Deploying AI Models with Speed, Efficiency, and Versatility

Inference on NVIDIA's AI Platform

Whitepaper

Table of Contents

List of Figures

NVIDIA's Inference Solution

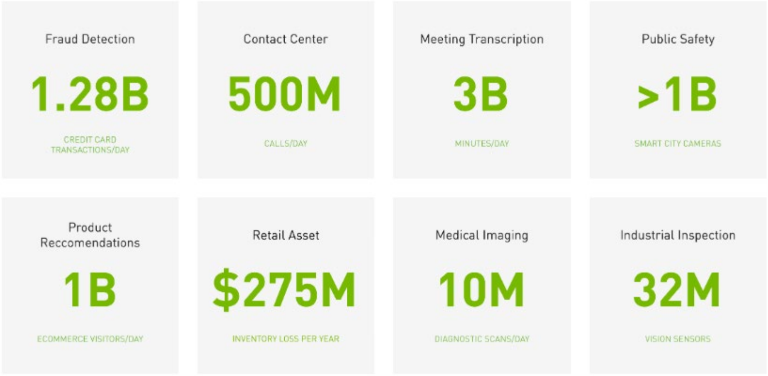
The AI Prototype to Production Gap in the Enterprise

Inference is where artificial intelligence (AI) delivers results, powering innovation across industries, including consumer internet, healthcare and life sciences, financial services, retail, manufacturing, and supercomputing. The rise of generative AI is requiring more powerful inference computing platforms. When modern AI applications, such as large language models (LLMs), are deployed at-scale, meeting real-time latency constraints across thousands of daily inference requests necessitates an optimized accelerated compute stack, starting from the hardware all the way to the software layer of any strong performing AI application. As researchers, MLOps engineers, data scientists, application developers, and more push the boundaries of what's possible in generative AI, computer vision, speech AI, natural language processing (NLP), and recommender systems, state-of-the-art AI models will continue to rapidly evolve and expand in size, complexity, diversity, and capability.

"The number of applications for generative AI is infinite, limited only by human imagination. Arming developers with the most powerful and flexible inference computing platform will accelerate the creation of new services that will improve our lives in ways not yet imaginable." - Jensen Huang, Founder and CEO of NVIDIA

The iPhone moment of AI, brought about by ChatGPT, has created overwhelming, unsurpassed demand for generative AI inference solutions across vertical markets and applications. However, for AI to have the utmost impact and deliver business results, trained AI models need to be integrated within applications and deployed on production systems - on premise, in the cloud, or at the edge - to "infer" things about new data that it's presented to it. AI inference performance at scale is critical for delivering the best end-user experience for your customers, minimizing the cost of AI deployments, and maximizing ROI for your AI projects. Imagine that your deployed AI models are trained to perfection for your use case but unable to deliver predictions or responses in real-time, or scale to support a spike in users requests? This is why AI inference requires accelerated computing.

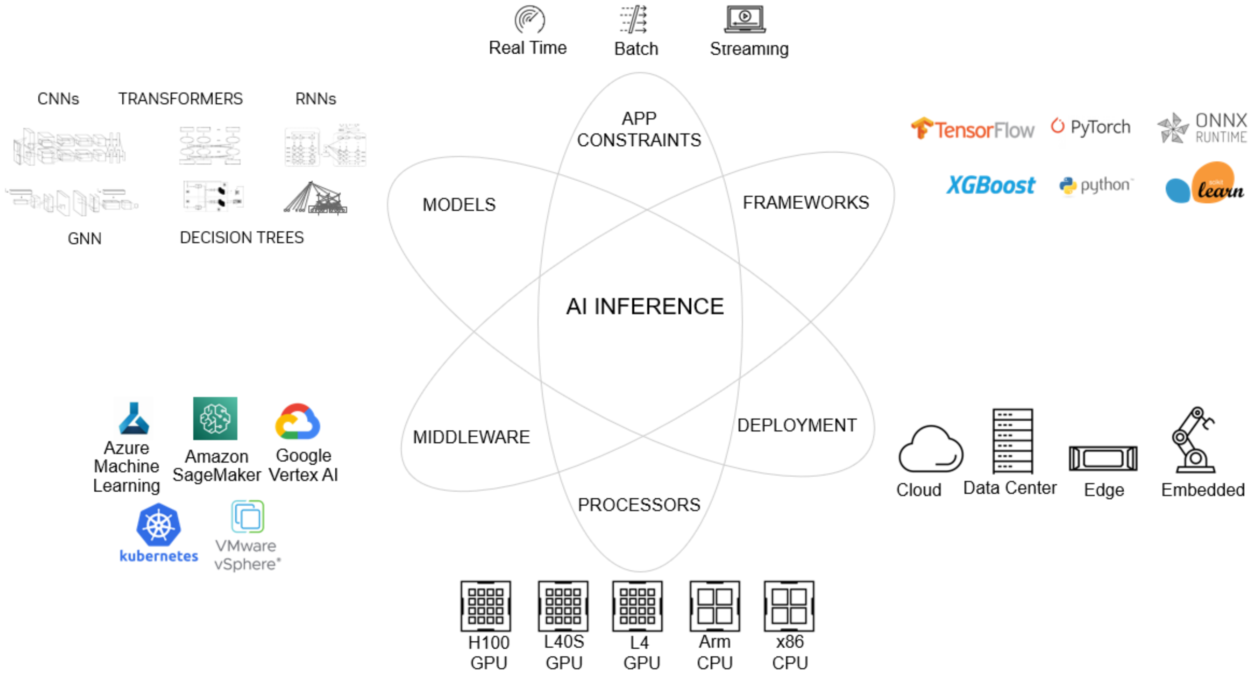
Figure 1. AI Inference Requests Across Example Use Cases



Based on NVIDIA analysis using public data and industry research reports

Operationalizing AI models within enterprise applications also poses several challenges due to the conflict between the nuances of model building and the operational realities of enterprise IT systems. Infrastructure for AI deployments requires the versatility to support diverse and ever-expanding AI model types and evolution, multiple AI frameworks, handling different types of inference query types, like batch, streaming and ensemble, and supporting multiple environments from edge to cloud. Many discrete and diverse pieces must work together in harmony to reach successful inference deployment, such as model selection, application constraints, framework training and optimization, deployment strategy, processor target, and orchestration and management software. The lack of a unified workflow for all of these areas in the inference equation creates an obstacle for enterprises and CSPs when it comes to meeting growing inference demand.

Figure 2. Challenges in AI Inference



In this paper, we begin with a view of the end-to-end deep learning workflow and move into the details of taking AI-enabled applications from prototype to production deployments. We'll cover the evolving inference usage landscape, architectural considerations for the optimal inference accelerator, and the NVIDIA AI inference platform, a complete end-to-end stack of products and services that delivers the performance, efficiency, and responsiveness critical to powering the next era of generative AI inference.

End-to-End AI Workflow Overview

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

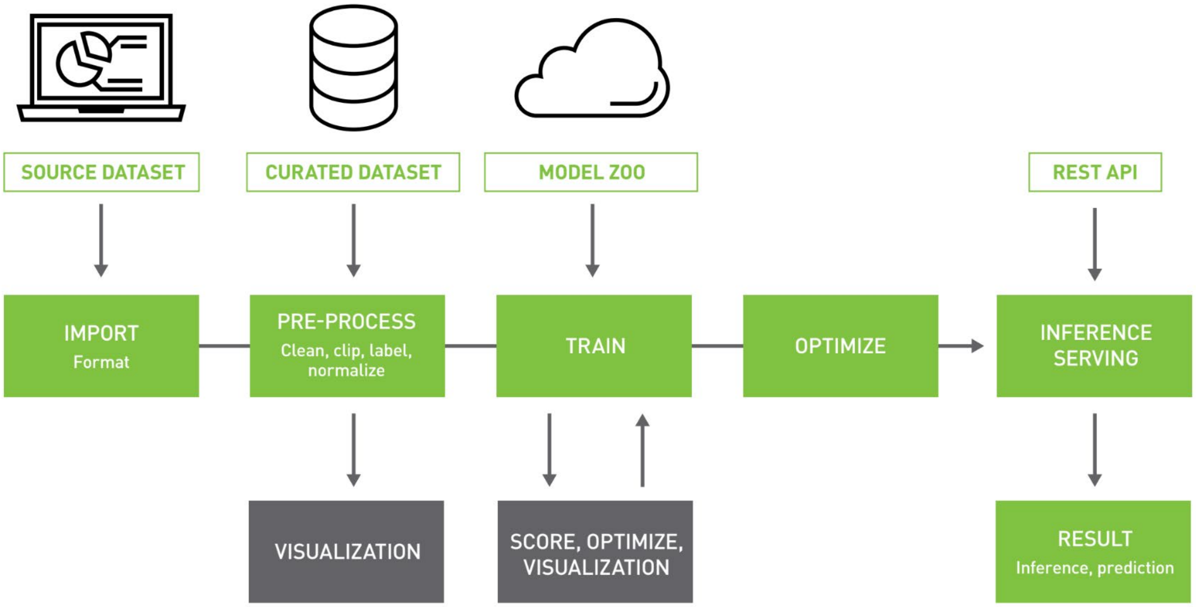
Building and deploying an AI-powered solution from idea to prototype to production is daunting. You need large volumes of data, AI expertise, and tools to curate, pre-process, and train AI models using this data, as well as to optimize for inference performance and finally deploy them into a usable, customer-facing application.

Additionally, training LLMs from scratch can be incredibly expensive and take weeks to months depending on the size of the task and underlying infrastructure. Off-the-shelf LLMs can also fall short in catering to the distinct requirements of organizations, whether it is the intricacies of specialized domain knowledge, industry jargon, or unique operational scenarios. This is precisely why modern end-to-end generative AI workflows take advantage of custom LLMs using various LLM customization techniques. These customized models provide enterprises with the means to create solutions personalized to match their brand voice and streamline workflows, for more accurate insights, and rich user experiences. Once the AI model, or ensemble of models, are ready for deployment, they must be optimized for inference in order to meet real-time latency requirements at scale.

This requires a full stack approach that solves for the entire workflow - start to finish - from importing and preparing data sets for training to deploying a trained network as an AI-powered service using inference.

See Figure 3 for end-to-end deep learning workflow, from training to inference.

Figure 3. End-to-End Workflow, from Training to Inference



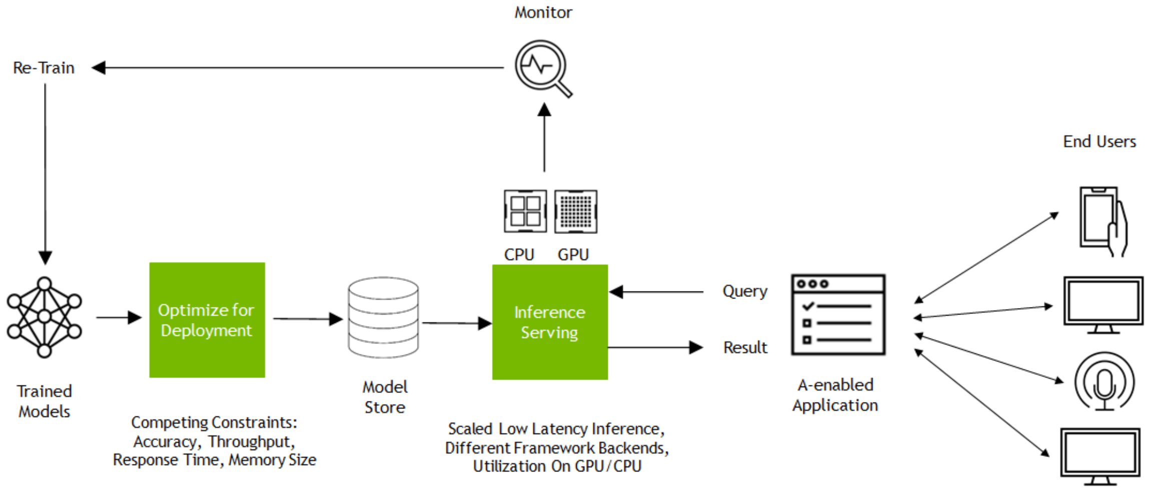
In many organizations, multiple teams are usually involved in AI development and deployment to production: data scientists, machine learning (ML) engineers, application developers, and IT operations. And while they work for the same organization, each has their own specific goals. Supporting the end-to-end lifecycle for AI requires both the developer tools and compute infrastructure to enable all teams to meet their goals.

