# Deploying AI Models with Speed, Efficiency and Versatility

## *Inference on NVIDIA's AI Platform*

Whitepaper

WP-11144-001_v01 | October 2022

# Table of Contents

# List of Figures

# List of Tables

# NVIDIA's Inference Solution

## 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 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 user requests? This is why AI inference requires accelerated computing.

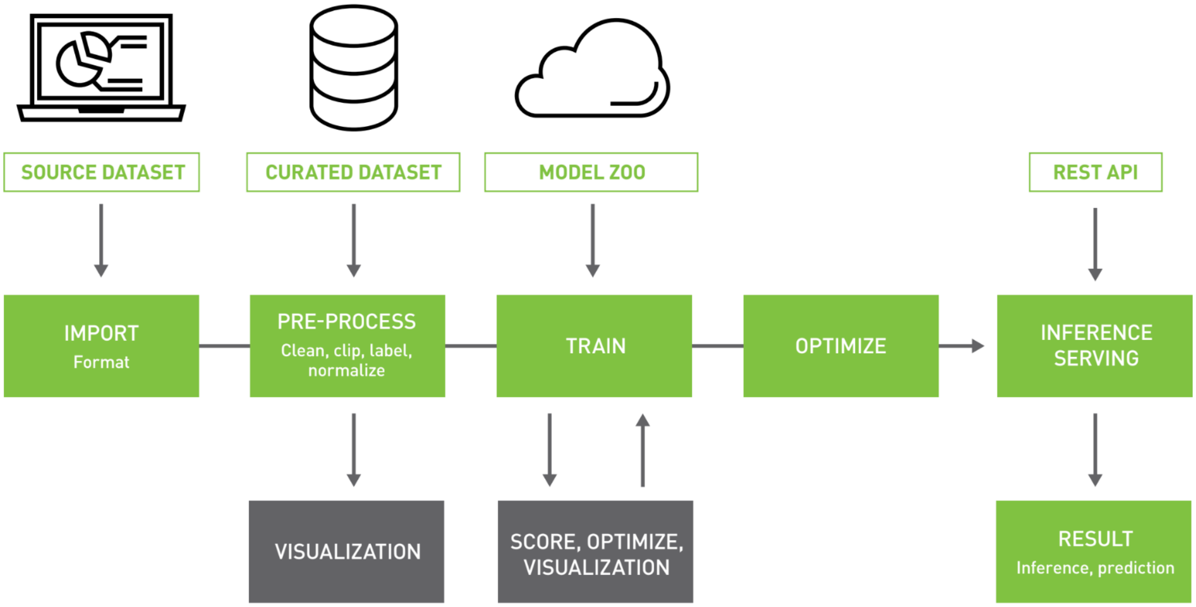
Operationalizing AI models within enterprise applications also poses a number of challenges due to the conflict between the nuances of model building and the operational realities of enterprise IT systems. Infrastructure for AI deployments requires the versatility to support diverse AI model architectures, multiple AI frameworks, handling different types of inference query types, like batch, streaming and ensemble, and supporting multiple environments from edge to cloud.

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 platform for inference.

# End-to-End AI 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. See Figure 1 for end-to-end deep learning workflow, from training to inference.

Figure 1. End-to-End Workflow, from Training to Inference



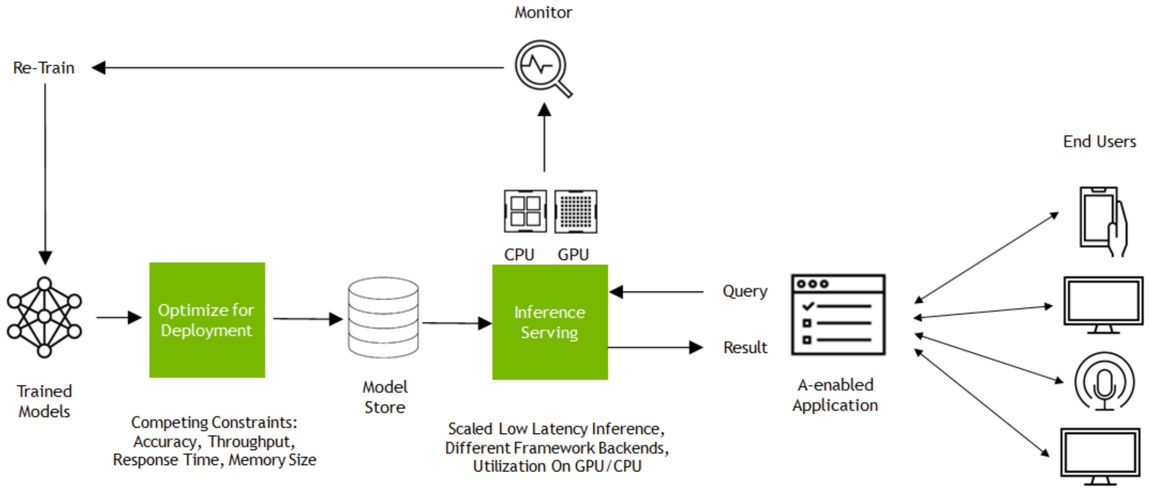
In many organizations, multiple teams are usually involved in AI development and deployment to production: data scientists, machine learning (ML) engineers, application developers, and IT operations. And while they work for the same organization, each has their own specific goals. Supporting the end-to-end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

In this paper, we 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 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 2. Generic AI Inference Workflow


Monitor
Re-Train
End Users
CPU
GPU
Optimize for Deployment
Inference Serving
Query
Result
Trained Models
Model Store
A-enabled Application
Competing Constraints: Accuracy, Throughput, Response Time, Memory Size
Scaled Low Latency Inference, Different Framework Backends, Utilization On GPU/CPU

## 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. 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 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.

