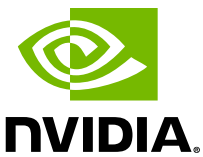
TECHNICAL OVERVIEW

NVIDIA AI FOR GPU-ACCELERATED DEEP LEARNING INFERENCE



THE AI PROTOTYPE TO PRODUCTION GAP IN THE ENTERPRISE

Artificial intelligence (AI) continues to drive breakthrough innovation across industries, including consumer internet, healthcare and life sciences, financial services, retail, manufacturing, and supercomputing. As researchers push the boundaries of what's possible in computer vision, speech, natural language processing (NLP), and recommender systems, state-of-the-art AI models continue to rapidly evolve and expand in size, complexity, and diversity. Training these AI models to convergence on a specified accuracy level and customizing them for your unique applications is a computationally-intensive, complex, and iterative process, where most time is spent during the research and prototype phase for enterprise AI projects.

However, for AI to have the utmost impact and deliver business results, these trained AI models need to be integrated within applications and deployed on production IT systems—on-premise, in the cloud, or at the edge—to "infer" things about new data that it's presented with based on its training. 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 user requests? This is why AI inference requires acceleration.

Operationalizing AI models within enterprise applications also poses a number of 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 AI model architectures for today's usages, as well as emerging usages that continue to evolve. In addition to infrastructure for deployment, there's a growing need to address the challenge of managing, monitoring, and scaling models trained in multiple frameworks, handling different types of inference query types, like batch, streaming and ensemble, and supporting multiple environments from edge to cloud.

NVIDIA GPUs and performance-optimized solution stack power a broad range of AI applications in production today, such as personalized shopping experiences, contact center automation, voice assistants, chatbots, visual search, and even assisted medical diagnostics. This full-stack AI platform from NVIDIA is accessible wherever you need to build and deploy—on-premise data centers, public cloud, desktops, laptops, and the world's fastest supercomputers.

In this paper, we will 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 Platforms, with an emphasis on data center deployments.

END-TO-END DEEP LEARNING WORKFLOW OVERVIEW

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. 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.

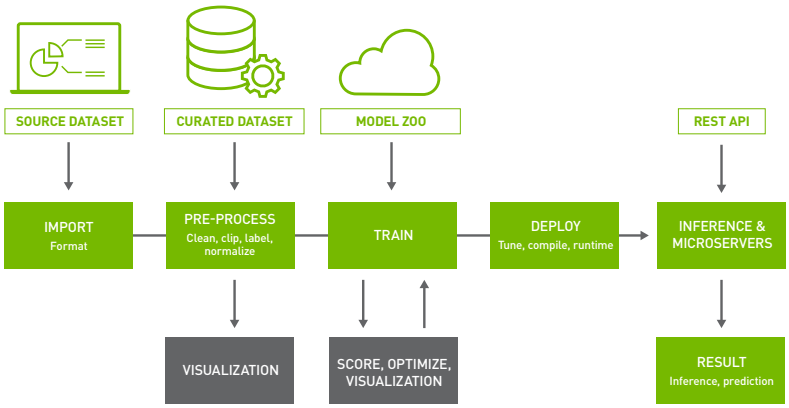


Figure 1: End-to-end deep learning workflow, from training to inference.

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.

In this paper, we will focus 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 converged AI models that are ready to take to production.

Accelerate Deep Learning Training

Building the most accurate AI models to solve your business problems is a complex, iterative, and computationally-intensive process. This is where data scientists and machine learning engineers spend a big chunk of the end-to-end AI workflow.. Also, today's state-of-the-art (SOTA) AI models have billions of parameters and the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time, compounding the time to solution.

With NVIDIA Tensor Core technology, TensorFloat-32 (TF32), and Automatic Mixed Precision support, NVIDIA platforms have been architected to accelerate deep learning training, at scale, for a wide range of applications, including conversational AI, recommendation systems, and computer vision. To enable

developers to kick-start their AI projects and harness the computational power of NVIDIA GPUs for model building and training, NVIDIA offers the NGC™ Catalog—a hub of GPU-optimized pre-trained AI models, enterprise-grade containers, and industry-specific SDKs.

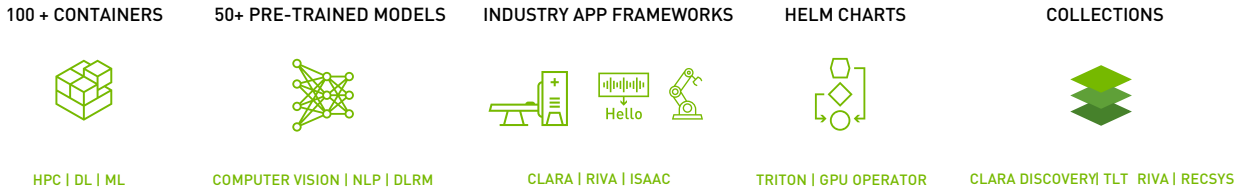


Figure 2: Simplify deep learning, machine learning and HPC workflows with GPU-optimized software from NGC Catalog

The NGC catalog provides a range of resources that meet the needs of data scientists, developers, and researchers with varying levels of AI expertise. These include:

- GPU-optimized deep learning frameworks (TensorFlow, PyTorch, MXNet, and others) to accelerate model building, training, and validation.
- State-of-the-art, pre-trained AI models, detailed code scripts with step-by-step instructions, and helper scripts for a variety of common AI tasks.
- End-to-end application-specific frameworks: NVIDIA Clara™ for healthcare, NVIDIA Merlin for recommendation systems, NVIDIA Riva for conversational AI, and more.
- Tools like NVIDIA TAO Toolkit, which enables developers to fine-tune on high quality NVIDIA pre-trained models, using only a fraction of the data and speeding up AI development by 10X.