The focus of this paper is mainly on the challenges of deploying trained AI models in production and how to overcome them to accelerate your path to production. However, a key prerequisite before you get to the deployment phase is, of course, to have completed the development phase of the AI workflow and have trained AI models that are ready to take to production.

AI Inference—Trained Model to Real Service

Trained AI models for your application only get you halfway there in terms of putting AI to work for your business. You need to integrate the trained models into actual applications, services, and products, and deploy them into the real-world to "infer" results on new data. Figure 4 shows the generic AI inference workflow.

Figure 4. Generic AI Inference Workflow



The Challenge of AI Inference Deployments at Scale

Community LLMs are growing at an explosive rate, with increased demand from companies to deploy these models into production. The size of these LLMs is driving the cost and complexity of deployment higher, requiring optimized inference performance for production generative AI applications. Higher performance not only helps decrease costs, but also improves user experiences. LLMs such as Llama, BLOOM, GPT3, Falcon, MPT, and Starcoder have demonstrated the potential of advanced architectures and operators. This has created a challenge in producing a solution that can efficiently optimize these models for inference.

Other AI-enabled applications like e-commerce product recommendations, voice-based assistants, and contact center automation require tens to hundreds of trained AI models, within the same deployed application, to deliver the desired user experience. It is important to consider the entire workflow of operationalizing trained models within production applications at scale.

The solution to deploy, manage, and scale these models with a guaranteed quality-of-service (QoS) in production is known as model or inference serving. Challenges of serving AI models at scale include supporting models trained in multiple deep learning and machine learning frameworks, handling different inference query types (real-time, batch, streaming, and ensemble, for example) and optimizing across multiple deployment platforms like CPUs and NVIDIA GPUs.

Additionally, you need to provision and manage the right compute infrastructure to deploy these AI models, with optimal utilization of compute resources and the flexibility to scale up or down to streamline operational costs of deployment. Deploying AI in production is both an inference serving and infrastructure management challenge, commonly referred to as the MLOps challenge. Clearly, taking AI from prototype to production and maximizing ROI on AI projects for your business requires a full-stack approach.

NVIDIA® AI Enterprise software suite offers a complete end-to-end software stack designed to tackle these challenges for MLOps engineers, data scientists, application developers, and infrastructure engineers, who are involved at different stages in the prototype to production process and with varying levels of AI expertise and experience.

Inference Performance of AI Models Matters

AI inference is where your end customers will interact with your AI-enabled applications and services, so inference performance of your trained AI model is crucial. The simplest inference method is to run samples through your model in-framework and turn off back propagation. However, this is far from optimal for production. Deployed AI services seek to deliver the highest level of service with the fewest number of servers. So, in-framework, by itself, is just a start. Inference deployments fall into one of two categories: high-batch/high-throughput "after-hours" workloads that can trade latency for high throughput, and real-time, latency-sensitive services that must immediately return the right answer.

If your AI models cannot deliver the right results fast enough and be deployed at scale with the fewest number of servers, it affects both the user experience and the ROI of your AI-powered applications. When considering a platform to deploy an AI-driven product or service, you must consider performance factors, including throughput, latency, accuracy, and efficiency. Let's break these considerations down one at a time:

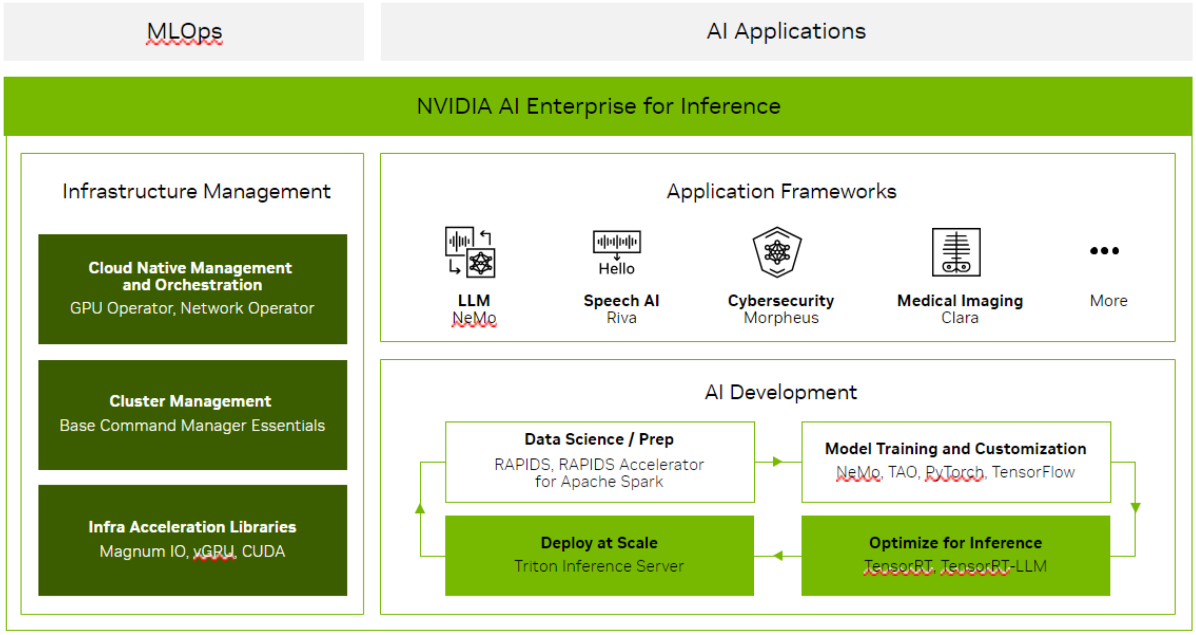
- LATENCY: Latency refers to how much time elapses from an input being presented to the AI model to an output being available. In some applications, low latency is a critical safety requirement. In other applications, latency is directly visible to users as a quality-of-service issue. For larger bulk processing, latency may not be important at all.
- THROUGHPUT: Throughput refers to how many inferences can be completed in a fixed unit of time. Higher throughput is better. Higher throughputs indicate a more efficient utilization of fixed compute resources. For "high-batch" offline inference applications that work on large amounts of data during off-peak hours, the total time will be determined by the throughput of the model.
- ACCURACY: While optimizing for inference performance, it's critical that an inference solution preserve the level of accuracy to ensure the AI model delivers the requisite results. Reduced precisions, such as FP16 and INT8, deliver 2-3x more performance compared to FP32 precision, with near-zero loss in accuracy. Additionally, NVIDIA's Hopper™ architecture advances fourth-generation Tensor Cores with the Transformer Engine using a new 8-bit floating point precision (FP8), which can enable massive speed ups in LLM inference (30x better performance than Ampere architecture).
- VERSATILITY: Hardware characteristics and speeds/feeds are important but are only useful if the enabling software allows developers to unlock the hardware's full potential. That takes the form of an end-to-end software stack that enables developers to optimize and deploy a broad range of AI model types, including image-based networks, language and speech networks, recommender systems, generative AI, and beyond.
- EFFICIENCY: Another important attribute of accelerated AI inference is the cost savings it can deliver around initial server cost (fewer server nodes), and the energy cost to power and cool this reduced number of servers throughout their lifecycle. This has multiple implications for on-premises deployments around rack efficiency, both in terms of power and number of rack slots occupied by these servers.

NVIDIA AI Inference Platform: The Full-Stack Approach

NVIDIA offers a full-stack approach to AI inference via NVIDIA AI software and GPU accelerated computing. They are the foundation for performance-optimized solution stacks that power a broad range of AI applications in production today, such as personalized shopping experiences, contact center automation, voice assistants, chatbots, visual search, and assisted medical diagnostics (see Figure 5).

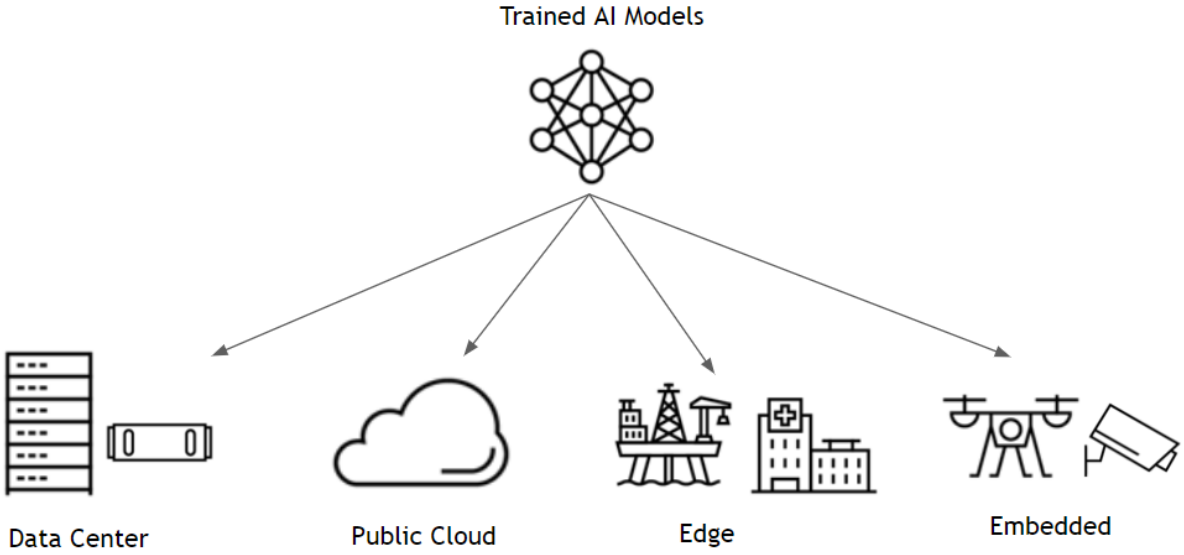
NVIDIA's inference platform delivers the performance, efficiency, and responsiveness critical to powering the next generation of AI products and services. The platform is a combination of architectural innovation, purpose-built to accelerate AI inference workloads, and an end-to-end software stack that is designed for data scientists, software developers, and infrastructure engineers, involved at different stages in prototype to production process and with varying levels of AI expertise and experience.

Figure 5. NVIDIA AI Inference Platform Stack


MLOps
AI Applications
NVIDIA AI Enterprise for Inference
Infrastructure Management
Cloud Native Management and Orchestration
GPU Operator, Network Operator
Cluster Management
Base Command Manager Essentials
Infra Acceleration Libraries
Magnum IO, vGPU, CUDA
Application Frameworks
Hello
LLM
NeMo
Speech AI
Riva
Cybersecurity
Morpheus
Medical Imaging
Clara
More
AI Development
Data Science / Prep
RAPIDS, RAPIDS Accelerator for Apache Spark
Model Training and Customization
NeMo, TAO, PyTorch, TensorFlow
Deploy at Scale
Triton Inference Server
Optimize for Inference
TensorRT, TensorRT-LLM

Depending on the service or product that you need to integrate your AI models into, and how your end customers will interact with it, the optimal place to execute AI inference can vary from inside the heart of the data center, on the public cloud, enterprise edge or in embedded devices (see Figure 6).

Figure 6. Diverse Use Cases Demand Diverse Deployment Environments for AI



Some industries, like healthcare for example, have well established rules about where data must be stored and how it can be accessed, and for these customers and industries, on-premises is likely the right call. Cloud deployments are a great choice, as well, since they provide on-demand compute as needed and allow organizations to ease into the AI transition before making larger IT investments.

NVIDIA Inference Infrastructure

The NVIDIA AI inference platform infrastructure consists of GPUs, CPUs, DPUs, networking solutions and complete systems to accelerate inference with maximum performance, security, and scalability.

> **GPU:** NVIDIA is the only vendor offering a complete portfolio with NVIDIA Hopper, Ada Lovelace, and NVIDIA Ampere™ GPUs from entry-level to mainstream to the highest performance, each providing the versatility to accelerate the widest range of AI inference applications, whether at the edge, in the cloud, or on premise. See the following section for a deeper dive into NVIDIA's GPU portfolio.