## 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:

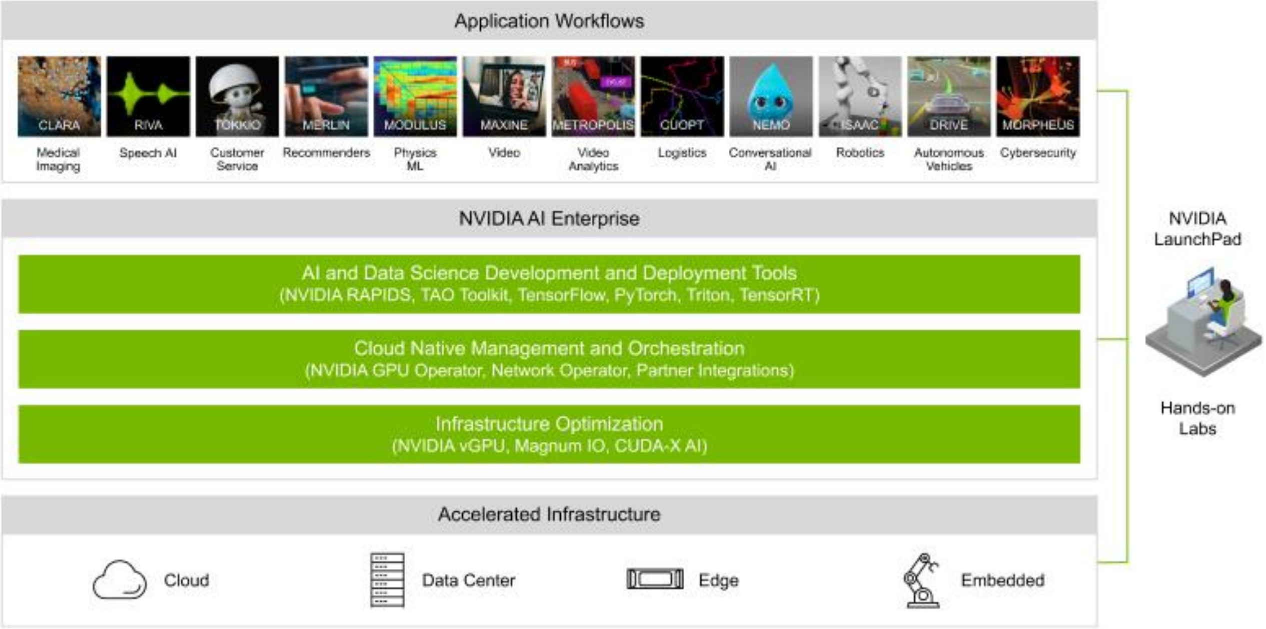
- LATENCY: Latency refers to how much time elapses from an input being presented to the AI model to an output being available. In some applications, low latency is a critical safety requirement. In other applications, latency is directly visible to users as a quality-of-service issue. For larger bulk 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 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.
- 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 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.

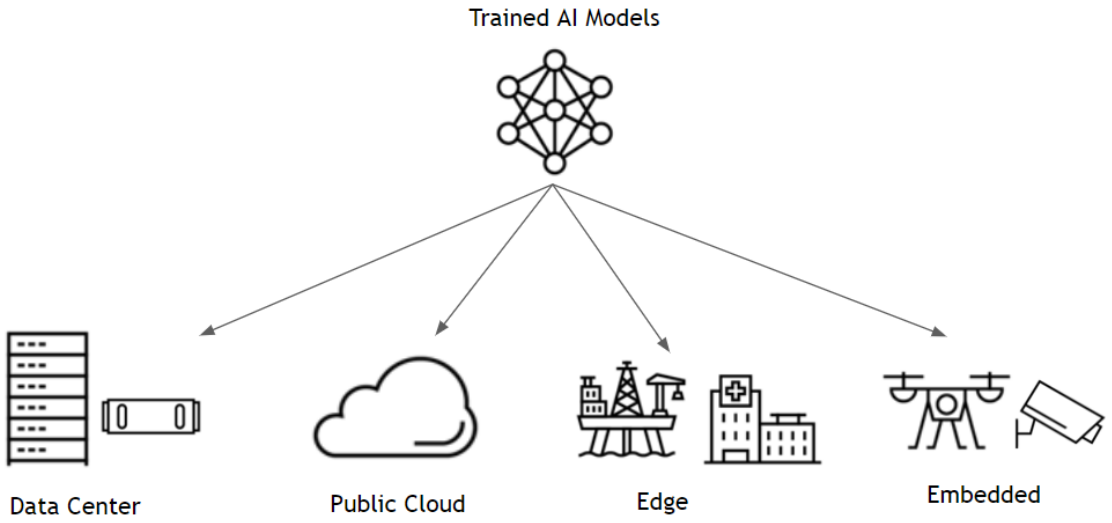
NVIDIA's inference platform delivers the performance, efficiency, and responsiveness critical to powering the next generation of AI products and services. The platform is a combination of architectural innovation, purpose-built to accelerate AI inference workloads, and an end-to-end software stack that 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 3. NVIDIA's Full Stack Approach to AI



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.

