



# NVIDIA DGX A100

## THE UNIVERSAL SYSTEM FOR AI IN FINANCE

The NVIDIA DGX™ A100 offers unmatched performance and versatility across the financial services industry, powering AI applications that can help organizations trade smarter, improve customer relationships, cut costs, and minimize risk. With today's volatile markets, the ability to generate deeper market insights and faster risk calculations, accelerate fraud detection, and enable conversational AI services can contribute directly to the bottom line. DGX A100 delivers the performance needed to power these calculations, as well as analytics at enterprise scale.

### Use Cases for Financial Services

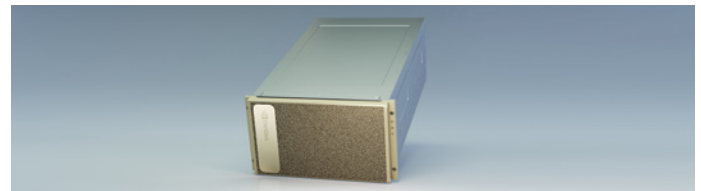
- > Accelerating computing for trading, risk management, fraud detection, and conversational AI
- > Enhancing and personalizing customer experiences
- > Achieving stronger fraud detection rates and lower false positives
- > Reducing operating costs
- > Generating faster risk calculations

### Driving Revenue, Mitigating Risk, and Enhancing Customer Experiences

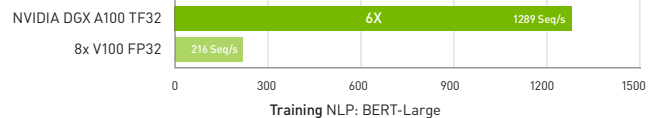
Today's traders need to be able to test and simulate model hypotheses with more speed and accuracy than the competition and extract deeper insights from more signals than ever before. DGX A100 dramatically shortens the time needed to develop and deploy models, creating both a depth of insight and a time-to-market advantage.

To identify fraud using large datasets, companies can use advanced machine learning techniques like boosted trees, which effectively cause fraudsters to identify themselves in the data. As organizations add more data, AI techniques like deep learning can help them achieve even better detection rates and lower false positives.

Conversational AI is powering services that enhance customer experiences, increasing satisfaction and loyalty. At the heart of conversational AI are deep learning models, but those require significant computing power to train chatbots that can communicate in the domain-specific language of financial services. DGX A100 delivers the fastest time to solution on the most complex models for incredible levels of language understanding.

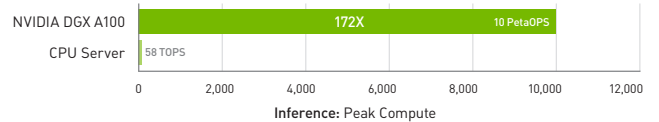


#### DGX A100 Delivers 6 Times The Training Performance



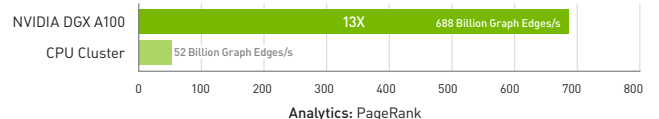
BERT Pre-Training Throughput using PyTorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 | V100: DGX-1 with 8x V100 using FP32 precision | DGXA100: DGX A100 with 8x A100 using TF32 precision

#### DGX A100 Delivers 172 Times The Inference Performance



CPU Server: 2x Intel Platinum 8280 using INT8 | DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity

#### DGX A100 Delivers 13 Times The Data Analytics Performance



3000x CPU Servers vs. 4x DGX A100 | Published Common Crawl Data Set: 128B Edges, 2.6TB Graph

## Fastest Time to Solution for Finance Services

NVIDIA DGX A100 features eight NVIDIA A100 Tensor Core GPUs, which deliver unmatched acceleration, and is fully optimized for NVIDIA CUDA-X™ software and the end-to-end NVIDIA data center solution stack. NVIDIA A100 GPUs bring a new precision, Tensor Float (TF32), which works just like FP32 but provides 20X higher floating operations per second (FLOPS) for AI compared to the previous generation. Best of all, no code changes are required to achieve the speedup. And when using NVIDIA's automatic mixed precision with FP16, A100 offers an additional 2X boost to performance with just one additional line of code.