AI INFERENCE – TRAINED MODEL TO REAL SERVICE

Converged and 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.

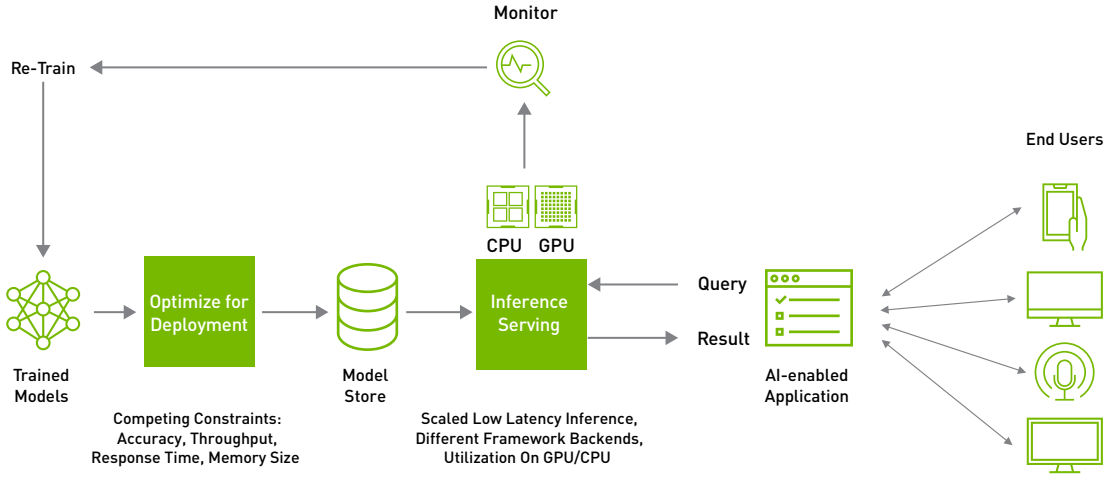


Figure 3: Challenges of Deploying Trained AI Models into Production

The Evolving Inference Landscape

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, or in remote disconnected environments to inside small, embedded devices.

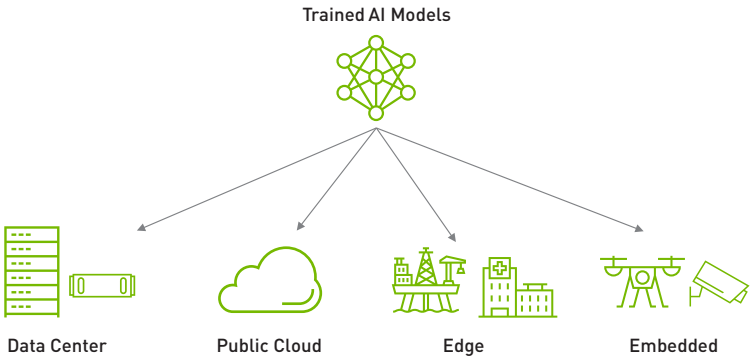


Figure 4: Diverse products and services demand diverse deployment environments for AI

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.

Inference Performance of AI Models

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 an accelerator 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 taken 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.
- **PROGRAMMABILITY/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, and beyond.
- **EFFICIENCY:** Another important attribute of accelerated AI inference is the economies 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.

The Challenge of AI Inference Deployments at Scale

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. Hence, beyond looking at inference performance on a per-model basis, 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 frameworks, handling different inference query types (real-time, batch, streaming, and ensemble, for example) and optimizing across multiple deployment platforms like CPUs and 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 INFERENCE PLATFORM: THE FULL-STACK APPROACH

NVIDIA's inference platform delivers the performance, efficiency, and responsiveness critical to powering the next generation of AI products and services—in the cloud, in the data center, at the network's edge, or in embedded devices. 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.