Figure 4. 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 GPUs

Following are the Inference GPUs:

- NVIDIA H100

  The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. With the NVIDIA fourth generation NVLINK, H100 accelerate 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. H100 uses breakthrough innovations in the NVIDIA Hopper architecture to deliver industry-leading conversational AI, speeding up large language models by 30X over the previous generation. 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 in inference performance.

- 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 L40 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 provides 120X better performance than CPU based server solutions. Refer to  www.nvidia.com/l40 for more information.
- 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 A10 Tensor Core GPU
  Built on the latest NVIDIA Ampere architecture, the NVIDIA A10 Tensor Core GPU combines second generation RT Cores, third-generation Tensor Cores, and new streaming microprocessors with 24 gigabytes (GB) of GDDR6 memory—all in a 150 W power envelope—for versatile graphics, rendering, AI, and compute performance. NVIDIA A10 is supported as part of NVIDIA-Certified Systems™, in the onprem data center, in the cloud, and at the edge. NVIDIA A10 builds on the rich ecosystem of AI frameworks from the NVIDIA NGC™ catalog, CUDA-X™ libraries, over 2.3 million developers, and over 1,800 GPU-optimized applications to help enterprises solve the most critical challenges in their business.

- 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 provides 120X better performance than CPU based server solutions. Refer to www.nvidia.com/l4 for more information.
- NVIDIA A2 Tensor Core GPU
  NVIDIA A2 Tensor Core GPU provides entry-level inference with low power, a small footprint, and high performance for NVIDIA AI at the edge. The NVIDIA A2 Tensor Core GPU is optimized for inference workloads and deployments in servers constrained by space and thermal requirements, such as 5G edge and industrial environments. Featuring a low-profile PCIe Gen4 card and a low 40-60 watt (W) configurable thermal design power (TDP) capability, the A2 brings adaptable inference acceleration to any organization.

# 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 A100, A30, A2 Tensor Core GPUs & H100 (in the future) 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

## 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. NVIDIA Triton is also integrated with the 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 A100, A30, and A2 GPUs, embedded platforms with NVIDIA® Jetson™, NVIDIA Triton, TensorRT, and Fleet Command , cloud service that securely deploys, manages, and scales AI applications across distributed edge infrastructure.

# NVIDIA AI

## NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud native suite of AI and data analytics software.  It's certified to deploy anywhere—from the enterprise data center to the public cloud—and includes global enterprise support to keep AI projects on track. It includes key enabling technologies and software from NVIDIA to accelerate data preparation, training at scale, optimized and scalable inference.

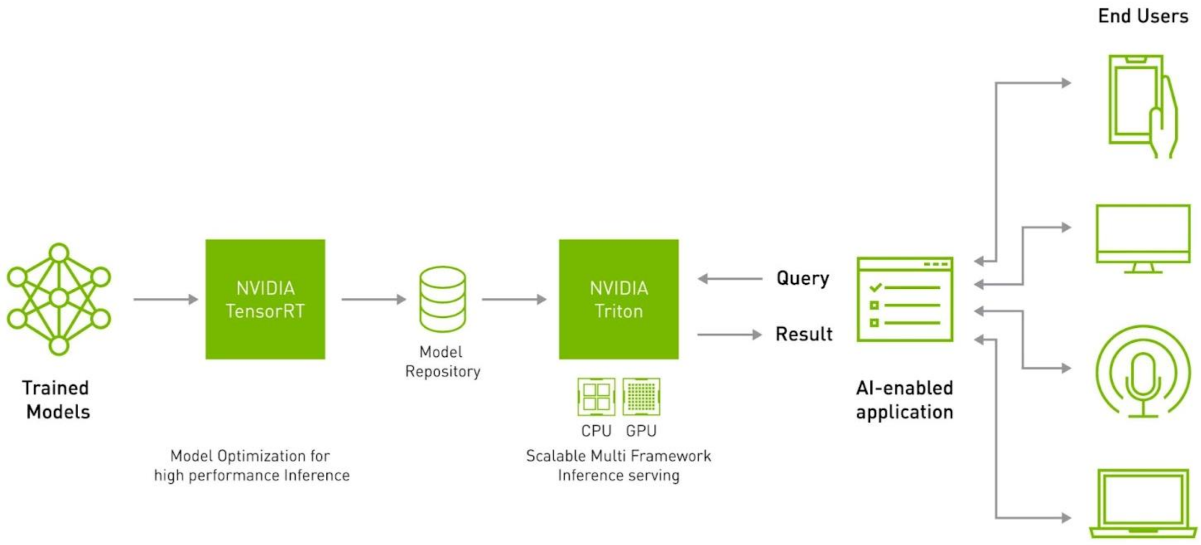
Optimized for AI development and deployment, NVIDIA AI Enterprise includes proven, open-sourced containers and frameworks that ease adoption of enterprise AI. Certified to run on mainstream NVIDIA-Certified servers accelerated by GPUs, or CPU-only, on NVIDIA DGX Systems and  in the public cloud, software in NVIDIA AI Enterprise can be deployed nearly anywhere and enables AI projects to be portable across today's increasingly hybrid data center.   NVIDIA AI Enterprise is also certified to run on common virtualization and container orchestration platforms such as VMware vSphere with Tanzu and Red Hat OpenShift so enterprise IT can integrate AI into the data center while still relying on familiar tools and management solutions. With NVIDIA enterprise support included, both the AI practitioner and IT administrative teams have access to NVIDIA experts globally, for coordinated support across the full solution including partner products, as well as control of upgrade and maintenance schedules with long term support (LTS) options, and access to instructor led customer trainings and knowledge base resources.