> **CPU:** NVIDIA AI software stack is designed to work with all major x86 CPU systems including AMD EPYC Milan and Genoa CPUs and Intel Emerald Rapids and Sapphire Rapids CPUs as part of our extensive NVIDIA-Certified Systems offering with all major OEMs.

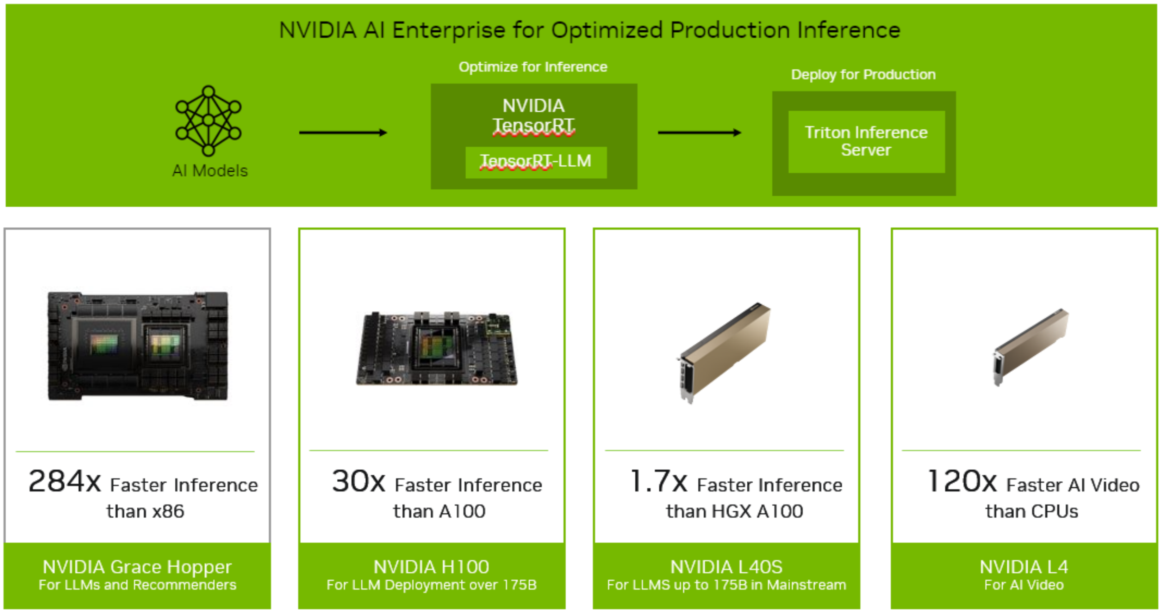
- **NVIDIA Grace™ CPU Superchip:** The NVIDIA AI software stack is also optimized for NVIDIA's Grace CPU, purpose built for giant scale AI and HPC applications. NVIDIA Grace CPU comes in two data center superchip products. The Grace Hopper Superchip pairs an NVIDIA Grace CPU with an NVIDIA H100 Tensor Core GPU in a coherent memory architecture for giant-scale NVIDIA AI applications. The NVIDIA Grace CPU Superchip combines 144 high-performance Arm cores with a fast NVIDIA designed fabric and LPDDR5X memory for HPC, demanding cloud and enterprise applications.
- **DPU:** The BlueField® DPU platform offloads, accelerates, and isolates a broad range of advanced infrastructure services, providing AI data centers with high-performance networking, robust security, and sustainability. For LLM AI inference, NVIDIA BlueField DPUs provide high-speed, low-latency connectivity between NVIDIA GPUs, which delivers consistent results as models grow in scale and complexity. It can handle networking tasks, storage tasks, security tasks, and AI acceleration.
- **NVIDIA DGX Systems:** NVIDIA DGX™ Systems are purpose-built for the unique demands of enterprise AI. Powered by NVIDIA Base Command software, and architecture for industry-leading performance and multi-node scale with DGX POD and DGX SuperPOD, DGX systems are the gold standard in AI infrastructure, delivering the fastest time-to-solution on the most complex AI workloads including natural language processing, recommender systems, data analytics and more, with direct access to NVIDIA AI experts.
- **NVIDIA-Certified Systems:** NVIDIA-Certified Systems brings NVIDIA GPUs and NVIDIA high-speed, secure network adapters to systems from leading NVIDIA partners in configurations validated for optimum performance, manageability, and scale. With an NVIDIA-Certified System, organizations can confidently choose enterprise-grade hardware solutions to power their accelerated computed workloads - from the desktop to the data center and edge.
- **Networking:** NVIDIA Quantum InfiniBand and Spectrum Ethernet networking platforms provide AI practitioners with advanced acceleration engines, and the fastest of interconnect speeds at up to 400Gb / s, enabling superior performance for inference at scale. These focus on enabling high-speed communication and data transfer between different parts of a distributed system. Focus on fast and efficient communication pathways between different nodes, allowing for rapid data exchange between servers, storage, and other components.

NVIDIA GPUs

NVIDIA offers a complete portfolio of GPUs, featuring Hopper and Ada Lovelace Tensor Core GPUs as the inference engine powering NVIDIA AI. Following are the inference GPUs (see Figure 7):

- **NVIDIA GH200 Grace Hopper Superchip**
  Enterprises need a versatile system to handle the largest models and realize the full potential of their inference infrastructure. The GH200 NVIDIA Grace Hopper™ Superchip delivers over 7x the fast-access memory to the GPU as traditional accelerated inference solutions and up to 284x more performance vs. CPUs to address LLMs, recommenders, vector databases and more.
- **NVIDIA H100 Tensor Core GPU**
  The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. The H100 GPU further extends NVIDIA's market-leading inference leadership with advancements in the NVIDIA Hopper architecture to deliver industry-leading conversational AI, speeding up large language models by 30x over the previous generation on LLMs over 175B parameters With the NVIDIA fourth generation NVLINK, H100 accelerates workloads, while the dedicated Transformer Engine supports trillion-parameter language models. NVIDIA H100 PCIe GPU configuration includes the NVIDIA AI Enterprise software suite to streamline development and deployment of AI workloads. For LLMs up to 175B parameters, systems equipped with H100 NVL GPUs can support inference on GPT3-175B with 12x more throughput in a fixed power data center than previous generation systems. For next generation trillion parameter LLMs, HGX H100 systems can scale for the highest inference performance.
- **NVIDIA L40S GPU**
  Combining NVIDIA's full stack of inference serving software with the compute capabilities of the L40S GPU provides a powerful platform for trained models ready for inference. With support for Transformer Engine, FP8, and a broad range of precisions, servers equipped with 8x L40S GPUs deliver up to 1.7x the inference performance of HGX A100 8-GPU systems.
- **NVIDIA L4 Tensor Core GPU**
  The NVIDIA Ada Lovelace L4 Tensor Core GPU delivers universal acceleration and energy efficiency for video, AI, virtual workstations, and graphics in the enterprise, in the cloud, and at the edge. With NVIDIA's AI platform and full-stack approach, L4 is optimized for video and inference at scale for a broad range of AI applications to deliver the best in personalized experiences. For AI Video pipeline applications using CV-CUDA®, servers equipped with L4 provide 120x better performance than CPU based server solutions, letting enterprises gain real-time insights to personalize content and implement cost effective smart-space solutions.

Figure 7. NVIDIA AI Enterprise for Optimized Production Inference



NVIDIA AI Enterprise for Optimized Production Inference

AI Models

Optimize for Inference

NVIDIA TensorRT

TensorRT-LLM

Deploy for Production

Triton Inference Server

284x Faster Inference than x86 — NVIDIA Grace Hopper, For LLMs and Recommenders

30x Faster Inference than A100 — NVIDIA H100, For LLM Deployment over 175B

1.7x Faster Inference than HGX A100 — NVIDIA L40S, For LLMS up to 175B in Mainstream

120x Faster AI Video than CPUs — NVIDIA L4, For AI Video

NVIDIA-Certified Systems

Deploying cutting-edge AI-enabled products and services in enterprise data centers needs computing infrastructure that provides performance, manageability, security, and scalability, while increasing operational efficiencies.

NVIDIA-Certified Systems™ enable enterprises to confidently deploy hardware solutions that securely and optimally run their modern accelerated workloads. NVIDIA-Certified Systems bring together NVIDIA GPUs and NVIDIA networking in servers, from leading NVIDIA partners, in optimized configurations. These servers are validated for performance, manageability, security, and scalability and are backed by enterprise-grade support from NVIDIA and our partners. With an NVIDIA-Certified System, enterprises can confidently choose performance-optimized hardware solutions to power their accelerated computing workloads—from the data center to the edge.

NVIDIA-Certified Systems with the GH200 Superchips, H100, L40S and L4 GPUs deliver breakthrough AI inference performance, ensuring that AI-enabled applications can be deployed with fewer servers and less power, resulting in faster insights with dramatically lower costs.

NVIDIA AI

NVIDIA AI Enterprise

Inference is where AI models are put to work and make predictions. It is a crucial process for enterprises who integrate AI into addressing questions and making evidence-based decisions. However, the complexity of maintaining security and stability of an AI software stack with increasing dependencies is a massive undertaking. A foundation AI stack consists of over 4500 open-source software that includes 3rd party and NVIDIA packages, representing 10,000 dependencies.

NVIDIA AI Enterprise, the enterprise-grade software that powers the NVIDIA inference platform, accelerates time to production with security, stability, manageability, and support. NVIDIA AI Enterprise includes proven, open-source NVIDIA frameworks, pretrained models, and development tools to streamline development and deployment of production-ready generative AI, computer vision, speech AI, data science, and more. To maintain uptime for mission-critical AI applications, NVIDIA AI Enterprise offers continuous monitoring and regular releases of security patches for critical and common vulnerabilities and exposures (CVEs), production releases that ensure API stability, management software for AI deployment at scale, and enterprise support with service-level agreements (SLAs).

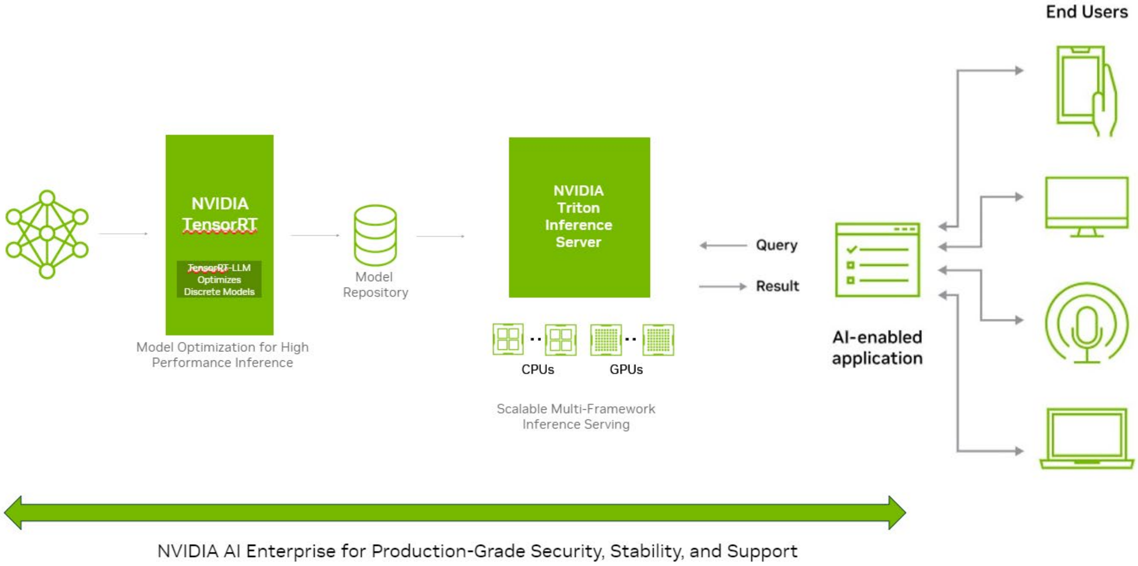
Key components of NVIDIA AI Enterprise that help optimize for AI inference performance and deployments:

- NVIDIA TensorRT™, an SDK for high-performance deep learning inference
- NVIDIA TensorRT-LLM, an SDK for high-performance deep learning inference that includes an inference optimizer and runtime GPUs deliver up to 1.7x the inference performance of HGX A100 8-GPU systems.
- NVIDIA Triton Inference Server, an open-source inference server for AI models

Inference Workflow with TensorRT and Triton

Figure 8 shows AI inference workflow with NVIDIA TensorRT and Triton inference server.