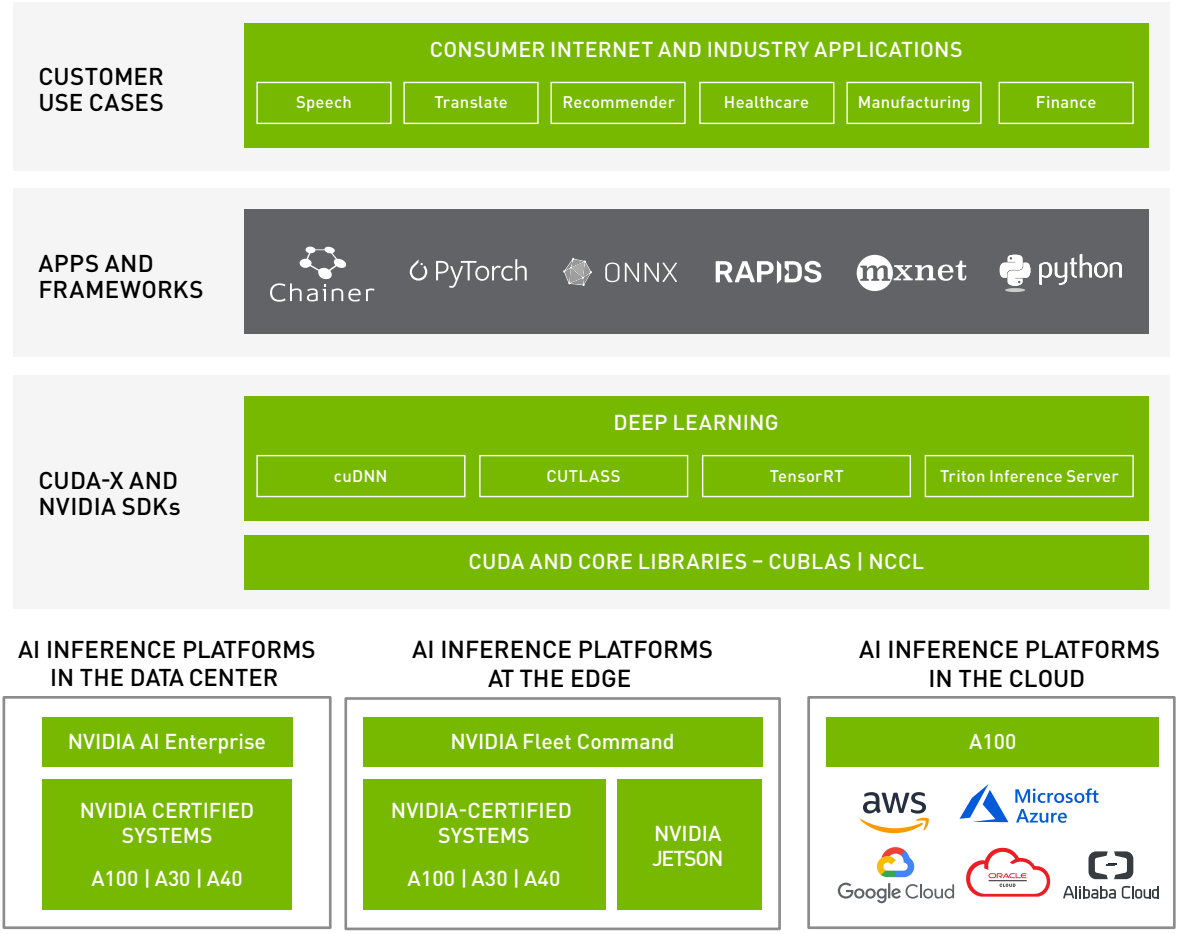


Figure 5: NVIDIA AI Inference Platform accelerates a wide array of usages, supports all frameworks and is available wherever you need to deploy AI - Data Center, Public Cloud and Edge

NVIDIA-Certified Systems for Enterprise Data Centers

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—both in smaller configurations and at scale.

NVIDIA-Certified Systems with the NVIDIA A100, A30, and A40 Tensor Core 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 A100 Tensor Core GPU: The A100 Tensor Core GPU delivers the next giant leap in our accelerated data center platform, providing unprecedented acceleration at every scale. It brings 10X more inference performance versus our previous generation, and third-generation Tensor Core technology that enables new levels of precision and acceleration. A breakthrough feature called Multi-GPU Instance (MIG) makes A100 an ideal inference accelerator, as it enables a single A100 to be partitioned into seven instances, where different neural networks can be run in each instance. A100 can additionally accelerate inference for sparse networks using a new feature called structural sparsity.

In addition to its market-leading inference capabilities, the NVIDIA A100 GPU also offers best-in-class training, high performance computing (HPC), and data analytics performance.

NVIDIA A30 Tensor Core GPU: The NVIDIA A30 Tensor Core GPU combines fast memory bandwidth and low power in a PCIe form factor, and leverages the Ampere architecture's groundbreaking features to optimize inference workloads. It accelerates a full range of precisions, from FP64 to TF32 and INT4. Supporting up to four MIG instances per GPU, A30 allows multiple networks to operate simultaneously in secure hardware partitions with guaranteed quality of service (QoS). And structural sparsity support delivers up to 2X more performance on top of A30's other inference performance gains.

NVIDIA A40 Tensor Core GPU: The NVIDIA A40 GPU is an evolutionary leap in performance and multi-workload capabilities from the data center, combining best-in-class professional graphics with powerful compute and AI acceleration to meet today's design, creative, and scientific challenges. Driving the next generation of virtual workstations and server-based workloads, NVIDIA A40 brings state-of-the-art features for ray-traced rendering, simulation, virtual production, and more to professionals anytime, anywhere.

AI Inference Acceleration in the Cloud

NVIDIA GPU platforms, including NVIDIA A100 Tensor Core GPUs, are also available globally through all major Cloud Service Providers (CSPs) like Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Oracle Cloud Infrastructure (OCI), and others. 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.

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 A100, A30, and A40 GPUs, embedded platforms with NVIDIA® Jetson™, AI software, and Fleet Command, a turnkey management service that allows enterprises to harness the power of AI at the edge.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud native suite of AI and data science applications and frameworks optimized and exclusively certified by NVIDIA to run on VMware vSphere with mainstream NVIDIA-Certified Systems. It includes key enabling technologies and software from NVIDIA for the rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud. NVIDIA AI Enterprise is licensed and supported by NVIDIA.

