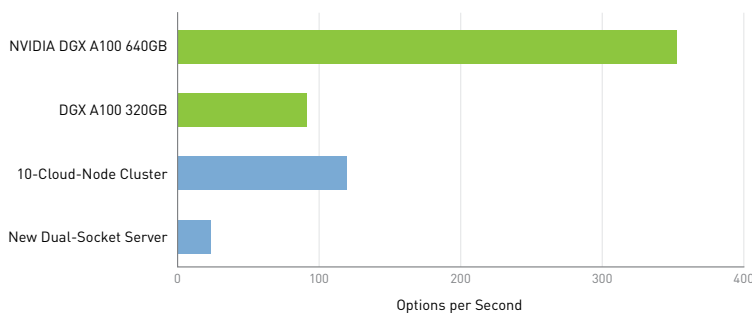
The latest NVIDIA DGX systems, with 640 gigabytes (GBs) of GPU memory each, landed eight performance records on the financial industry’s widely watched **STAC-A2 benchmark** of financial risk models, including the top honors for energy and space efficiency. Some of the largest firms on Wall Street and the broader global financial industry rely on STAC-A2 as a key risk model benchmark for measuring compute platform performance.

The latest NVIDIA DGX A100 systems with 640GB of GPU memory deliver 14.8X greater throughput (number of options priced per second) than a recently tested solution based on a single standard CPU server—a dual-socket CPU-based system—as measured by the STAC-A2 benchmark.

These were also the first publicly released STAC-A2 results for a solution using containers that have been audited. Kubernetes, an open-source container orchestration platform, has become the de facto standard for managing containerized workloads, which are instrumental in deploying complex multi-stage workflows.

Over 3X the Throughput

NVIDIA with OpenShift Completed the Most Calculations per Second in STAC-A2



STAC-A2.02.HPORTFOLIO.SPEED
 NVIDIA DGX A100 640GB - SUT ID NVDA210914 | DGX A100 320GB - SUT ID NVDA200909 | 10-Cloud-Node Cluster - SUT ID INTC210331 | New Dual-Socket Server - SUT ID INTC210315

Proof Points: Pricing, Risk, and Portfolio Optimization

- > **J.P. Morgan Chase** used NVIDIA T4 GPUs to deliver a **40X increase** in the end-to-end speed of their risk calculations, while reducing the cost of ownership by 75 percent. Risk calculations now run in minutes instead of hours. By integrating GPUs into the global computing infrastructure, they've experienced nearly **70 percent** GPU-utilization rates, 24 hours a day.
- > **CBOE**, a leading provider of real-time risk analytics on global derivatives markets, uses the CUDA C++ programming language on NVIDIA T4 GPUs to calculate implied volatilities for the entire Options Price Reporting Authority (OPRA) feed in real time.
- > **Munich Re Markets** has developed an AI-based analytics and quality assurance tool for diversified portfolio construction. Working with NVIDIA, Munich Re Markets rebuilt their models, which resulted in a **50X** performance boost.
- > **Riskfuel**, a startup developing fast and accurate derivatives models, tested a deep neural network-based model that prices over-the-counter (OTC) derivatives in real time. Compared to a CPU approach that yielded 32 valuations per second, Riskfuel was able to achieve 915,000,000 valuations per second—**28 million times faster**—utilizing a Microsoft Azure instance featuring NVIDIA A100.
- > **Federal Reserve** analysis showed speedups of **over 250X** for GPU versus CPU in running Monte Carlo simulations for European and American options pricing.
- > **Oxford-Man Institute of Quantitative Finance (OMI) and University of Toronto (U of T)** researchers have recently published limit order book (LOB) studies, and the latter was accelerated on GPUs at the University of Iowa's Interactive Data Analytics Service. It was measured to be **22X faster** in training time and projected to be **40X faster** on large data center GPUs for the very accurate Random Forest classifier. LOBs are available for many of the foreign exchange currency markets, as well as many equity names and cryptocurrencies.

For example, for Google, bid-and-ask prices by volume appear on the market on a nanosecond timestamp basis. There can be as many as ten bids and ten ask prices. Predicting the market direction of the mid-quoted prices has shown to be in the 70–80 percent range by the OMI and U of T studies when applying machine learning. By speeding up the machine learning training time for this kind of financial market research with GPUs, researchers save valuable time to production.

NVIDIA Roadmap: Grace and Hopper

By bringing together the NVIDIA Grace™ and NVIDIA Hopper™ architectures with NVIDIA® NVLink®-C2C, the **NVIDIA Grace Hopper Superchip** delivers a CPU+GPU coherent memory model for accelerated AI and HPC applications. AI models are exploding in complexity and size as they improve conversational AI with hundreds of billions of parameters, enhance deep recommender systems containing tens of terabytes of data, and enable new scientific discoveries. These massive models are pushing the limits of today's systems. Continuing to scale them for accuracy and usefulness requires fast access to a large pool of memory and a tight coupling of the CPU and GPU.

The Leader in MLPerf and Algorithmic Model Deployment

Two years after the debut of NVIDIA A100, NVIDIA AI is the only platform to have completed all benchmark tests in MLPerf and continues to accelerate algorithmic model deployment. In fact, 90 percent of MLPerf submissions are powered by NVIDIA AI. The platform is universal and performant for every model and every framework, scales to any size, accelerates end-to-end AI from data preparation to training to inference, and is available on every major cloud and server maker to deliver smarter, securer financial services.

Ready to Get Started?

To learn more about NVIDIA AI and HPC solutions for financial services, visit www.nvidia.com/finance

To access NVIDIA's full software suite in the NVIDIA NGC catalog, visit catalog.ngc.nvidia.com

To learn more about NVIDIA Grace Hopper, visit www.nvidia.com/grace-cpu