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, which will be discussed next.

# Inference Workflow with TensorRT and Triton

Figure 5 shows AI inference workflow with NVIDIA TensorRT and Triton inference server.

Figure 5. AI Inference Workflow with NVIDIA TensorRT and Triton Inference Server



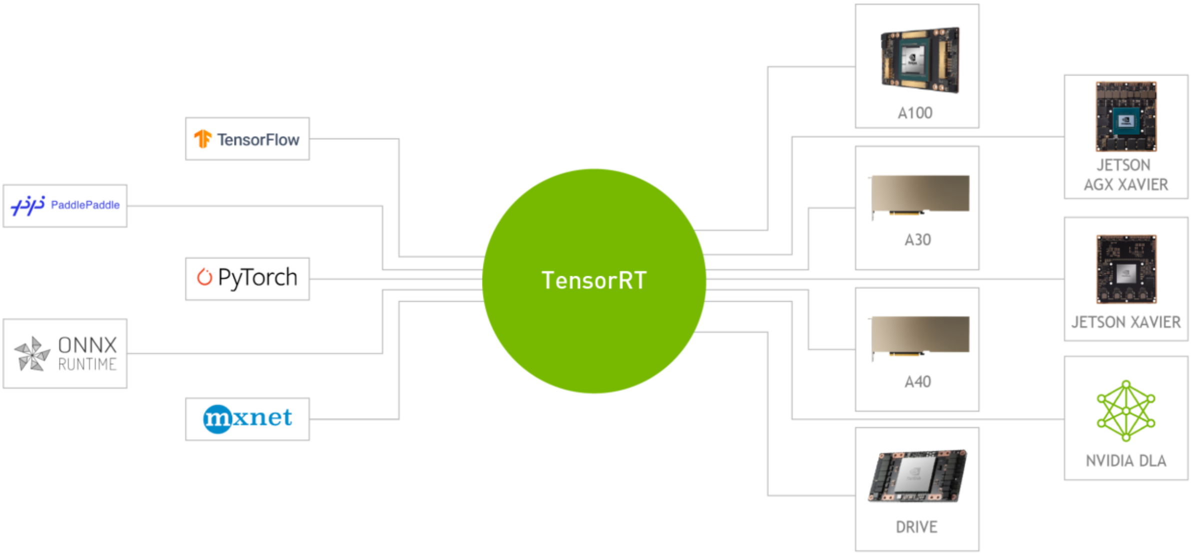
# Inference Optimization with TensorRT

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.

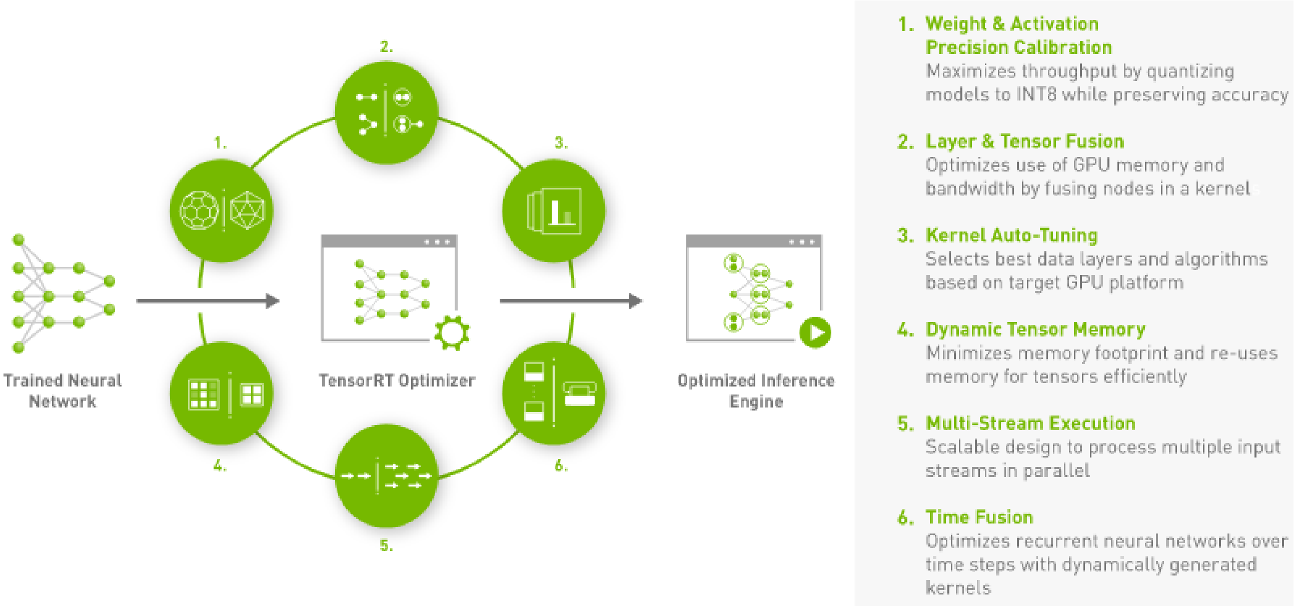
NVIDIA TensorRT is an SDK for high-performance deep learning inference that includes an inference optimizer and runtime. It enables developers to import trained models from all major deep learning frameworks and optimize them for deployment with the highest throughput and lowest latency, while preserving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

Figure 6. 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 or INT8), for improved latency, throughput, and efficiency.

Figure 7. 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

The 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.