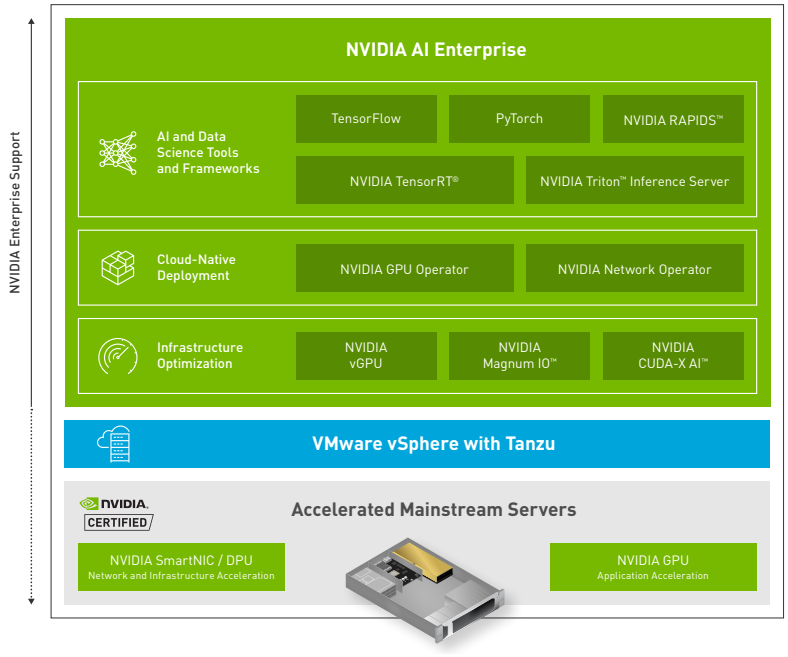


Figure 6: The NVIDIA AI Enterprise suite includes the applications, frameworks, and tools used by AI researchers, data scientists, and developers.

The NVIDIA AI Enterprise Suite is certified to run on mainstream 1U/2U servers with A100 and A30 GPUs. Through the joint development with VMware, IT teams can accelerate the speed at which developers can build, deploy, and scale AI-enabled applications on the same VMware vSphere

infrastructure they've already invested in, and deliver enterprise-class manageability, security, and availability. Performance optimizations enable workloads running in GPU accelerated virtual machines to achieve near bare metal performance for AI training and inference. For example, inference workloads in a virtualized environment running on the NVIDIA A100 GPU achieved up to 266x better performance over a CPU-only server when running a natural language processing (NLP) model, which is similar performance gains to running in bare metal.

Two key components of NVIDIA AI Enterprise that help optimize for AI inference performance and deployments at scale include NVIDIA TensorRT™ and NVIDIA Triton™ Inference Server, as shown in the AI inference workflow diagram below.

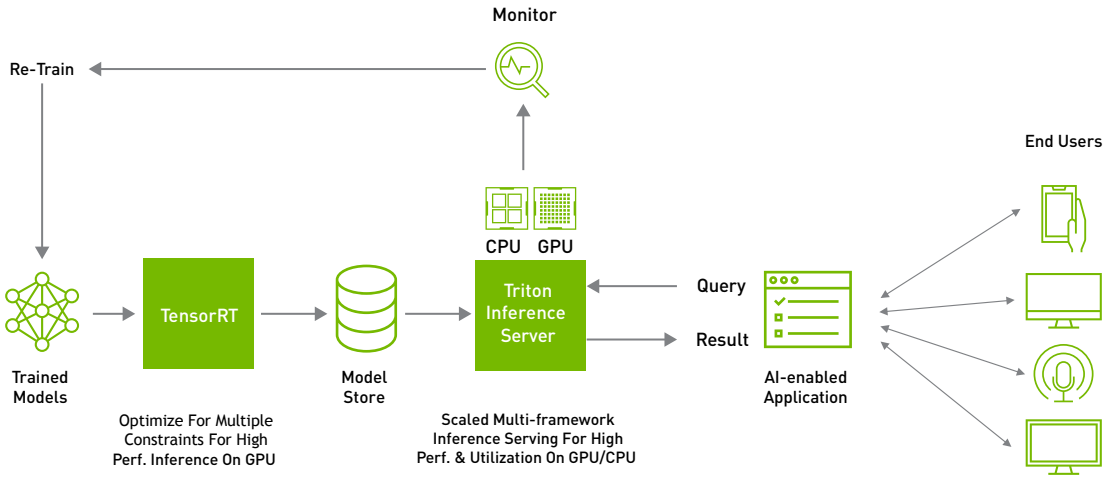


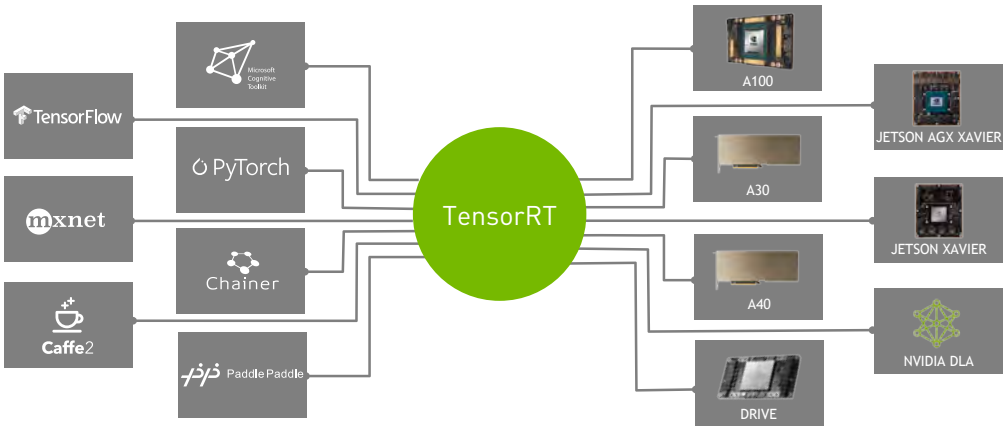
Figure 7: Accelerate path to production deployments with NVIDIA TensorRT and NVIDIA Triton Inference Server

Inference Optimization with TensorRT

As more applications use deep learning in production, demands on accuracy and performance have led to strong growth in model complexity and size. 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.

Leaving performance on the table for AI inference leads to poor utilization of infrastructure, more servers for deployment, 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.

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.



5

Figure 8: NVIDIA TensorRT accelerates inference of networks trained in any major deep learning framework, and deploys to a wide array of NVIDIA GPU accelerators.