Figure 8. AI Inference Workflow with NVIDIA TensorRT and Triton Inference Server


End Users
NVIDIA TensorRT
TensorRT-LLM Optimizes Discrete Models
Model Optimization for High Performance Inference
Model Repository
NVIDIA Triton Inference Server
CPUs
GPUs
Scalable Multi-Framework Inference Serving
Query
Result
AI-enabled application
NVIDIA AI Enterprise for Production-Grade Security, Stability, and Support

Inference Optimization with TensorRT and TensorRT-LLM

As more applications use deep learning in production, demands on accuracy and performance have led to strong growth in model complexity and size. Among the most notable trends is the rise in popularity of large language models (LLMs), which are increasingly being adopted by both consumers and businesses for a variety of applications, from building creative writing chatbots for marketing teams to document summarization tools for legal teams, code writing for software development, and much more. Additionally, safety-critical applications, like those in the automotive industry, place strict requirements on throughput and latency expected from deep learning models. The same holds true for some consumer applications, including recommendation systems and conversational AI.

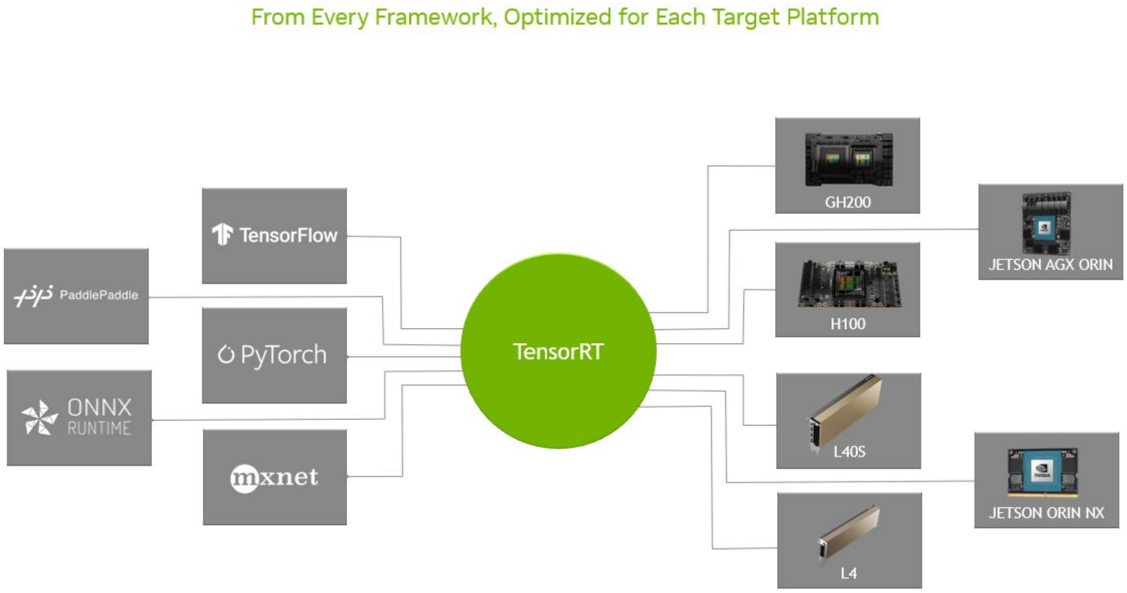
AI inference deployments that are not optimized can lead to poor utilization of infrastructure, require more servers for deployment, lead to higher operational costs and “sluggish” user experiences. For edge and embedded deployments, optimization is key for fitting models into device memory and meeting tight performance constraints.

TensorRT, TensorRT-LLM, and Triton Inference Server were developed as part of the NVIDIA AI Enterprise Suite with enterprise support from NVIDIA, to help tackle these challenges and ensure optimal inference deployment.

TensorRT

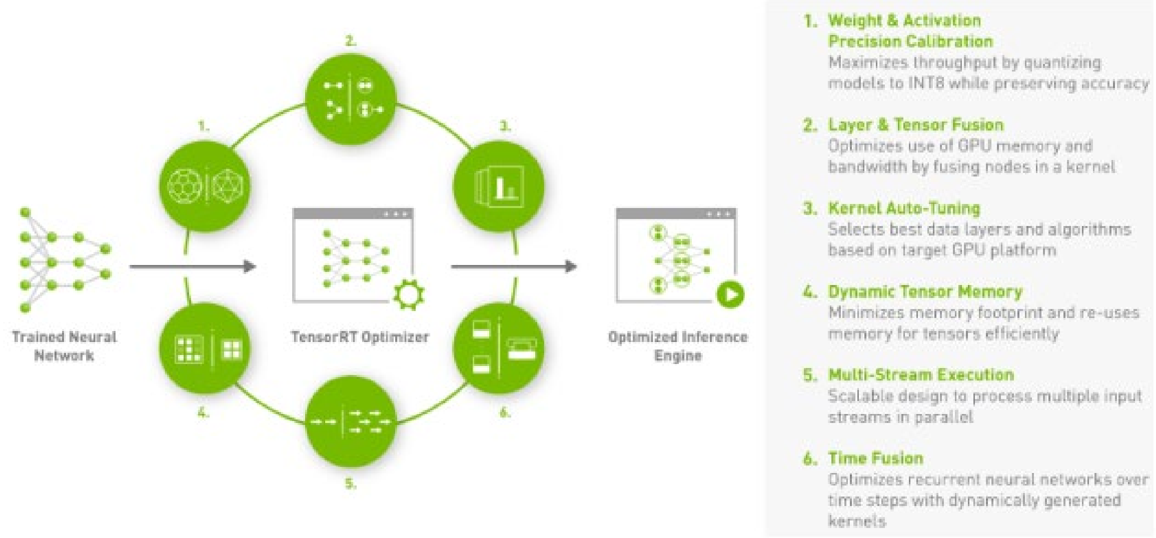
NVIDIA TensorRT™ is an SDK for high-performance deep learning inference that includes an inference optimizer and runtime. It enables developers to import trained models from all major deep learning frameworks and optimize them for deployment with the highest throughput and lowest latency, while preserving the accuracy of predictions.

Figure 9. NVIDIA TensorRT SDK



TensorRT-optimized applications perform orders of magnitude faster on NVIDIA GPUs than CPU-only platforms during inference. To realize this performance gain, TensorRT offers a range of optimizations that can be automatically applied to fine-tune trained AI models for production deployment on NVIDIA GPUs. These include combining model layers, optimizing kernel selection, and performing normalization and conversion to optimized matrix math, depending on the specified precision (FP32, FP16, FP8 or INT8), for improved latency, throughput, and efficiency. TensorRT also includes support for Sparse Tensor Cores on NVIDIA's latest architectures and Quantization-Aware Training (QAT) to achieve FP32 accuracy for INT8 inference. TensorRT's optimizations can also scale for boosting multi-GPU multi-node inference performance.

Figure 10. Multi-step Model Optimization with TensorRT



1. **Weight & Activation Precision Calibration**
   Maximizes throughput by quantizing models to INT8 while preserving accuracy
2. **Layer & Tensor Fusion**
   Optimizes use of GPU memory and bandwidth by fusing nodes in a kernel
3. **Kernel Auto-Tuning**
   Selects best data layers and algorithms based on target GPU platform
4. **Dynamic Tensor Memory**
   Minimizes memory footprint and re-uses memory for tensors efficiently
5. **Multi-Stream Execution**
   Scalable design to process multiple input streams in parallel
6. **Time Fusion**
   Optimizes recurrent neural networks over time steps with dynamically generated kernels

In addition to performance, TensorRT is designed for versatility tightly integrated with popular frameworks like TensorFlow, PyTorch, and ONNX Runtime to achieve optimized accuracy for inference. TensorRT optimizes across multiple classes of AI models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer-based models, covering a broad range of inference use cases, including computer vision, fraud detection, search, product/ad recommendation engines, as well as LLM-powered chat bots, language services, and more.

To keep up with the latest TensorRT features and developer resources, check out the TensorRT Getting Started zone.

"Hugging Face Accelerated Inference API already delivers up to 100x speedup for Transformer models served with NVIDIA A100 GPUs. With TensorRT 8, Hugging Face achieved sub 1ms inference latency on BERT." - Jeff Boudier, Product Director, Hugging Face

"TensorFlow's integration with NVIDIA TensorRT delivers up to 4.5x higher inference throughput (compared to regular GPU execution) on NVIDIA A100 GPUs. By leveraging TensorRT, TensorFlow achieves state-of-the-art performance for GPU inference." - Francois Chollet, Creator of Keras, Software Engineer, Google

TensorRT-LLM

The LLM ecosystem is innovating rapidly, developing new and diverse model architectures for new capabilities and use cases. As LLMs evolve, developers need high-accuracy results for optimal production inference deployments. However, increases in LLM size drive up the costs and complexities of optimal deployment.

That's why NVIDIA has expanded TensorRT's optimizations into LLMs. NVIDIA has been working closely with leading LLM companies including Amazon, Meta, Anyscale, Cohere, Deci, Grammarly, and many more to accelerate and optimize LLM inference. The innovations that have come out of these strategic partnerships have now been integrated into NVIDIA.

TensorRT-LLM is an open-source library that accelerates and optimizes inference performance on the latest LLMs on NVIDIA GPUs. It wraps TensorRT's deep learning compiler and includes the latest optimized kernels made for cutting-edge implementations of FlashAttention and masked multi-head attention (MHA) for the context and generation phases of LLM model execution. TensorRT-LLM also consists of pre- and post-processing steps and multi-GPU/multi-node communication primitives in a simple, open-source Python API for groundbreaking LLM inference on GPUs. It enables developers to experiment with new LLMs, offering peak performance and quick customization capabilities, without requiring deep knowledge of C++ or NVIDIA CUDA™.

All of these capabilities will help developers create customized LLMs faster and more accurately to meet the needs of virtually any industry.

Higher Performance, Better TCO, and Energy Efficient

On the GPT-J-6B, the H100 alone is 4x faster than A100, and adding TensorRT-LLM and its benefits, results in a whopping 8x boost in throughput using CNN/Daily Mail, a well-known dataset for evaluating summarization performance. The performance boost translates to over 5x lower total cost of ownership (TCO) and nearly 6X lower energy usage compared to the previous generation without TensorRT-LLM.

For larger models, like Llama 2- a popular LLM from Meta and used widely by organizations looking to incorporate generative AI workflows into their business. H100 GPUs with TensorRT-LLM can accelerate inference performance by 4.6x relative to A100 GPUs. This reduces both the TCO and the energy usage by 3X relative to A100 without TensorRT-LLM benefits.

Tensor Parallelism

Previously, developers looking to achieve the best performance for LLM inference had to rewrite and manually split the AI model into fragments and coordinate execution across GPUs. Now with TensorRT-LLM, developers can use tensor parallelism, a type of model parallelism in which individual weight matrices are split across devices. This enables efficient inference at scale - with each model running in parallel across multiple GPUs connected through NVLink and across multiple servers - without developer intervention or model changes.

In-flight Batching

The versatility of today’s LLM workloads can make it difficult to batch requests and execute them in parallel effectively - a common optimization for serving neural networks - which could result in some requests finishing much earlier than others.

To manage dynamic loads, TensorRT-LLM includes an optimized scheduling technique called in-flight batching. This takes advantage of the fact that the overall text generation process for an LLM can be broken down into multiple iterations of execution on the model.

With in-flight batching, rather than waiting for the whole batch to finish before moving on to the next set of requests, the TensorRT-LLM runtime immediately evicts finished sequences from the batch. It then begins executing new requests while other requests are still in flight. In-flight batching and the additional kernel-level optimizations enable improved GPU usage and minimally double the throughput on a benchmark of real-world LLM requests on H100 Tensor Core GPUs, helping to minimize TCO).