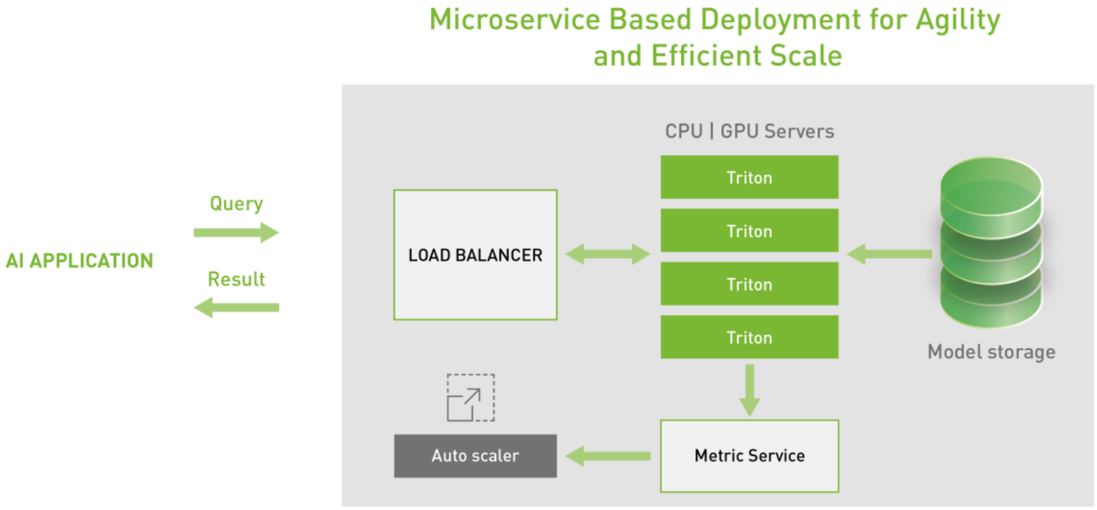
To keep up with the latest TensorRT features and developer resources, check out the TensorRT Getting Started zone.

# 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/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 opensource 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.

Figure 8. High Level Triton Deployment 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, 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.

## 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 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. The NVIDIA TensorRT SDK and NVIDIA Triton are both available as part of the NVIDIA AI Enterprise Suite, with enterprise support from NVIDIA.

# 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 9 provides a quick overview of a few of these.

Figure 9. Application-specific Frameworks to Accelerate Developer Productivity

| | |
|---|---|
| **NVIDIA Clara \| Healthcare**<br>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.<br>Learn More | **NVIDIA Isaac \| Robotics**<br>A toolkit that includes building blocks and tools to accelerate robot developments that require the increased perception and navigation features enabled by AI.<br>Learn More |
| **NVIDIA Driveworks \| Automotive**<br>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.<br>Learn More | **NVIDIA Aerial \| Telco**<br>An application framework for building high performance, software defined, cloud native 5G applications to address increasing consumer demand.<br>Learn More |
| **NVIDIA Morpheus \| Cybersecurity**<br>An application framework that enables cybersecurity developers to create optimized AI pipelines for filtering, processing, and classifying large volumes of real-time data.<br>Learn  More | **NVIDIA Maxine \| Video Conferencing**<br>Suite of GPU accelerated SDKs for AI enabled audio, video and augmented reality in communications applications.<br>Learn More |
| **NVIDIA Riva \| Speech AI**<br>GPU-accelerated speech AI SDK for building applications like live captioning, human-like voice interfaces for virtual assistants, and branded voices that can deliver world-class accuracy and run in real-time.<br>Learn More | **NVIDIA Merlin \| Recommender Systems**<br>An opensource framework for building large scale recommender systems, from ingesting and training to deploying a production-quality pipeline.<br>Learn More |
| **NVIDIA Metropolis \| Intelligent Video Analytics**<br>An application framework that simplifies the development, deployment, and scaling of AI-enabled video analytics applications from edge to cloud.<br>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

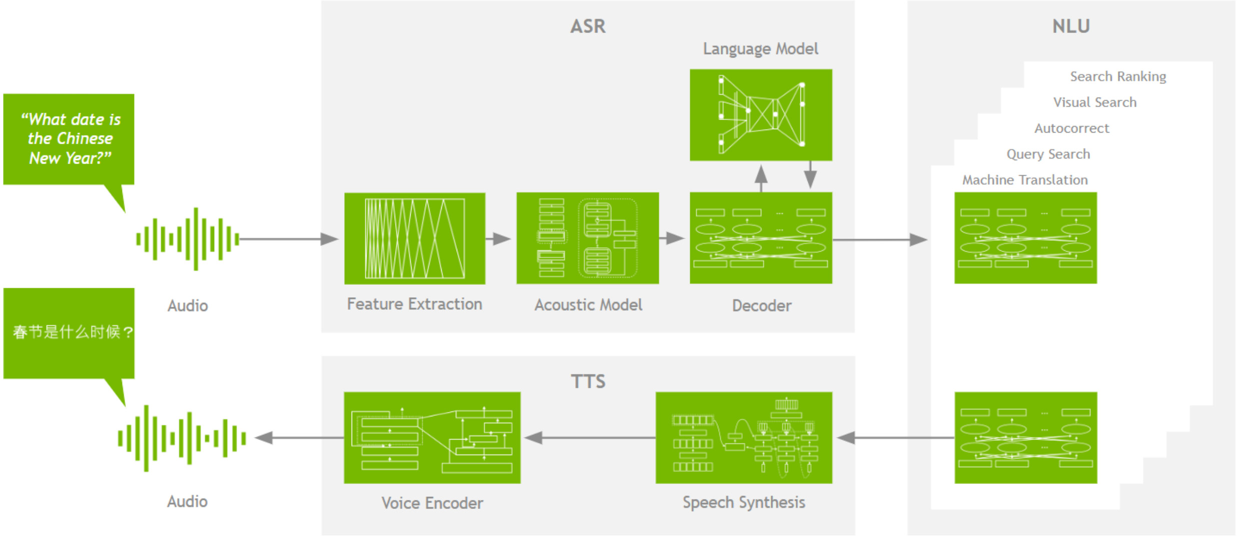
Conversational AI is the application of machine learning to develop language-based applications that allow humans to interact naturally with devices, machines, and computers using speech. In the last few years, deep learning has improved the state-of-the-art in conversational AI and offered superhuman 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.