TensorRT-optimized applications perform up to 40X 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 or INT8), for improved latency, throughput, and efficiency.

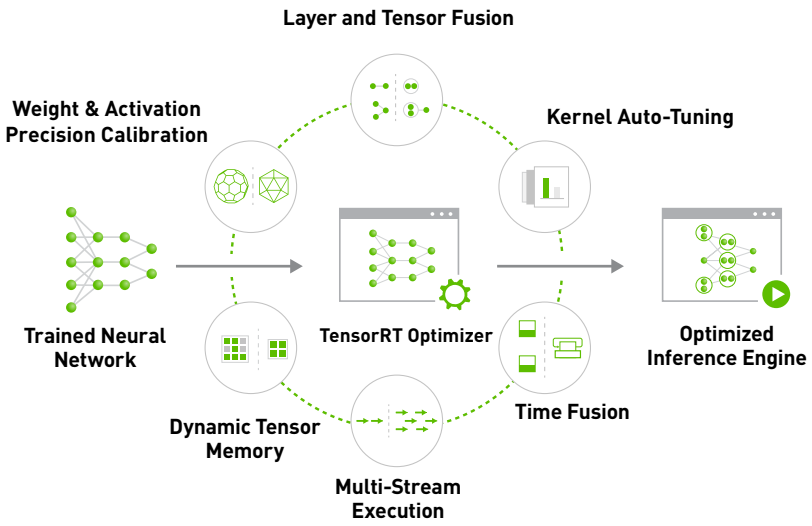


Figure 9: TensorRT features include fusion of layers and tensors, and kernel auto-tuning that deploys architecture-specific kernels to run optimally on that particular platform.

The latest transformer optimizations in TensorRT slash inference latency for BERT-Large, a 340 million parameter model for natural language understanding (NLU), down to 1.2 milliseconds—a major stride towards making production deployment of real-time conversational AI services a reality for a broad range of customers—cloud to edge. Other recent enhancements include

support for Sparse Tensor Cores on NVIDIA Ampere architecture GPUs and quantization-aware training (QAT) to achieve FP32 accuracy for INT8 inference.

In addition to performance, TensorRT is designed for versatility, optimizing 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, chat bots, language services, and more. TensorRT is tightly integrated with popular frameworks like TensorFlow, PyTorch, and ONNX Runtime to achieve optimized performance for inference.

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

AI Inference at Scale with Triton Inference Server

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 managers, who 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, NVIDIA TensorRT, PyTorch, ONNX Runtime, OpenVINO, RAPIDS™, Forest Inference Library (FIL), or a custom C++/Python framework) from local or public cloud storage on any GPU- or CPU-based infrastructure.

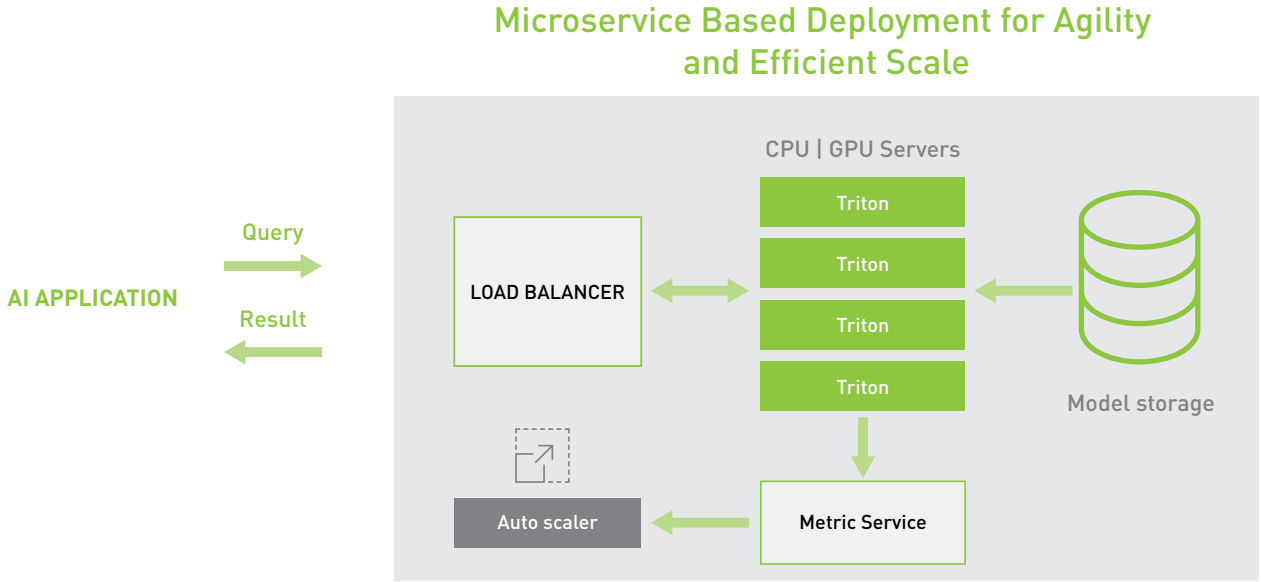


Figure 10: High-level Triton Inference Server Architecture

High-Performance Inference on CPUs and GPUs

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), 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, concurrent models) to meet specified performance targets, dynamic batching, multi-GPU support, ragged input batching for streaming inputs, and advanced scheduling that help deliver high performance inference. It can also automatically convert a trained AI model from any framework to TensorRT, optimizing for performance on specific deployment targets using the Triton Model Navigator feature.