Transformer Engine with FP8

LLMs contain billions of model weights and activations, typically trained and represented with 16-bit floating point precision (FP16 or BF16) values where each value occupies 16 bits of memory. At inference time, however, most models can be effectively represented at lower precision, like 8-bit or even 4-bit integers (INT8 or INT4), using modern quantization techniques.

Quantization is the process of reducing the precision of a model’s weights and activations without sacrificing accuracy. Using lower precision means that each parameter is smaller, and the model takes up less space in GPU memory. This enables inference on larger models with the same hardware while spending less time on memory operations during execution.

NVIDIA H100 and L40S GPUs with TensorRT-LLM give users the ability to convert their model weights into a new FP8 format easily and compile their models to take advantage of optimized FP8 kernels automatically. This is made possible through Transformer Engine technology and done without having to change any model code.

The FP8 data format, introduced in H100 and available in L40S, enables developers to quantize their models and radically reduce memory consumption without degrading model accuracy. FP8 quantization retains higher accuracy compared to other data formats like INT8 or INT4 while achieving the fastest performance and offering the simplest implementation.

AI Inference at Scale with Triton

Extracting measurable business value from AI requires a bridge between the world of data scientists, ML researchers – who build and optimize AI models – and the world of DevOps and infrastructure/platform team, who build and maintain the production IT environments that need to run at minimum cost and maximum utilization. From right-sizing the compute needed to host the AI-enabled service, to being able to dynamically load-balance applications running on multiple servers to meet SLAs and drive the best user experiences, the path to AI inference in production has many challenges.

To bridge this gap and simplify the deployment of AI-enabled services, NVIDIA offers Triton Inference Server—an open source inference serving software—to deploy trained AI models from any framework (TensorFlow, PyTorch, ONNX, OpenVINO, XGBoost and others or a custom C++/Python framework) on any GPU- or CPU-based infrastructure from cloud to edge.

Triton Inference Server

The Triton Inference Server provides a standardized inference platform that can run multiple models concurrently on GPU servers or CPU-only servers in the public cloud, in the data center, at the edge, and in embedded devices (e.g., NVIDIA Jetson™ Orin), eliminating the need to support disparate serving solutions and maximizing CPU/GPU utilization.

Triton packs in many features like automatically finding the best model configurations (batch size, model concurrency, precision) to meet specified performance targets, dynamic batching, multi-GPU support, streaming inputs, model pipelines with business logic and advanced scheduling that help deliver high performance inference.

Table 1. Triton Inference Server Features

Utilization	Usability	Customization	Performance
Concurrent Model Execution Multiple models (or multiple instances of same model) may execute on GPU simultaneously. **Dynamic Batching** Inference requests can be batched up by the inference server to: 1) the model allowed maximum or 2)the User-defined latency SLA	**Multiple Model Format Support** - TensorRT - PyTorch Jit (.pt) - TensorFlow 1.x - GraftDef/SavedModel - TensorFlow+TensorRT 1.x GraftDef - TensorFlow + TensorRT 2.x - SavedModel - ONNX graft (ONNX Runtime) - OpenVino - RAPIDS FIL. (Forrest Inference Library) **CPU Model Inference Execution** Framework native models can execute inference requests on the CPU. **Metrics** Utilization, count, memory, and latency. **Model Controlled API** Explicitly load/unload models into and out of Triton based on changes made in the model-control configurations.	**Custom Backend for C++ and Python** Custom backends allows the user more flexibility by providing their own implementation of an execution engine through the use of a shared library. **Model Ensemble** Pipeline of one or more models and the connection of input and output tensors between those models (can be used with custom backend) **Streaming API** Built in support for streaming inputs. Accommodates stateful/sequence models that have a sequence of inputs to keep track of (speech, translation, ect.) **Decoupled Inference Servicing** Engages a model once sufficient but not all inputs are received e.g., speech recognition and synthesis.	**System/CUDA Shared Memory** Inputs/outputs needed to be passed RPC overhead to/from Triton are stored in system/CUDA shared memory. Reduces HTTP/gRPC overhead. **Library Version** Link against libtritonserver.so so that you can include all the inference swerver functionality directly in your application. **KFServing Data Plane v2 protocal** Using community standard gRPC and HTTP/REST protocol designed to be performant. Enables integration with KFServer. **Data Loading Library (DALI) Backend** Allows pre-processing and augmentation pipelines for images, videos, and speech within Triton.

To find the best Triton model configuration, developers can use Model Analyzer to find optimal parameter settings for batch size, model concurrency, and precision to deploy efficient inference. This process can optimize hardware usage, maximize model throughput, increase reliability, and allow for better hardware sizing by better managing GPU memory footprint. However, this is just one step in the end-to-end deployment process. Model Navigator automates the steps between going from a trained model to instance deployment by converting models to TensorRT, validating accuracy after TensorRT conversion, optimizing configurations with Model Analyzer, generating model config files for Triton's model repository, and finally generating a helm chart to deploy on Kubernetes.

Designed for IT, DevOps, and MLOps

Triton Inference Server simplifies the path to deploy and maintain AI models within standard production IT infrastructure. The PyTriton feature provides a simple interface to use Triton in Python code. Triton is also available as a Docker container on NGC and GitHub (updated monthly), and can integrate with Kubernetes, the container management platform for orchestration, metrics, and autoscaling. It also integrates with KServe, and public cloud-managed Kubernetes services like Amazon Elastic Kubernetes Service (EKS), Azure Kubernetes Service (AKS), and Google Kubernetes Engine (GKE), for an end-to-end AI workflow.

Triton Inference Server also exports Prometheus metrics for monitoring and supports the standard HTTP/gRPC interface to connect with other applications like load balancers. It's also integrated in MLOps platforms like Amazon SageMaker, Azure Machine Learning, Google Vertex AI and many others. All these integrations help the production team deploy a streamlined inference-in-production platform with lower complexity, higher visibility into resource utilization, and scalability.

Triton Inference Server Examples

To learn more about running inference on Triton Inference Server, explore the following open source examples across model types: Computer Vision,

AI Inference Acceleration in the Cloud

The NVIDIA AI inference platform is available in the cloud from all major cloud service partners (CSPs), including AWS, Azure, Google Cloud, and Oracle Cloud Infrastructure. NVIDIA AI Enterprise, including NVIDIA Triton and TensorRT, is available on major cloud marketplaces as an enterprise-ready deployment solution. In addition, NVIDIA's full software stack is available from the NGC catalog, NVIDIA's hub of GPU-optimized AI, high-performance computing (HPC), and data analytics software that simplifies building an AI workflow. These containers can easily be run on cloud instances.

Additionally, NVIDIA GPU platforms, including NVIDIA H100 Tensor Core GPUs, are also available through all major Cloud Service Providers (CSPs). With access to NVIDIA GPUs in the cloud, you can provision the right-sized GPU resources for your inference workloads on-demand with flexible pay-as-you-go pricing options. NVIDIA GPUs are also widely supported in Managed Kubernetes services offered by cloud service providers (CSPs), offering the flexibility to rent the GPU resources needed and automatically scale up or down as AI inference workload requirements change. NVIDIA Triton Inference Server is also integrated with cloud AI platforms like Amazon SageMaker, Azure ML, Google Vertex AI and Alibaba PAI-EAS.

AI Inference Acceleration at the Edge

From portable medical devices to automated delivery drones, intelligent edge solutions demand advanced inference to solve complex problems. But these use cases can't rely on network connections back to the data center or the public cloud due to latency constraints or the need to function in a disconnected environment. Edge computing is tailored for real-time, always-on solutions that have low-latency requirements. Always-on solutions are sensors or other pieces of infrastructure that are constantly working or monitoring their environments.

Faster insights can equate to saving time, costs, and even lives. That's why enterprises in every industry are looking to tap into the data generated from billions of IoT sensors. NVIDIA edge computing solutions bring together NVIDIA-Certified Systems with NVIDIA H100, L40S, L4, and GH200, embedded platforms with NVIDIA Jetson™ and NVIDIA Orin™, NVIDIA Triton Inference Server, TensorRT, and Fleet Command, a cloud service that securely deploys, manages, and scales AI applications across distributed edge infrastructure.

Application-specific Frameworks

Given the diversity of AI use cases across industries, a one size fits all approach to accelerated AI inference is far from optimal. To that end, NVIDIA has created application-specific frameworks to accelerate developer productivity and address the common challenges of deploying AI within those specific applications. Figure 11 provides a quick overview of a few of these.

To help convey how NVIDIA's application specific frameworks accelerate the path to developing and deploying AI in production, we will zoom into four use cases: generative AI / LLMs, conversational AI, recommender systems, and computer vision, including the challenges inherent within each and how to address them using a full-stack approach.

Figure 11. Application-specific Frameworks to Accelerate Developer Productivity

NVIDIA Clara | Healthcare Healthcare application framework for AI-powered imaging, genomics, and the development and deployment of smart sensors. It includes full-stack GPU-accelerated libraries, SDKs, and reference applications. Learn More	**NVIDIA Isaac™ | Robotics** A toolkit that includes building blocks and tools to accelerate robot developments that require the increased perception and navigation features enabled by AI. Learn More
NVIDIA Driveworks | Automotive An SDK for autonomous vehicle (AV) software development, with an extensive set of capabilities, including the processing modules, tools, and frameworks for advanced AV development. Learn More	**Aerial | Telco** An application framework for building high performance, so NVIDIA Maxine is a suite of GPU-accelerated AI SDKs and cloud-native microservices for deploying AI features that enhance audio, video, and augmented reality effects in real time.ftware defined, NVIDIA Maxine is a suite of GPU-accelerated AI SDKs and cloud-native microservices for deploying AI features that enhance audio, video, and augmented reality effects in real time.ss networks (RAN) to address increasing demand for a more flexible, scalable and Open RAN compliant network. Learn More
NVIDIA Morpheus | Cybersecurity An application framework that enables cybersecurity developers to create optimized AI pipelines for filtering, processing, and classifying large volumes of real-time data. Learn More	**NVIDIA Maxine | Video Conferencing NVIDIA** NVIDIA Maxine is a suite of GPU-accelerated AI SDKs and cloud-native microservices for deploying AI features that enhance audio, video, and augmented reality effects in real time. Learn More
NVIDIA Riva | Speech and Translation AI Multilingual speech and translation AI for building and deploying fully customizable voice interfaces for real-time applications such as chatbots, intelligent virtual assistants and digital avatars, and live captioning for broadcast events and video-conferencing meetings. . Learn More	**NVIDIA Merlin | Recommender Systems** An open source framework for building high performing modern recommender systems at any scale. Learn More
NVIDIA Metropolis | AI Video Analytics An application framework that simplifies the development, deployment, and scaling of AI-enabled video analytics applications from edge to cloud. Learn More	**NVIDIA NeMo | Generative AI** End-to-end, cloud-native enterprise framework for developers to build, customize, and deploy generative AI models with billions of parameters. Learn More

Generative AI / LLMs

Generative AI has become a transformative force, empowering organizations spanning every industry to achieve unparalleled levels of productivity, elevate customer experiences, and deliver superior operational efficiencies.

Large language models (LLMs) are the brains behind generative AI. Access to incredibly powerful and knowledgeable foundation models, like Llama and Falcon LLM, has opened the door to amazing opportunities. However, these models lack the domain-specific knowledge required to serve enterprise use cases.

