Solution Overview

NVIDIA

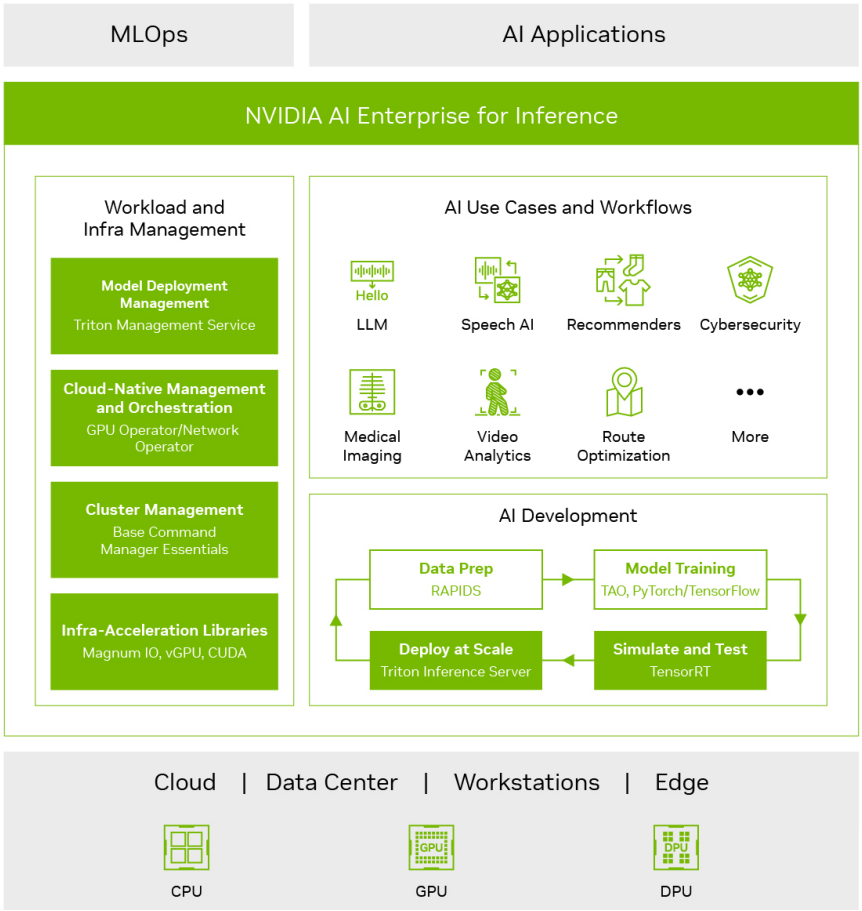
Faster, More Accurate AI Inference

Drive breakthrough performance in your AI-enabled applications and services.

The Challenges of AI Inference Deployments

Deploying and running models in production applications is complex. Organizations must account for constantly evolving model frameworks with different teams using different frameworks. The computing processors, accelerators, and software platforms are also evolving at an unprecedented pace. The number and type of models are expanding as AI is making its way into many areas of business.

For the successful use and growth of AI, organizations need a full-stack approach to AI inference that supports the end-to-end AI lifecycle, with tools that help all teams reach their goals.



NVIDIA AI inference platform workflow, from trained model to end-user application.

Key Challenges for AI In Production

- **Latency:** Real-world applications often require low latency and high throughput.
- **Integration with production data:** Different feature transformations in model training and live inference can deliver inaccuracy in production.
- **Different platforms:** Cloud (AWS, GCP, Azure, etc.), on-premises servers, mobile devices, and IoT devices are all possible locations to deploy inference.
- **Different models and machine learning frameworks:** As the models and frameworks differ, best practices for deploying and integrating also differ.
- **Scaling:** Multiple models can be involved in inference, resulting in complex prediction flows.
- **Diverse CPUs and GPUs:** Models can be executed on a CPU or GPU, and there are different types of GPUs and CPUs.

Often, organizations end up having multiple, disparate inference serving solutions—per model, per framework, or per application.

1 "Gartner Identifies the Top Strategic Technology Trends for 2021," Gartner, October 2020

Deploy Next-Generation AI Inference With the NVIDIA AI Platform

NVIDIA AI offers a complete end-to-end stack and suite of products and services to deliver the performance, efficiency, and responsiveness that is critical to powering the next generation of AI inference—in the cloud, in the data center, at the network edge, or in embedded devices.

Accelerated Production AI With NVIDIA AI Enterprise

NVIDIA® TensorRT™, **NVIDIA TensorRT-LLM**, **NVIDIA Triton Inference Server**, and **NVIDIA Triton Management Service (TMS)** are part of **NVIDIA AI Enterprise**, an end-to-end software platform that streamlines AI development and deployment and provides support for enterprise-level production inference.