The A100 GPU has 1.6 terabytes per second (TB/s) of memory bandwidth, a greater than 70 percent increase over the last generation. It also has significantly more on-chip memory, including a 40 megabyte (MB) level 2 cache that's nearly 7X larger than the previous generation. DGX A100 also debuts the third generation of NVIDIA® NVLink®, which doubles the GPU-to-GPU direct bandwidth to 600 gigabytes per second (GB/s), almost 10X higher than PCIe Gen 4. This unprecedented power delivers the fastest time to solution, allowing users to tackle challenges that weren't possible or practical before, such as generating faster risk calculations or achieving stronger fraud detection rates.

## The World's Most Secure AI System for Financial Services

NVIDIA DGX A100 delivers the most robust security posture for financial services firms, with a multi-layered approach that secures all major hardware and software components. Stretching across the baseboard management controller, CPU board, GPU board, self-encrypted drives, and secure boot, DGX A100 has security built in, allowing IT to focus on operationalizing AI rather than on threat assessment and mitigation.

## Unmatched Data Center Scalability with Mellanox

NVIDIA DGX A100 is the foundational building block for large AI clusters like NVIDIA DGX SuperPOD™, the enterprise blueprint for scalable AI infrastructure. DGX A100 features eight single-port Mellanox ConnectX-6 VPI HDR InfiniBand adapters for clustering and one dual-port ConnectX-6 VPI Ethernet adapter for storage and networking, all capable of 200 gigabits per second (Gb/s). The combination of GPU-accelerated compute with state-of-the-art networking means DGX A100 can scale to thousands of nodes to meet the biggest challenges, such as conversational AI and large-scale image classification.

## Proven Infrastructure Solutions Built with Trusted Data Center Leaders

In combination with leading storage and networking technology providers, a portfolio of infrastructure solutions are available that incorporate the best of the NVIDIA DGX POD™ reference architecture. Delivered as fully integrated, ready-to-deploy offerings through the NVIDIA Partner Network (NPN), these solutions simplify and accelerate data center AI deployments.

## Direct Access to NVIDIA DGXperts

NVIDIA DGX A100 is more than a server. It's a complete hardware and software platform built upon the knowledge gained from the world's largest DGX proving ground—NVIDIA DGX SATURNV—and backed by thousands of DGXperts at NVIDIA. DGXperts are AI-fluent practitioners who offer guidance and design expertise to fast-track AI transformation. They've built a wealth of experience over the last decade to help maximize the value of DGX investments, ensuring that applications get up and running quickly and stay running smoothly, for dramatically improved time to insights.

To learn more about NVIDIA DGX A100, visit [www.nvidia.com/dgx-a100](http://www.nvidia.com/dgx-a100)

To learn more about NVIDIA technology's top use cases for financial services, visit [www.nvidia.com/finance](http://www.nvidia.com/finance)

© 2020 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA-X, DGX A100, DGX POD, DGX SuperPOD, NVLink, and NVSwitch are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners. AUG20

## SYSTEM SPECIFICATIONS

GPUs	<b>8x NVIDIA A100 Tensor Core GPUs</b>
GPU Memory	<b>320 GB total</b>
Performance	<b>5 petaFLOPS AI 10 petaOPS INT8</b>
NVIDIA NVSwitches <b>6</b>	
System Power Usage	<b>6,500 W max</b>
CPU	<b>Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)</b>
System Memory	<b>1TB</b>
Networking	<b>8x Single-Port Mellanox ConnectX-6 VPI 200 Gb/s HDR InfiniBand 1x dual-port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet</b>
Storage	<b>OS: 2x 1.92 TB M.2 NVME drives Internal storage: 15 TB (4x 3.84 TB) U.2 NVME drives</b>
Software	<b>Ubuntu Linux OS</b>
System Weight	<b>271 lbs (123 kgs)</b>
Packaged System Weight	<b>315 lbs (143 kgs)</b>
System Dimensions	<b>Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) max Length: 35.3 in (897.1 mm) max</b>
Operating Temperature Range	<b>5°–30°C (41°–86°F)</b>