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 10).

Figure 10. Overview of a Conversational AI Pipeline

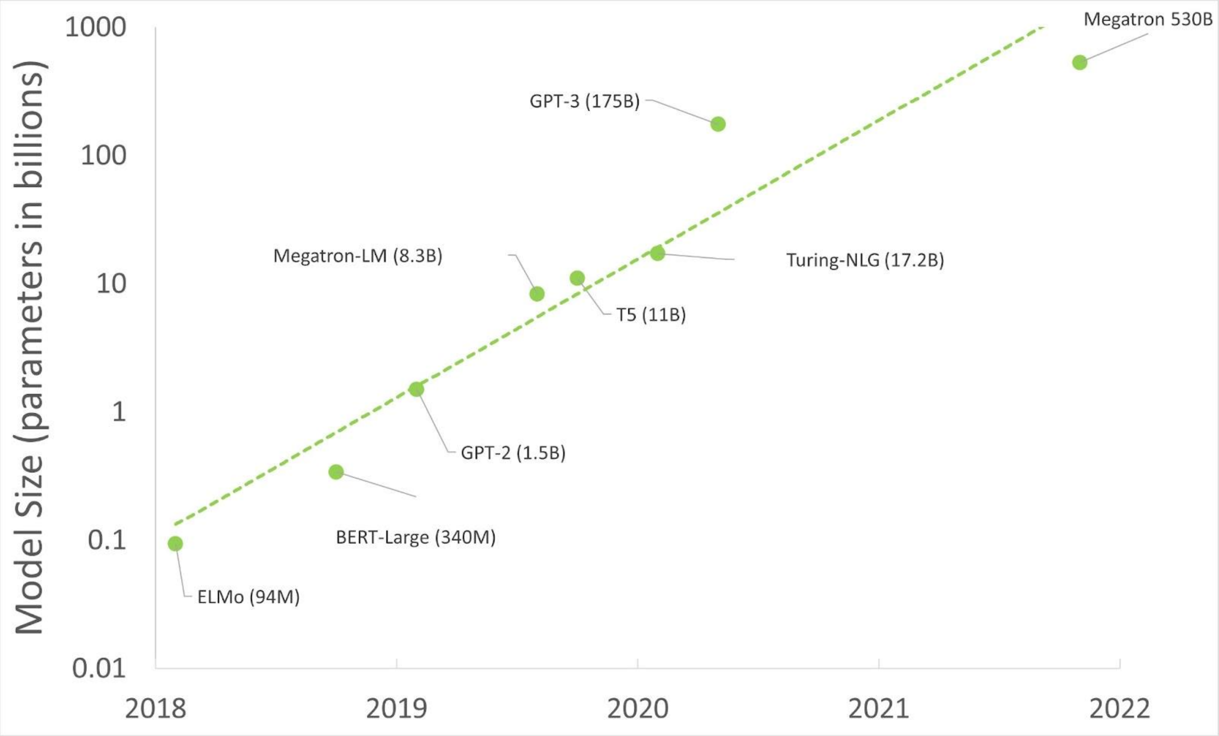


A Conversational AI pipeline has three sections: Automatic speech recognition, Natural Language understanding, and Text-to-Speech. Within each of these sections there can be multiple natural networks off working together to quickly deliver an answer to a posed question.

## 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. Transformer based 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 (see Figure 11).

Figure 11. State-of-the-art NLP Model Sizes Increasing with Time



Researchers at NVIDIA also released Megatron, a PyTorch-based framework for training giant language models based on the transformer architecture. Using Megatron, researchers can 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. 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 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.

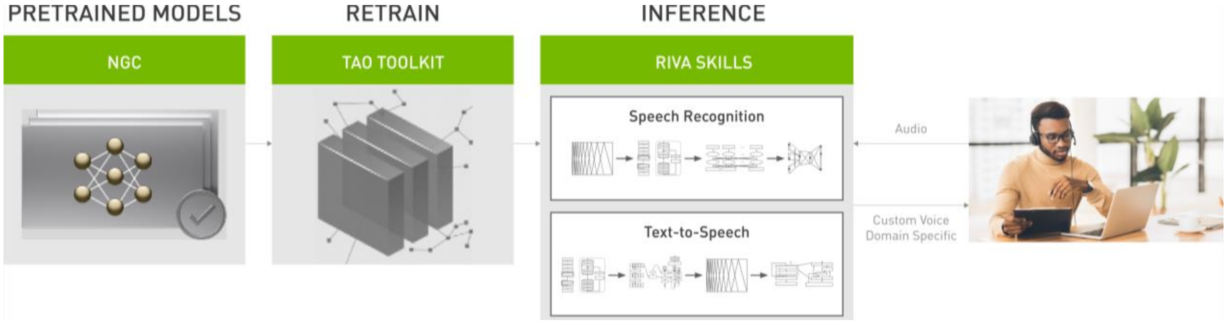
Hence, the platform needed to deliver a conversational AI service must be both performant and programmable so that AI developers can accelerate time to solution, build new services, and continuously push the boundaries of conversational AI.