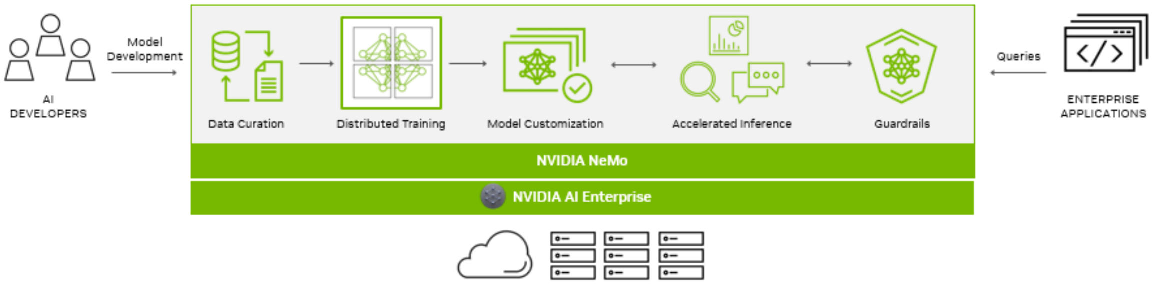
Developers have three choices for powering their generative AI applications:

- Use a pre-trained LLM: the easiest lift is to use a foundation model, as it works very well for use cases that relies on general-purpose knowledge.
- Customize a pre-trained LLM: pre-trained models customized with domain-specific knowledge, task-specific skills, and connected to enterprises' knowledge bases perform tasks and provide responses based on the latest proprietary information.
- Develop an LLM from scratch: organizations with specialized data (for example, models catering to regional languages) cannot use pre-trained foundation models and must build their models from scratch.

NVIDIA NeMo - Build, Customize, and Deploy LLMs

NVIDIA NeMo is an end-to-end, cloud-native framework for building, customizing, and deploying generative AI models. It includes training and inferencing frameworks, guardrails, and data curation tools, for an easy, cost-effective, and fast way to adopt generative AI (see Figure 12).

Figure 12. NVIDIA NeMo



Exploding Transformer-Based Language Model Size and Complexity

See Figure 13 for an illustration of the evolutionary tree for transformer-based language model size and complexity.

Figure 13. Evolutionary Tree for Transformer-Based Language Model Size and Complexity

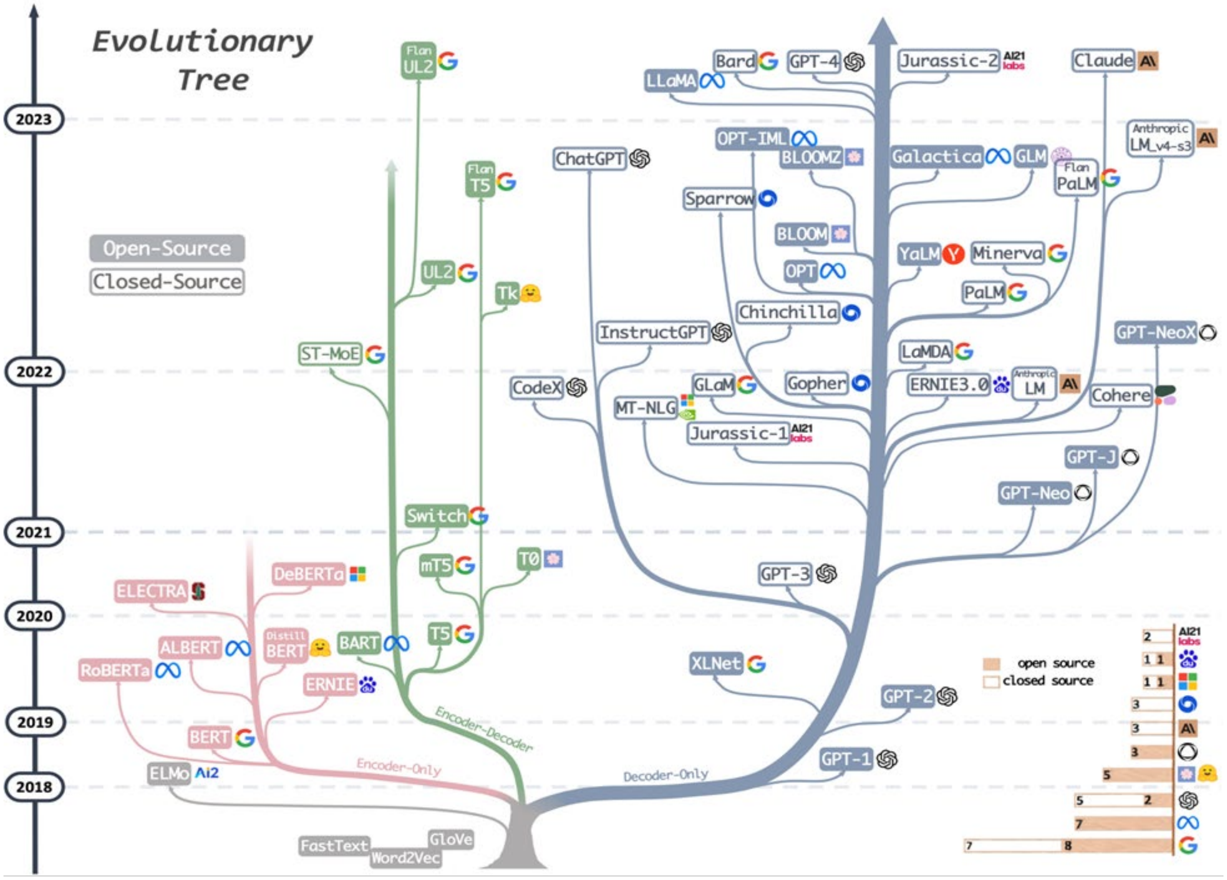


Image from Mooler0410/LLMsPracticalGuide

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., ... Hu, X. (2023). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2304.13712

The domain of natural language processing has witnessed an unprecedented escalation in Transformer-based language model size and complexity. These models, like BERT, GPT, and their various derivatives, have experienced exponential growth in parameters, with models like GPT-4 reaching the trillions. This explosion in size and complexity necessitates substantial computational resources, thereby elevating the complexity and cost of inference. The NVIDIA AI inference platform enables developers to meet the substantial hardware and software demands posed by these massive LLMs all while maintaining costs, starting with Conversational AI.

Conversational AI

Conversational AI is the application of machine learning to develop language-based applications that allow humans to interact naturally with devices, machines, and computers using speech. In the last few years, deep learning has improved the state-of-the-art in conversational AI and offered superhuman accuracy on certain tasks. Deep learning has also reduced the need for deep knowledge of linguistics and rule-based techniques for building language services, which has led to widespread adoption across industries like retail, healthcare, and finance.

However, the technology behind Conversational AI is complex, involving a multi-step process that requires a massive amount of computing power and computations that must happen in less than 300 milliseconds to deliver an optimal user experience. .

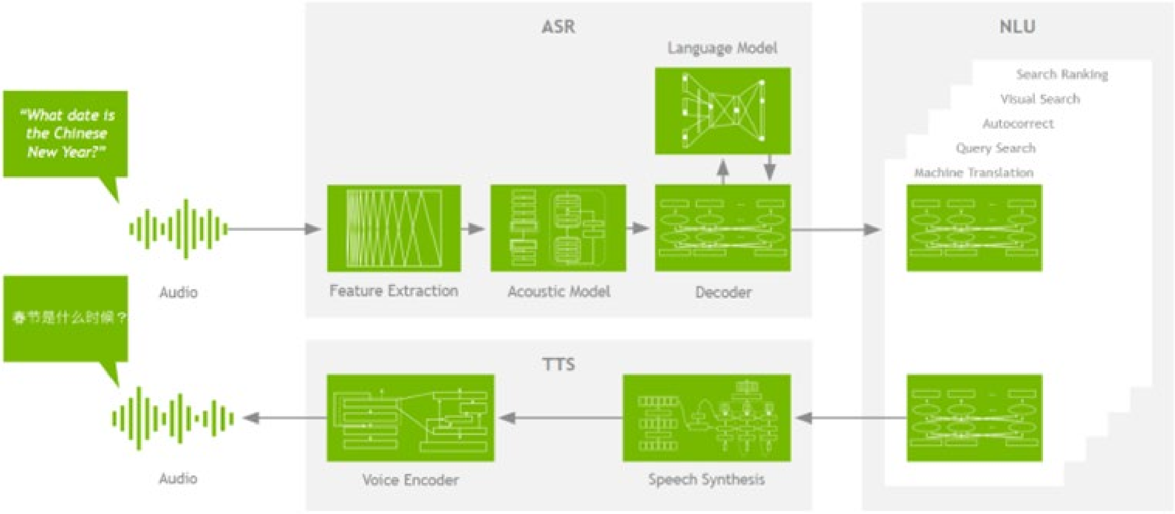
> **Automatic Speech Recognition (ASR):** Speaking into a device like a smartphone, having the system understand the words, and converting the audio into text.

> **Natural Language Processing (NLP) or Natural Language Understanding (NLU):** When spoken content is parsed for meaning so that an AI service can search for and return a relevant and useful response.

> **Text-to-Speech (TTS) with voice synthesis:** When the answer is then converted into an audio signal that speaks the answer, but is processed to sound like a human voice, including pitch changes, timbre, and cadence.

Conversational AI consists of Speech AI + NLP, where Speech AI includes ASR and TTS. Within these three steps, however, there can be over a dozen deep learning models that are connected to deliver a single response back to the end user (as shown in Figure 14).

Figure 14. Overview of a Conversational AI Pipeline



A Conversational AI pipeline has three sections: Automatic speech recognition, Natural Language understanding, and Text-to-Speech. Within each of these sections there can be multiple natural networks off working together to quickly deliver an answer to a posed question.

Delivering Conversational AI Services: What It Takes

The parallel processing capabilities and Tensor Core architecture of NVIDIA GPUs allow for higher throughput and scalability when working with complex language models – enabling record-setting performance for both the training and inference of BERT.

To deliver conversational AI services in production, several language models need to work together to generate a response for a single query in less than 300 milliseconds. Meeting this tight end-to-end latency budget requires the latency for a single model, within the conversational AI pipeline, to be only a few milliseconds. NVIDIA GPUs can deliver the latency required. This makes it practical to use the most advanced transformer-based language models in production.

The conversational AI domain continues to be an intensive focus area for AI researchers and these neural networks and datasets keep growing at significant rates. What is not changing is the requirement to deliver conversational AI in a way that's actually conversational. This means initial questions are understood, relevant and useful answers are delivered in real-time and follow-up questions are inferred in the context of the questions that preceded them. It also means the voice speaking the answers feel natural and human.

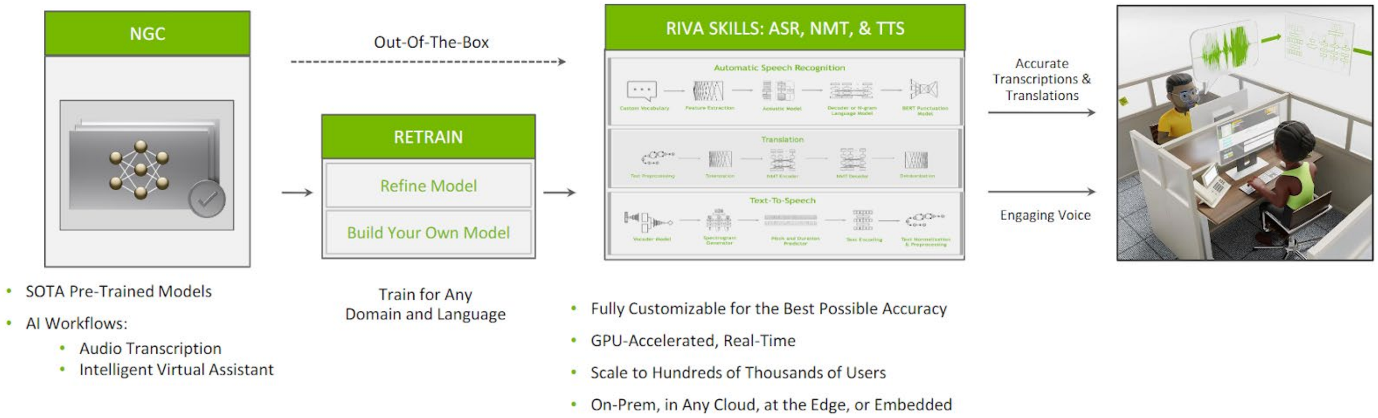
Hence, the platform needed to deliver a conversational AI service must be both performant and programmable so that AI developers can accelerate time to solution, build new services, and continuously push the boundaries of conversational AI.

NVIDIA Riva – Build and Deploy Multilingual Speech and Translation AI Applications

Speech and translation AI are powering conversational AI applications across the globe. As speech- and translation-based applications are adopted globally, solutions need to interact with humans across many languages. Speech and translation AI apps need to understand industry specific jargon, translate speech in different languages, and respond naturally in real-time.