Key features and benefits:

- Over 100 frameworks, pretrained models, AI workflows, and development tools accelerate and simplify an enterprise's journey to AI.
- Security and API stability are ensured for interdependent software packages within the inference software stack.
- Software optimizations increase performance and lower TCO.
- Infrastructure and inference management are simplified with NVIDIA Base Command™ Manager Essentials and Triton Management Service.
- Industry ecosystem certifications are available across the cloud, data center, workstations, and edge.
- The software platform comes with enterprise support with service-level agreements and access to AI experts.

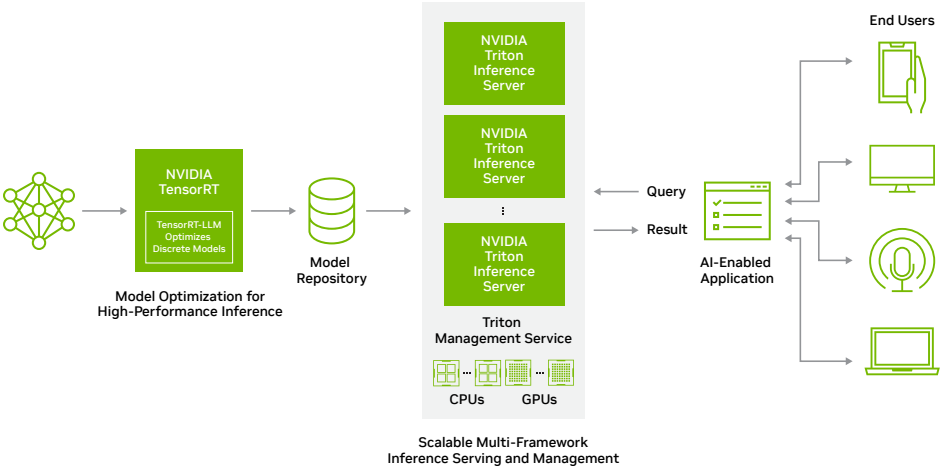


Diagram displaying how NVIDIA's AI inference software works.

NVIDIA TensorRT

NVIDIA TensorRT, an SDK for high-performance deep learning inference, includes a deep learning inference optimizer and a runtime that deliver low latency and high throughput for inference applications. TensorRT can be deployed, run, and scaled with Triton.

The End-to-End NVIDIA AI Inference Platform

- **NVIDIA AI Enterprise:** An end-to-end AI software platform consisting of NVIDIA Triton and TensorRT to simplify building, sharing, and deploying AI applications. With enterprise-grade support, stability, manageability, and security, organizations can accelerate time to value while eliminating unplanned downtime.
- **NVIDIA TensorRT:** An SDK for high-performance deep learning inference, key for delivering low latency and high throughput to inference applications.
- **NVIDIA TensorRT-LLM:** An open-source library enabling developers to accelerate and optimize inference performance on the latest LLMs on NVIDIA GPUs.
- **NVIDIA Triton Inference Server:** An open-source inference serving software that's key for fast and scalable inference in every application.
- **NVIDIA Triton Management Service:** A software application that automates the deployment of multiple Triton Inference Server instances in Kubernetes with resource-efficient model orchestration on GPUs and CPUs.
- **Wide Ecosystem Integration:** Triton and TensorRT are tightly integrated into and work with all major platforms, including Amazon Sagemaker, Azure Machine Learning, Google Vertex AI, TensorFlow, Pytorch, and more.
- **Accelerated Computing Infrastructure:** Hardware systems designed for optimizing AI workloads.

Key features:

- Built on the **NVIDIA CUDA®** parallel programming model, TensorRT optimizes techniques such as quantization, layer and tensor fusion, kernel tuning, and many more on NVIDIA GPUs.
- TensorRT provides INT8 using quantization-aware training and post-training quantization and floating point 16 (FP16) optimizations for deployment of deep learning inference applications, such as video streaming, recommendations, anomaly detection, and natural language processing.
- Integrations with all major frameworks, including Pytorch and TensorFlow, allow users to achieve 6X faster inference with a single line of code.

NVIDIA TensorRT-LLM is an open-source library that accelerates and optimizes inference performance of the latest LLMs on NVIDIA GPUs. It enables developers to experiment with new LLMs, offering speed-of-light performance with quick customization capabilities without deep knowledge of C++ or CUDA optimization.

TensorRT-LLM wraps TensorRT's Deep Learning Compiler, optimized kernels from FasterTransformer, pre- and post-processing, and multi-GPU/multi-node communication in a simple open-source Python API for defining, optimizing, and executing LLMs for inference in production.