Designed for IT, DevOps, and MLOps

Triton Inference Server simplifies the path to deploy and maintain AI models within standard production IT infrastructure. Available as a Docker container, Triton integrates with Kubernetes, the container management platform for orchestration, metrics, and autoscaling. It also integrates with Kubeflow, KFServing, 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, Seldon, and ClearML. All these integrations help IT deploy a streamlined inference-in-production platform with lower complexity, higher visibility into resource utilization, and scalability.

The NVIDIA TensorRT SDK and Triton Inference Server are both available as part of the NVIDIA AI Enterprise Suite, which is optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems

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. While each of these industry-specific platforms has its unique attributes, they support NVIDIA TensorRT and Triton Inference Server to achieve optimal inference performance for their particular tasks. Here's a quick overview of a few of these:

NVIDIA CLARA | HEALTHCARE

> A healthcare application framework for AI-powered imaging and genomics that 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 developements 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 >

NVIDIA AERIAL | TELCO

> An application framework for building high performance, software defined, cloud native 5G applications to address increasing consumer demand.

LEARN MORE >

NVIDIA MAXINE | VIDEO CONFERENCING

> An SDK with state-of-the-art features for developers to build virtual collaboration and content creation solutions, including video conferencing and streaming applications.

LEARN MORE >

NVIDIA MERLIN | RECOMMENDER SYSTEMS

> An open source framework for building large scale deep learning recommender systems, from ingesting and training to deploying a production-quality pipeline.

LEARN MORE >

NVIDIA RIVA | CONVERSATIONAL AI

> A GPU-accelerated SDK for building multimodal conversational AI applications like virtual assistants, multi-user diarization, and call center assistants that deliver real-time performance on GPUs.

LEARN MORE >

NVIDIA METROPOLIS | INTELLIGENT VIDEO ANALYTICS

> An application framework, set of developer tools, and partner ecosystem for transforming data from trillions of AI and IoT devices into valuable insights.

LEARN MORE >

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

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. Typically, the conversational AI pipeline, in real-time speech applications, consists of three stages:

- **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.

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 11).

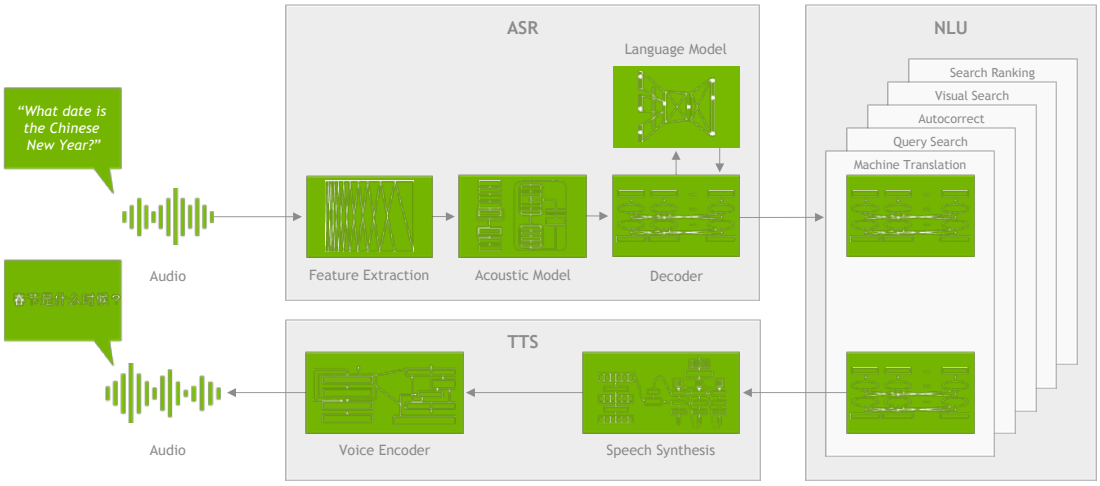


Figure 11: Overview of a conversational AI pipeline

Exploding Transformer-Based Language Model Size and Complexity

True conversational AI is a voice assistant that can engage in human-like dialogue, capturing context and providing intelligent responses. Such AI models are massive and highly complex. Recent breakthroughs in both training and inference on language networks like the Bidirectional Encoder Representations from Transformers, known more commonly as BERT, have demonstrated superhuman levels of accuracy in NLU. Since BERT was initially released by Google in 2018, researchers have built upon its capabilities to continue refining both performance and accuracy

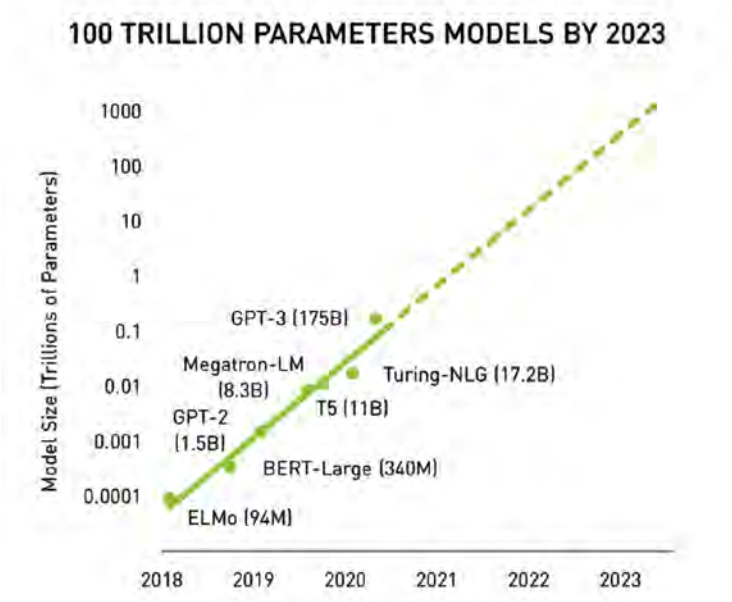


Figure 12: Trend of State-of-the-art NLP model sizes with time

Researchers at NVIDIA also released Megatron, a PyTorch-based framework for training giant language models based on the transformer architecture. Using Megatron, NVIDIA researchers efficiently train very large language models—from one billion parameters all the way to one trillion parameters—using both model and data parallelism, achieving an almost-perfect linear scaling of the NVIDIA GPUs required to train these massive models. Megatron GPT2 achieves state-of-the-art accuracy across multiple speech benchmarks, as seen on the RACE Leaderboard, which tracks NLP model accuracy.