NVIDIA Riva is a GPU-accelerated speech and translation SDK with automatic speech recognition (ASR), text-to-speech (TTS), and neural machine translation (NMT) skills for conversational applications. Riva offers out-of-the-box (OOTB), state-of-the-art speech and translation models that are trained for millions of hours on thousands of hours of audio data. The ASR, TTS, and NMT pipelines are optimized for real-time performance, with inference running far below the natural conversation threshold of 300 milliseconds (see Figure 15).

Figure 15. ASR, TTS, and NMT Pipelines


NGC
Out-Of-The-Box
RETRAIN
Refine Model
Build Your Own Model
RIVA SKILLS: ASR, NMT, & TTS
Automatic Speech Recognition
Custom Vocabulary
Feature Extraction
Acoustic Model
Decoder or N-gram Language Model
BERT Punctuation Model
Translation
Text Preprocessing
Tokenization
NMT Encoder
NMT Decoder
Detokenization
Text-To-Speech
Vocoder Model
Spectrogram Generator
Pitch and Duration Predictor
Text Encoding
Text Normalization & Preprocessing
Accurate Transcriptions & Translations
Engaging Voice
• SOTA Pre-Trained Models
• AI Workflows:
• Audio Transcription
• Intelligent Virtual Assistant
Train for Any Domain and Language
• Fully Customizable for the Best Possible Accuracy
• GPU-Accelerated, Real-Time
• Scale to Hundreds of Thousands of Users
• On-Prem, in Any Cloud, at the Edge, or Embedded

NVIDIA Riva provides state-of-the-art models, fully accelerated pipelines, and tools to easily add speech and translation AI capabilities to real-time applications like intelligent virtual assistants, call center agent assists, and video conferencing. Riva components are fully customizable, so you can adapt the applications for your use case and industry and deploy them in all clouds, on-premises, at the edge, and on embedded systems. In addition, Riva offers packaged AI workflows for audio transcriptions and intelligent virtual assistants.

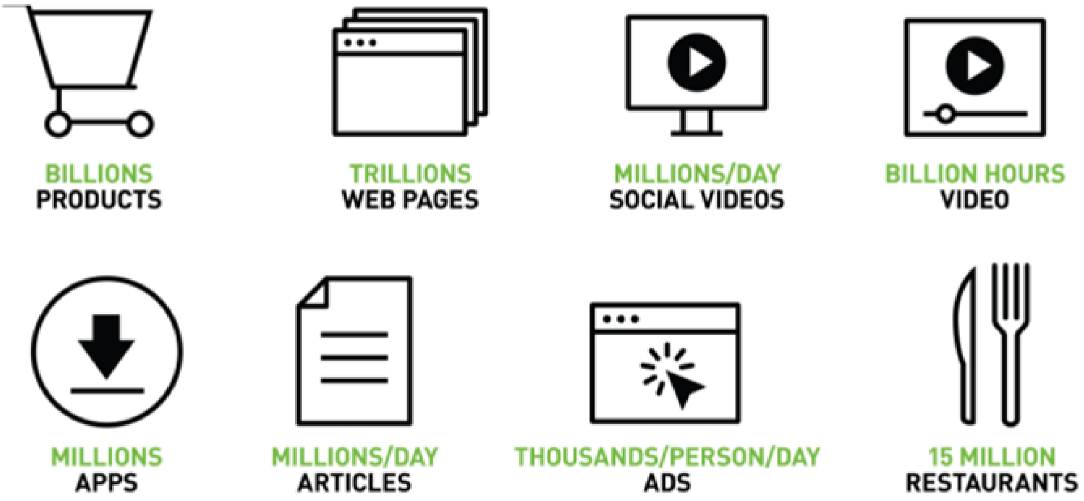
Under the hood, Riva applies powerful NVIDIA TensorRT optimizations to models, configures the NVIDIA Triton Inference Server for model serving, and exposes the models as a service through a standard API that can easily be integrated into applications. For domain-specific data, users can fine-tune Riva speech and translation models with the NVIDIA NeMo to achieve the best possible accuracy.

NVIDIA Riva is part of the NVIDIA AI Enterprise software platform, and can be purchased for production deployments with unlimited usage on all clouds, access to NVIDIA AI experts, and long-term support. Riva containers are also available for free for development for 90 days to all members of the NVIDIA Developer Program on NGC. Organizations looking to deploy Riva-based applications can apply to try Riva on NVIDIA Launchpad, a program that provides short-term access to enterprise-grade NVIDIA hardware and software via a web browser.

Recommender Systems

It is simply impossible for enterprises to connect billions of users in the world with the products, services, and even expertise – among trillions of things that matter to them. Recommender systems learn user preferences and “recommend” relevant consumer products from the exponential number of available options, significantly improving conversion. From Amazon’s shopping recommendations to Netflix’s content suggestions, recommender systems can influence every action consumers take, from visiting a web page to using social media for shopping (see Figure 16).

Figure 16. Recommender Systems Connect Trillions of Users to Millions of Products and Services



As the growth in the volume of data available to power these systems accelerates, data scientists and ML engineers are increasingly turning to more traditional ML methods to highly expressive DL models to improve the quality of their recommendations. In the future, they will rely upon an ensemble of tools, techniques, and frameworks to deploy at scale.

Recommenders work by collecting information, such as what movies you tell your video streaming app you want to see, ratings and reviews you’ve submitted, purchases you’ve made, and other actions you’ve taken in the past. These data sets are often huge and tabular, with multiple entries of metadata, including product and customer interactions. They can be hundreds of terabytes in size and require massive compute, connectivity, and storage performance to train effectively.

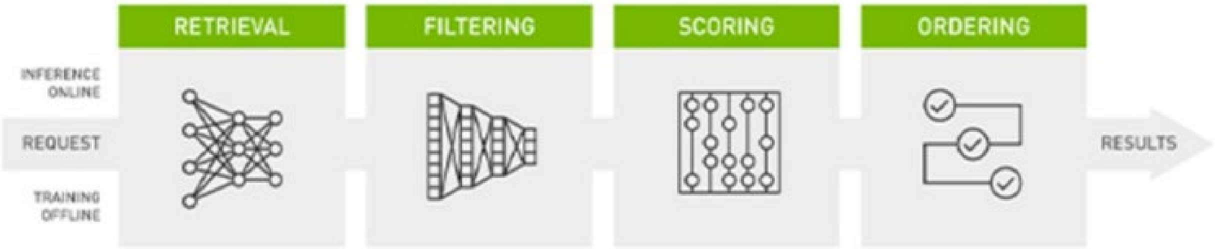
With NVIDIA GPUs, you can exploit data parallelism through columnar data processing instead of traditional row-based reading designed initially for CPUs. This provides higher performance and cost savings. Current DL-based models for recommender systems like DLRM, Wide and Deep (W&D), Neural Collaborative Filtering (NCF), Variational AutoEncoder (VAE) are part of the NVIDIA GPU-accelerated DL model portfolio that covers a wide range of network architectures and applications in many different domains beyond recommender systems, including image, text, and speech analysis.

NVIDIA Merlin – Build Large-Scale Recommender Systems for Production Readiness

NVIDIA Merlin is an open-source framework for building high-performing recommender systems at scale. It empowers data scientists, machine learning engineers, and researchers to build high-performing recommenders at scale. Merlin includes libraries, methods, and tools that streamline the building of recommenders by addressing common preprocessing, feature engineering, training, inference and deploying to production challenges.

Merlin components and capabilities are optimized to support the retrieval, filtering, scoring and ordering of hundreds of terabytes of data, all accessible through easy-to-use APIs (see Figure 17). With Merlin, better predictions, increased click-through rates, and faster deployment to production are within reach.

Figure 17. Recommender Workflow with NVIDIA Merlin



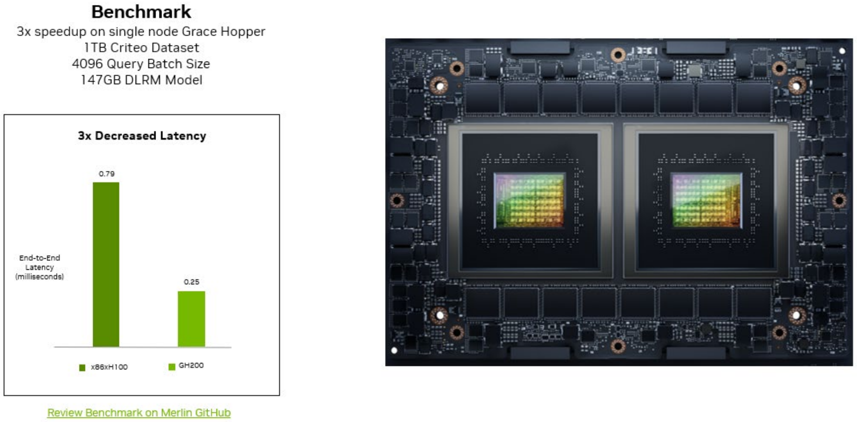
From ingesting and training to deploying GPU-accelerated recommender systems in production, NVIDIA Merlin libraries accelerates the recommender workloads across the entire pipeline. It offers open-source components to simplify building, training, optimizing, and deploying a production-quality recommender pipeline.

- **Merlin NVTabular**
  Merlin NVTabular is a feature engineering and preprocessing library designed to effectively manipulate terabytes of recommender system datasets and significantly reduce data preparation time.
- **Merlin HugeCTR**
  Merlin HugeCTR is a deep neural network framework designed for recommender systems on GPUs. It provides distributed model-parallel training and inference with hierarchical memory for maximum performance and scalability.
- **Merlin Transformers4Rec**
  Merlin Transformers4Rec is a library that streamlines the building of pipelines for session-based recommendations. The library makes it easier to explore and apply popular transformer architectures when building recommenders.

- **Merlin Distributed Training**
  Merlin provides support for distributed training across multiple GPUs. Components include Merlin SOK (SparceOperationsKit) and Merlin Distributed Embeddings (DE). TensorFlow (TF) users are empowered to use SOK (TF 1.x) and DE (TF 2.x) to leverage model parallelism to scale training.
- **Merlin Systems**
  Merlin Systems is a library that eases new model and workflow deployment to production. It enables ML engineers and operations to deploy an end-to-end recommender pipeline with 50 lines of code.

Merlin can deliver exceptional performance. A single Grace Hopper node enables a 3x speedup for a 147 GB DLRM model compared to a Hopper x86 based on the Criteo 1TB dataset (see Figure 18).

Figure 18. Merlin HPS Performance Benchmark
3x Speedup for Continued Unmatched Performance



Computer Vision

Image-centric use cases have been at the center of the DL phenomenon, going back to AlexNet, which won the ImageNet competition in 2012, signaling what we refer to as the “Big Bang” of DL and AI. Computer vision has a broad range of applications, including smart cities, agriculture, autonomous driving, consumer electronics, gaming, healthcare, manufacturing, and retail services to name a few. In all these applications, computer vision is the technology that enables the cameras and vision systems to perceive, analyze, and interpret information in images and videos.

Modern cities are dotted with video cameras that generate a massive amount of data every day. Deep learning-based computer vision is the best way to turn this raw video data into actionable insights, and NVIDIA GPU-based inference is the only way to do it in real-time. To enable developers, NVIDIA offers a variety of different GPU-accelerated libraries, SDKs and application frameworks for every stage of the computer vision pipeline, including codecs, data processing, training and inference from the edge to the cloud.

Codecs

nvJPEG/nvJPEG2000/nvTIFF - Provides high-performance GPU-accelerated libraries for encoding and decoding JPEG, JPEG2000, TIFF type of images.

Video Codec - Offers a comprehensive set of APIs including high-performance tools for hardware accelerated video encoding, transcoding, and decoding videos on Windows and Linux

VPF - Provides easy-to-use python bindings for HW accelerated video decoding, encoding, transcoding and GPU-accelerated color space and pixel format conversions.

Data Processing

CV-CUDA - An open-source, low-level library that easily integrates into existing custom CV applications to accelerate video and image processing.