NVIDIA Triton

NVIDIA Triton Inference Server is open-source inference serving software that helps standardize AI model deployment and execution in production from all major AI frameworks on any GPU- or CPU-based infrastructure.

Key features:

- High performance and utilization are achieved on both GPU and CPU systems through request batching and concurrent model execution.
- Stringent application latency service-level agreements (SLAs) for real-time and offline batched inference and large language models (LLMs) come with support for multi-GPU, multi-node execution and model ensembles.
- PyTriton provides a simple interface that lets Python developers use Triton Inference Server to serve anything, be it a model, a simple processing function, or an entire inference pipeline.
- Triton standardizes AI model deployment for all applications across cloud and edge, and it's in production at world-leading companies—Amazon, Microsoft, American Express, and thousands more.
- Triton's model analyzer can shrink model deployment time from weeks to days. It helps select the optimal deployment configuration to meet the application's latency, throughput, and memory requirements.

NVIDIA Triton Management Service (TMS) automates the deployment of multiple Triton Inference Server instances in Kubernetes with resource-efficient model orchestration on GPUs and CPUs. TMS, available exclusively with NVIDIA AI Enterprise, enables large-scale inference deployment with high performance and hardware utilization.

Key Benefits of NVIDIA AI Inference

- **Standardized deployment:** Standardize model deployment across applications, AI frameworks, model architectures, and platforms.
- **Easy integration:** Integrate easily with tools and platforms on public clouds, in on-premises data centers, and at the edge.
- **Lower cost:** Achieve high throughput and utilization from AI infrastructure, thereby lowering costs.
- **Seamless scalability:** Scale inference jobs seamlessly across one or more GPUs.
- **High performance:** Experience incredible performance with the NVIDIA inference platform, which has set records across multiple categories in MLPerf, the leading industry benchmark for AI.

Key features:

- Deployment is simplified through automation and management of instances on Kubernetes and grouping models by framework for efficient memory use.
- Resources are maximized by loading models on demand, unloading when not in use via a lease system, and placing as many models as possible on a single GPU.
- Each Triton Inference Server instance is monitored for health and capacity, and autoscales based on latency and hardware utilization.
- Inference deployment is efficiently managed, from a single model to hundreds of models, on premises or in any public cloud.

NVIDIA Accelerated Computing Infrastructure

NVIDIA offers a complete infrastructure portfolio for AI workloads, with support for full systems that include GPUs, **NVIDIA Grace™ CPUs**, NVIDIA GH200 Grace Hopper™ SuperChips, and accelerated networking with NVIDIA BlueField® data processing units (DPUs).

Key features:

- A complete GPU portfolio with **NVIDIA Hopper™**, **Ada Lovelace**, and **Ampere** architecture GPUs, from entry level to mainstream to the highest performance. Each GPU has the versatility to accelerate AI applications, whether at the edge, in the cloud, or on premises.
- The NVIDIA AI software stack is designed to work with all major x86 CPU systems, including AMD and Intel CPUs, and is optimized for the Grace CPU and the GH200 Grace Hopper Superchip, which combines the NVIDIA Grace and Hopper architectures to accelerate AI and high-performance computing (HPC) applications.
- The BlueField DPU platform offloads, accelerates, and isolates a broad range of advanced infrastructure services, providing AI data centers with high performance, robust security, and sustainability.

- NVIDIA Quantum InfiniBand and NVIDIA Spectrum™ Ethernet networking platforms provide AI practitioners with advanced acceleration engines and the fastest speeds at 400 gigabits per second (Gb/s), enabling superior performance for inference at scale.
- Designed for industry-leading performance and multi-node scale with NVIDIA DGX POD™ and DGX SuperPOD™, NVIDIA DGX™ systems are the gold standard in AI infrastructure, delivering the fastest time to solution on the most complex AI workloads.
- NVIDIA-Certified Systems™ bring NVIDIA GPUs and NVIDIA high-speed, secure network adapters to systems from leading NVIDIA partners in configurations validated for optimum performance, manageability, and scale.

Ready to Get Started?

To learn more about NVIDIA AI Enterprise, visit:

nvidia.com/ai-enterprise-suite

To contact us about purchasing our AI inference software, visit:

nvidia.com/ai-enterprise-sales

© 2023 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, Base Command, BlueField, CUDA, DGX, DGX POD, DGX SuperPOD, Grace, Grace Hopper, Hopper, NVIDIA-Certified Systems, Spectrum, TensorRT, and Triton are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 2730427. SEP23

NVIDIA