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. Even highly optimized CPU code results in a processing time of more than 40 milliseconds. NVIDIA GPUs can deliver 40X higher inference

performance than CPU-only platforms. 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 isn't 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 feels 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 Conversational AI Applications

Given the complexity of multi-stage, multi-network conversational AI pipelines, deploying a service with conversational AI can seem daunting. To make this process easier and maximize the performance benefits of NVIDIA GPUs for training and inference, NVIDIA offers pre-trained conversational AI models and developer toolkits to customize and deploy end-to-end pipelines.

NVIDIA Riva is a GPU-accelerated SDK for building multimodal conversational AI applications that use an end-to-end deep learning pipeline. Developers at enterprises can easily fine-tune state-of-art-models on their data to achieve a deeper understanding of their specific context and optimize for inference to offer end-to-end real-time services that run in less than 300 milliseconds (ms) and deliver 7X higher throughput on GPUs compared with CPUs.

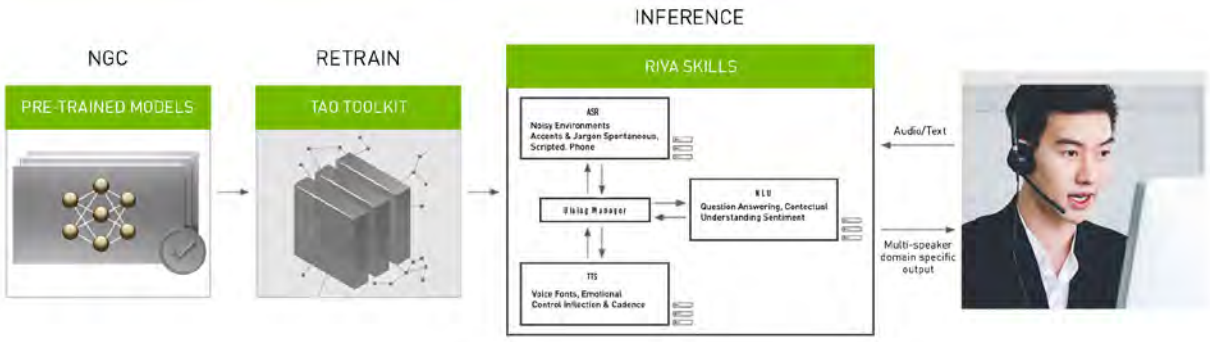


Figure 13: The Riva SDK includes pre-trained conversational AI models, the NVIDIA TAO Toolkit, and optimized end-to-end skills for speech, vision, and natural language processing (NLP) tasks

The Riva SDK includes pre-trained conversational AI models and optimized end-to-end skills for speech, vision, and natural language processing (NLP) tasks. Fusing vision, audio, and other sensor inputs simultaneously provides capabilities such as multi-user, multi-context conversations in applications like virtual assistants, multi-user diarization (the process of partitioning an input audio stream into homogeneous segments), and call center assistants.

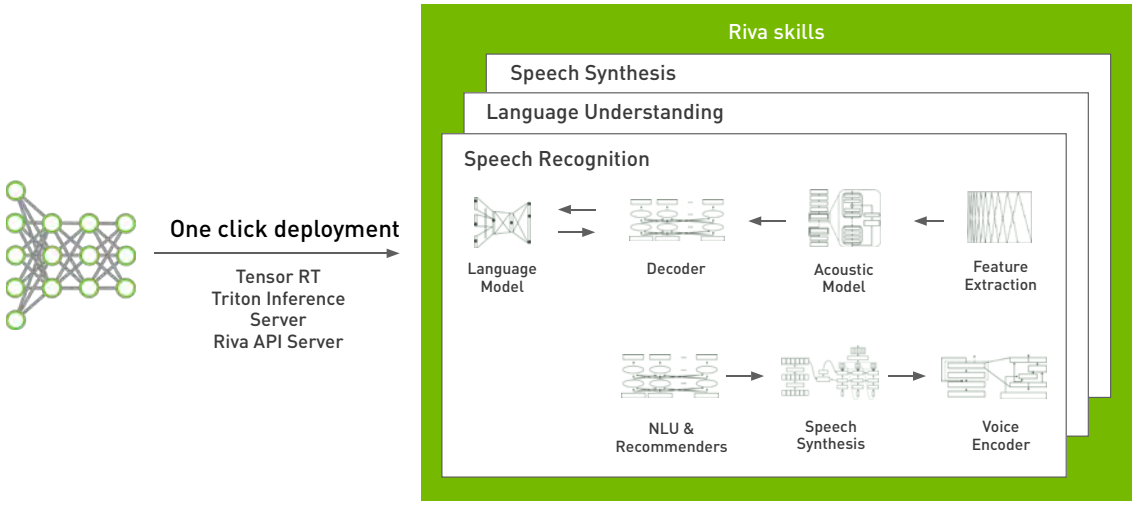


Figure 14: Riva AI skills

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 be easily integrated into applications.