## NVIDIA Riva – Build and Deploy Speech AI Applications

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

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

Figure 12. Train and Deploy an End-to-End Speech AI Pipeline using NVIDIA Riva



NVIDIA Riva provides state-of-the-art models, fully accelerated pipelines, and tools to easily add Speech AI capabilities to real-time applications like virtual assistants, call center agent assist, and video conferencing. Riva components are customizable, so you can adapt the applications for your use case and industry and deploy them in any cloud, on-premises, and at the edge.

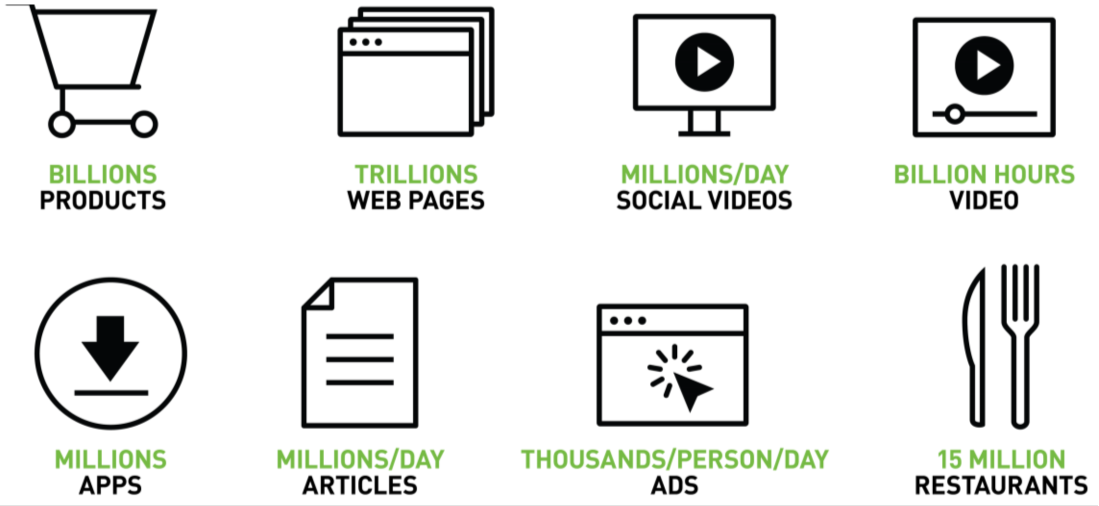
Under the hood, Riva applies powerful NVIDIA TensorRT optimizations to models, configures the NVIDIA Triton Inference Server for model serving, and exposes the models as a service through a standard API that can be easily integrated into applications. For domain-specific data, users can fine-tune Riva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

NVIDIA Riva is available freely to all members of the NVIDIA Developer Program on NGC. Riva can be deployed free of charge up to a certain usage limit. Organizations looking to deploy Riva-based applications beyond the free limit with enterprise-grade support, can leverage the Riva Enterprise paid subscription program.

# Recommender Systems

It is simply impossible for enterprises to connect billions of users in the world with the products, services, even expertise—among trillions of things[1]—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.

Figure 13. 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 from 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

[1] https://time.com/5955412/artificial-intelligence-nvidia-jensen-huang/

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.

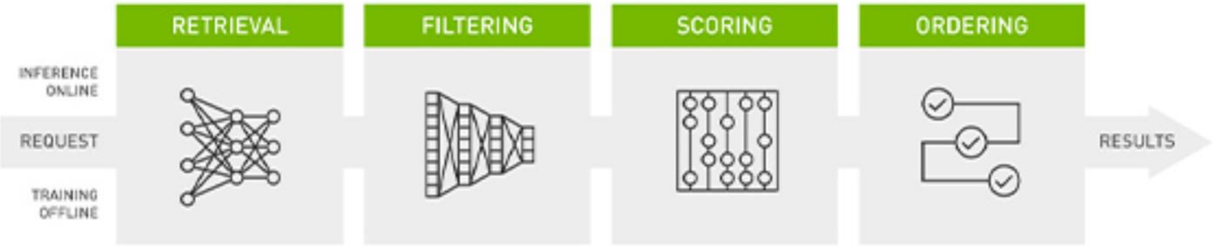
With NVIDIA GPUs, you can exploit data parallelism through columnar data processing instead of traditional row-based reading designed initially for CPUs. This provides higher performance and cost savings. Current DL–based models for recommender systems like DLRM, Wide and Deep (W&D), Neural Collaborative 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 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 14). With Merlin, better predictions, increased click-through rates, and faster deployment to production are within reach.

Figure 14. Recommender workflow with NVIDIA Merlin



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 Models
  Merlin Models is a library that provides standard models for recommender systems and high-quality implementations from ML to more advanced DL models on CPUs and GPUs. Train models for retrieval and ranking within 10 lines of code.
- 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 transfers.
- Merlin Distributed Training
  Merlin provides support for distributed training across multiple GPUs. Components include Merlin SOK (SparseOperationsKit) 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.