DALI - A holistic framework for loading, decoding, and processing data while offering operators for augmenting 2D and 3D image, video, and audio data. Providing the convenient transfer of data processing between training and inference.

NPP - A comprehensive set of image/video/signal processing ops in a closed source library for efficient processing of large images (e.g., 20kx20k pixels). Performs up to 30x faster than CPU-only implementations.

Optical Flow - Detect and track objects in successive video frames, interpolate, or extrapolate video frames to improve smoothness of video playback and compute flow vectors

VPI - A low-level library providing CV operators for use in AI/CV pipelines running on embedded devices like Jetson or dGPUs and in thermal and energy constrained environments. Also supports CPUs, GPUs, PVA, VIC and OFA.

Training

NVIDIA Omniverse Replicator - A core extension of the Omniverse platform, replicator allows developers to bootstrap the model training process by generating photo-realistic, physically-aware training datasets

NVIDIA TAO - Create highly accurate, customized, and enterprise-ready AI models with this low-code toolkit and deploy them on any device - GPUs, CPUs, and MCUs—whether at the edge or in the cloud.

End-to-End Computer Vision Frameworks:

NVIDIA Metropolis is an end-to-end application framework that makes it easier for developers to combine common video cameras and sensors with AI-enabled video analytics to provide operational efficiency and safety applications across a broad range of industries, including retail analytics, city traffic management, airport operations, and automated factory inspections.

DeepStream SDK, a foundational layer of the NVIDIA Metropolis framework, is a streaming analytic toolkit for building AI-powered applications. It takes the streaming data as input – from a USB/CSI camera, video from file, or streams over RTSP – and uses AI and computer vision to generate insights from pixels for a better understanding of the environment.

Fraud Detection

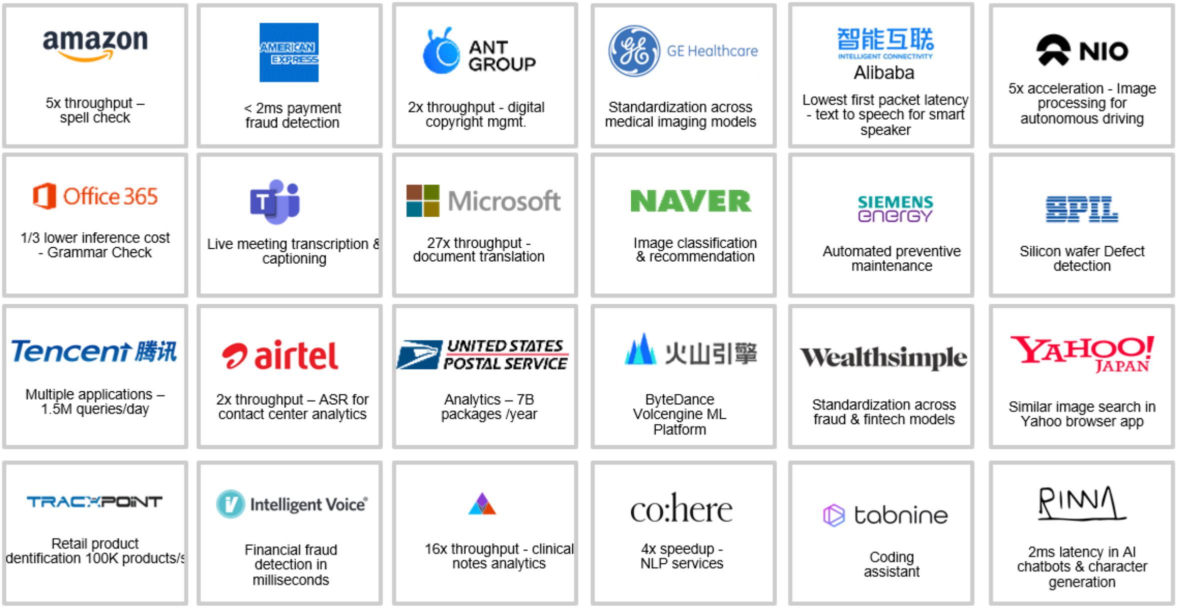
Transaction fraud is a multi-billion-dollar problem. Detecting true fraud is critical, but traditional systems have historically generated many more false-positive than true-fraud signals. Now, advanced machine learning and deep learning techniques are improving detection and, at the same time, drastically cutting false-positive rates. AI is revolutionizing multi-trillion-dollar industries and powering the growth of nations around the world to make a significant impact on an organization's bottom line.

Leveraging NVIDIA's full-stack platform, leading banks are deploying enterprise AI capabilities that reduce operational costs, drive higher revenues, improve customer satisfaction, and create long-term competitive advantage. NVIDIA Triton Inference Server and the NVIDIA TensorRT SDK, part of NVIDIA AI Enterprise software, can help with easy deployment, running, and scaling of AI models to deliver meaningful outcomes in financial services. Accelerated inference can benefit modern-day inference pipelines across cloud and on premises.

World-Leading Inference Performance

The NVIDIA AI inference platform is already powering a range of cutting-edge customer applications in production today, including predictive healthcare, online product and content recommendations, voice-based search, contact center automation, fraud detection, and others deployed across on-prem, cloud, and at the edge. Thousands of companies worldwide, like the ones shown in Figure 19, are using the NVIDIA AI inference platform to transform their businesses.

Figure 19. NVIDIA AI Inference Platform Customer Success


amazon
5x throughput – spell check
AMERICAN EXPRESS
< 2ms payment fraud detection
ANT GROUP
2x throughput - digital copyright mgmt.
GE Healthcare
Standardization across medical imaging models
智能互联
INTELLIGENT CONNECTIVITY
Alibaba
Lowest first packet latency - text to speech for smart speaker
NIO
5x acceleration - Image processing for autonomous driving
Office 365
1/3 lower inference cost - Grammar Check
Live meeting transcription & captioning
Microsoft
27x throughput - document translation
NAVER
Image classification & recommendation
SIEMENS energy
Automated preventive maintenance
SPIL
Silicon wafer Defect detection
Tencent 腾讯
Multiple applications – 1.5M queries/day
airtel
2x throughput – ASR for contact center analytics
UNITED STATES POSTAL SERVICE
Analytics – 7B packages /year
火山引擎
ByteDance Volcengine ML Platform
Wealthsimple
Standardization across fraud & fintech models
YAHOO! JAPAN
Similar image search in Yahoo browser app
TRACXPOINT
Retail product dentification 100K products/s
Intelligent Voice
Financial fraud detection in milliseconds
16x throughput - clinical notes analytics
co:here
4x speedup - NLP services
tabnine
Coding assistant
RINNA
2ms latency in AI chatbots & character generation

MLPerf Inference

The full-stack approach has also been instrumental in NVIDIA achieving top-place finishes in MLPerf Inference, an industry-standard benchmark that measures AI inference performance across a broad range of use cases like large language models, computer vision, medical imaging, natural language, and recommender systems. The NVIDIA AI platform delivers this leadership performance using a combination of the world's most advanced GPUs with Tensor Core technology and ongoing software optimizations, NVIDIA TensorRT, TensorRT-LLM and Triton Inference Server for AI inference deployments in the data center, in the cloud, or at the edge.

Recent MLPerf Inference benchmarks highlight the exceptional performance and versatility of the NVIDIA AI platform. NVIDIA HGX H100 systems with eight H100 GPUs delivered the highest throughput on every data center inference test, continuing NVIDIA's record of performance leadership in AI inference. Grace Hopper Superchips (GH200), which combines a Hopper GPU with a Grace CPU providing more memory, bandwidth, and power shifting capabilities between the CPU and GPU, also led across all MLPerf's data center tests including inference for LLMs, computer vision, speech recognition, medical imaging, and recommendation systems. Additionally, L4 GPUs ran the full range of workloads and delivered up to 6x more performance than high-end CPUs. At the edge, the NVIDIA Jetson Orin system-on-module showed significant performance improvements of up to 84% driven by new software that takes advantage of the latest version of the chip's programmable vision accelerator (PVA), the NVIDIA Ampere architecture GPU, and a dedicated deep learning accelerator.

The NVIDIA AI Inference Platform has continuously evolved over the last several years and inference performance has scaled significantly. Continuous software optimizations, such as NVIDIA TensorRT-LLM, bring more performance to existing platforms, delivering ongoing ROI.

Many of the optimizations and advances that enabled these MLPerf Inference results are available from the NGC Catalog container and the NVIDIA GitHub repository . You can find the latest MLPerf Inference results for the NVIDIA AI Inference Platforms on the NVIDIA MLPerf webpage.

Conclusion

Deployment and integration of trained AI models in production remains a complex challenge, both for application developers and the platform/infrastructure teams supporting them. Taking AI from prototype to production to revenue demands overcoming issues related to diverse frameworks, different model architectures, underutilized infrastructure for inference, and lack of standardized implementations across multiple deployment environments that cause many enterprise AI projects to fail. Additionally, these AI-powered services will be deployed across a wide range of industries, each with its own particular requirements and constraints. So, an effective AI inference acceleration platform is much more than just the hardware.

The NVIDIA AI Inference Platform provides a full stack approach to address these challenges and supports a wide range of AI inference use cases through a combination of architectural optimization, reduced precision, and comprehensive developer solutions to power through high-batch workloads, and low latency to deliver optimal real-time performance in time-constrained applications. It also offers the versatility to accelerate rapidly evolving AI model architectures and a unified solution to maximize performance and utilization, as well as to simplify AI inference deployments within on-prem enterprise data centers, in the public cloud, at the edge, or even in embedded devices. With enterprise support available with NVIDIA AI Enterprise, organizations can focus on harnessing the business value of enterprise AI.

Learn how you can benefit from the NVIDIA AI Inference Platform and take your AI projects from prototype to production. Try it today on NVIDIA LaunchPad.

NVIDIA LaunchPad

NVIDIA LaunchPad is a universal proving ground, offering extensive testing of the latest NVIDIA enterprise hardware and software. It expedites short-term trials, facilitates long-term proofs of concepts (POCs), and accelerates the development of both managed and standalone services.

Begin with a tailored development environment or explore a wide selection of hands-on labs. These labs offer guided experiences across use cases ranging from AI and data science to 3D design and infrastructure optimization. Enterprises can access essential hardware and software stacks through private hosted infrastructure.

LaunchPad resources deployed in partner data centers are globally accessible directly from NVIDIA. They feature mainstream NVIDIA-Certified servers running comprehensive NVIDIA software stacks. This platform supports developers, designers, and IT professionals by accelerating the creation and deployment of modern data-intensive applications. Quick testing and prototyping pave the way for confident decisions in software and infrastructure deployment for production workflows.

There are several labs like image classification, chatbots and scaling data science with Triton Inference Server. LaunchPad covers a wide spectrum of use cases, from complex AI development and training on one end to low-latency data analytics and inference on the other.

Access step-by-step guided labs for inference with ready-to-use hardware, software, sample data, and applications on NVIDIA LaunchPad.

Enterprise Support

As AI initiatives move into the production stage, the need for a trusted, scalable support model for enterprise becomes vital to ensuring AI projects stay on track. NVIDIA Enterprise Support is offered through NVIDIA AI Enterprise and includes:

- Broad Platform Support: Full enterprise grade support for multiple deployment options across on-prem, hybrid and multi-cloud environments
- Access to NVIDIA AI Experts: Local business hours(e.g. 9 a.m. - 5 p.m.) support includes guidance on configuration and performance, and escalations to engineering
- Priority notification: Latest security fixes and maintenance releases
- Long term support: Up to 3-years for designated software branches

Customized support upgrade option: Designated Technical Account Manager (TAM) or Business Critical support for 24x7 live agent access with a One hour response time for severe issues

Request for a free 90-day NVIDIA AI Enterprise evaluation license that includes access to the NVIDIA enterprise-grade inference platform and NVIDIA enterprise support.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA TensorRT, NVIDIA Triton, NVIDIA Merlin, NVIDIA's Hopper, NVIDIA Grace, NVIDIA Ampere , BlueField, DGX, NVIDIA Grace Hopper, NVIDIA-Certified Systems, NVIDIA Jetson, NVIDIA CUDA, NVIDIA Jetson, and NVIDIA Orin are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com