RECOMMENDER SYSTEMS

Understanding consumer behavior has never been more critical for enterprises because it's simply impossible for the billions of users in the world to connect with the products, services, even expertise—among hundreds of billions of things—that matter to them. Recommender systems help 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 usual social media for shopping. On some of the largest commercial platforms, recommendations account for as much as 30% of revenue, which can translate into billions of dollars in sales.

Figure 15: Recommender systems connect billions of users to millions of products and services.

As the growth in the volume of data available to power these systems accelerates, data scientists are increasingly turning from more traditional ML methods to highly expressive DL models to improve the quality of their recommendations.

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

NVIDIA Merlin™ is an open-source framework that empowers data scientists, machine learning engineers, and researchers to build large-scale deep learning recommender systems. Merlin includes libraries, methods, and tools that democratize building deep learning recommenders by addressing common preprocessing, feature engineering, training, and inference challenges. Each component of the Merlin pipeline is optimized to support hundreds of terabytes of data, all accessible through easy-to-use APIs.

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

- 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 is a deep neural network training framework designed for recommender systems. It provides distributed training with model-parallel embedding tables and data-parallel neural networks across multiple GPUs and nodes for maximum performance.
- **TensorRT and Triton:** Leverage TensorRT and Triton Inference Server, within the Merlin framework to optimize models for inference and deploy recommender systems efficiently on GPUs by maximizing throughput with the right combination of latency and GPU utilization.

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 to build computer vision-related applications from edge to cloud.

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.

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 the edge.

Actively Detect Diseases in 145M Hearts per Year

GE Healthcare

Identify Trends in Over 300B Pins for Better Search Results

Pinterest

Award-Winning Customer Service

T-Mobile

Office Grammar Checker Reduced Cost by 70%

Microsoft

Real-Time Analytics on 7B Packages per Year

United States Postal Service

Intelligent Search with SOTA NLU for 1.28 Users

WeChat

Enhanced Real-Time Fraud Detection

American Express

The full-stack approach has also helped ensure that NVIDIA finishes top-place in MLPerf Inference, an industry-standard benchmark that measures AI inference performance across a broad range of use cases like 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 Multi-Instance GPUs (MIG), powerful and scalable interconnect technologies, and ongoing software optimizations, in NVIDIA TensorRT and Triton Inference Server for AI inference deployments in the data center, in the cloud, or at the edge.

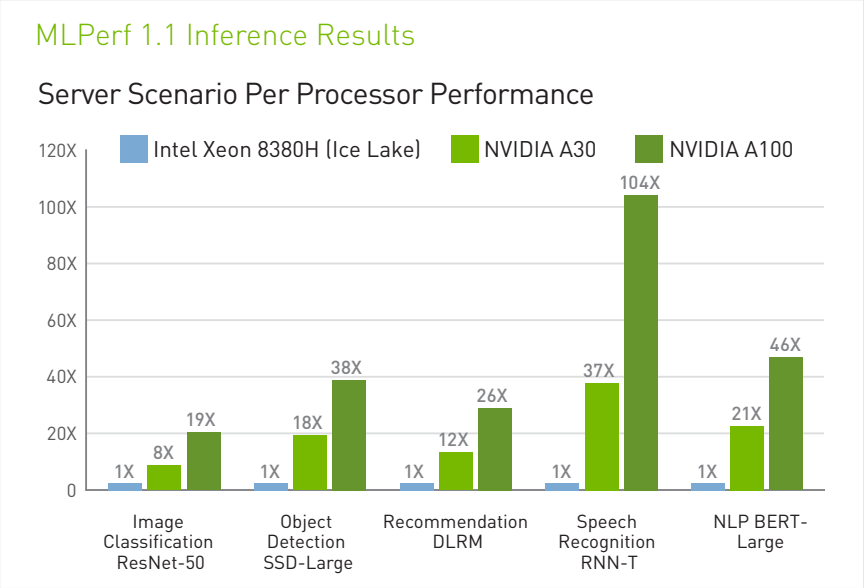


Figure 16: The above comparisons show relative performance on a per-chip basis, normalized to CPU. NVIDIA delivers up to 104X more inference performance than CPU-based platforms.NVIDIA delivers up to 104X more inference performance than CPU-based platforms.

	INTEL XEON 8380 (ICE LAKE)	NVIDIA A30	NVIDIA A100
Image Classification ResNet-50	1,713	14,502	32,505
Object Detection SSD-Large	25	447	948
Recommendation DLRM	10,123	125,066	287,833
Speech Recognition RNN-T	125	4,625	13,002
NLP BERT-Large	70	1,438	3,224

Table 1: Raw data showing per-chip inference performance across all workloadsNVIDIA delivers up to 104X more inference performance than CPU-based platforms.

MLPerf v1.1 Inference Closed; Per-accelerator performance derived from the best MLPerf results for respective submissions using reported accelerator count in Data Center Offline and Server. Qualcomm AI 100: 1.1-057 and 1.1-058, Intel Xeon 8380: 1.1-023 and 1.1-024, NVIDIA A30: 1.1-43, NVIDIA A100 (Arm): 1.1-033, NVIDIA A100 (X86): 1.1-047. MLPerf name and logo are trademarks. See www.mlcommons.org for more information.

The NVIDIA AI Inference Platform has continuously evolved over the last several years and inference performance has scaled by nearly 190X in the last five years. Continuous software optimizations bring more performance to existing platforms, delivering ongoing ROI. 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 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 about much more than just the hardware

The NVIDIA AI Inference Platform is uniquely capable of addressing 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.

Find out more about how you can benefit from the NVIDIA AI Inference Platform and take your AI projects from prototype to production:

NVIDIA