# 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. Thousands of companies wordwide, like the ones show in Figure 15, are using the NVIDIA AI inference platform to transform their businesses

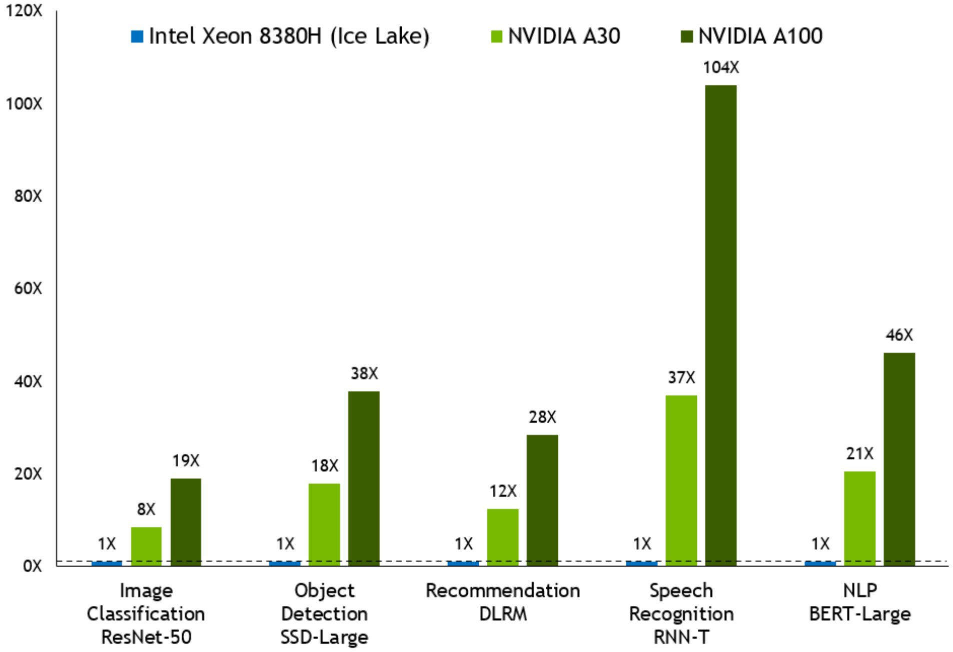
Figure 15. NVIDIA AI Inference Platform

| | | | |
|---|---|---|---|
| Actively detect diseases in 145 million hearts per year | 5X faster spell check for enhanced product search<br>amazon | Award-winning customer service<br>T Mobile | |
| Office grammar checker reduced cost by 70%<br>Microsoft | Real-time analytics on 7 billion packages per year<br>UNITED STATES POSTAL SERVICE | Intelligent search with SOTA NLU for 1.2 billion users<br>We Chat | Enhanced Realtime fraud detection<br>AMERICAN EXPRESS |

## MLPerf Inference

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 shows the elative performance of GPUs and CPUs on a per-chip basis, normalized to CPU. Table 1 shows the data used to generate Figure 16.

Figure 16. MLPerf Inference 2.0—Elative Performance of GPUs and CPUs on a per-chip basis, Normalized to CPU



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

Table 1. Raw Data Showing per-chip Inference Performance Across all Workloads

| Cell Heading | Intel Xeon 8380 (Ice Lake) | NVIDIA A30 | NVIDIA A100 |
|---|---|---|---|
| Image Classification ResNet-50 | 1X | 8X | 19X |
| Object Detection SSD-Large | 1X | 18X | 38X |
| Recommendation DLRM | 1X | 12X | 28X |
| Speech Recognition RNN-T | 1X | 37X | 104X |
| NLP BERT-Large | 1X | 21X | 46X |

Note: MLPerf v2.0 Inference Closed; Per-accelerator performance derived from the best MLPerf results for respective submissions using reported accelerator count in Data Center Offline and Server. Intel Xeon 8380 from MLPerf v.1.1 submission: 1.1-023 and 1.1-024, Intel Xeon 8380H 1.1-026, NVIDIA A30: 2.0-090, NVIDIA A100 (X86): 2.0-094.
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 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 about 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.

## NVIDIA Launchpad

NVIDIA LaunchPad is a free program that provides users with short-term access to a large catalog of hands-on labs. Now enterprises and organizations can immediately tap into the necessary hardware and software stacks to experience end-to-end solution workflows in the areas of AI, data science, 3D design collaboration and simulation, and more.

NVIDIA LaunchPad resources deployed across partner data centers are available directly from NVIDIA in nine regions globally. These resources include both NVIDIA DGX™ supercomputers and mainstream NVIDIA-Certified Servers™ running complete NVIDIA software stacks.

LaunchPad helps developers, designers, and IT professionals speed up the creation and deployment of modern, data-intensive applications, including AI inference. After quick testing and prototyping, the same complete stack can be deployed for production workflows, so more confident software and infrastructure decisions can be made.

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.

Explore the free, hands-on AI labs with Triton 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 every deployment option Facross on-prem, hybrid and multi-cloud environments
- Access to NVIDIA AI Experts: 8-5 local business hours for guidance on configuration and performance, including access to engineering
- Priority notifications of the latest security fixes and maintenance releases
- Long term support for up to 3-years for designated software branches
- Customized support upgrade option: designated Technical Account Manager (TAM) and Business Critical support for 24x7 live agent access

Learn how you can benefit from the NVIDIA AI Inference Platform and take your AI projects from prototype to production.

## 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 NGC, NVIDIA-Certified Systems, NVIDIA Clara, NVIDIA TensorRT, NVIDIA Jetson, NVIDIA Triton, NVIDIA RAPIDS, and NVIDIA Merlin are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and/or other countries. Othercomp any 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