



NVIDIA

GTC S41755 FAST, SCALABLE, STANDARDIZED AI INFERENCE DEPLOYMENT

March 2022

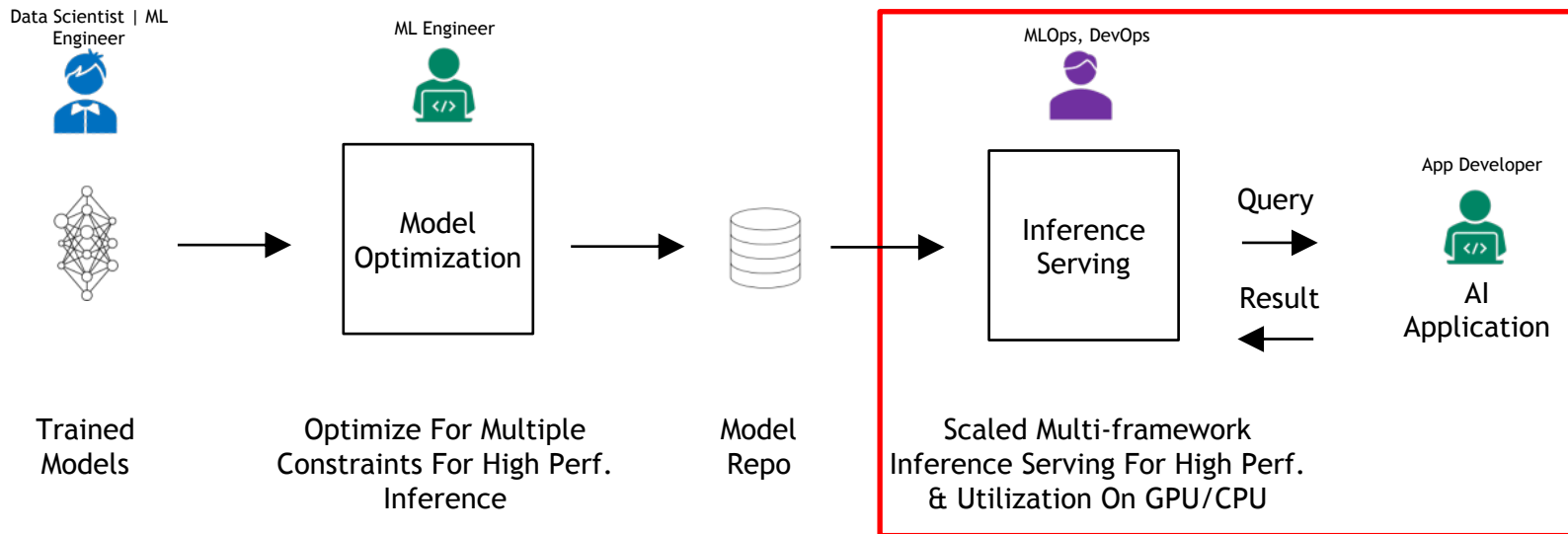
Shankar Chandrasekaran, Triton Product Marketing Manager

Mahan Salehi, Triton Product Manager

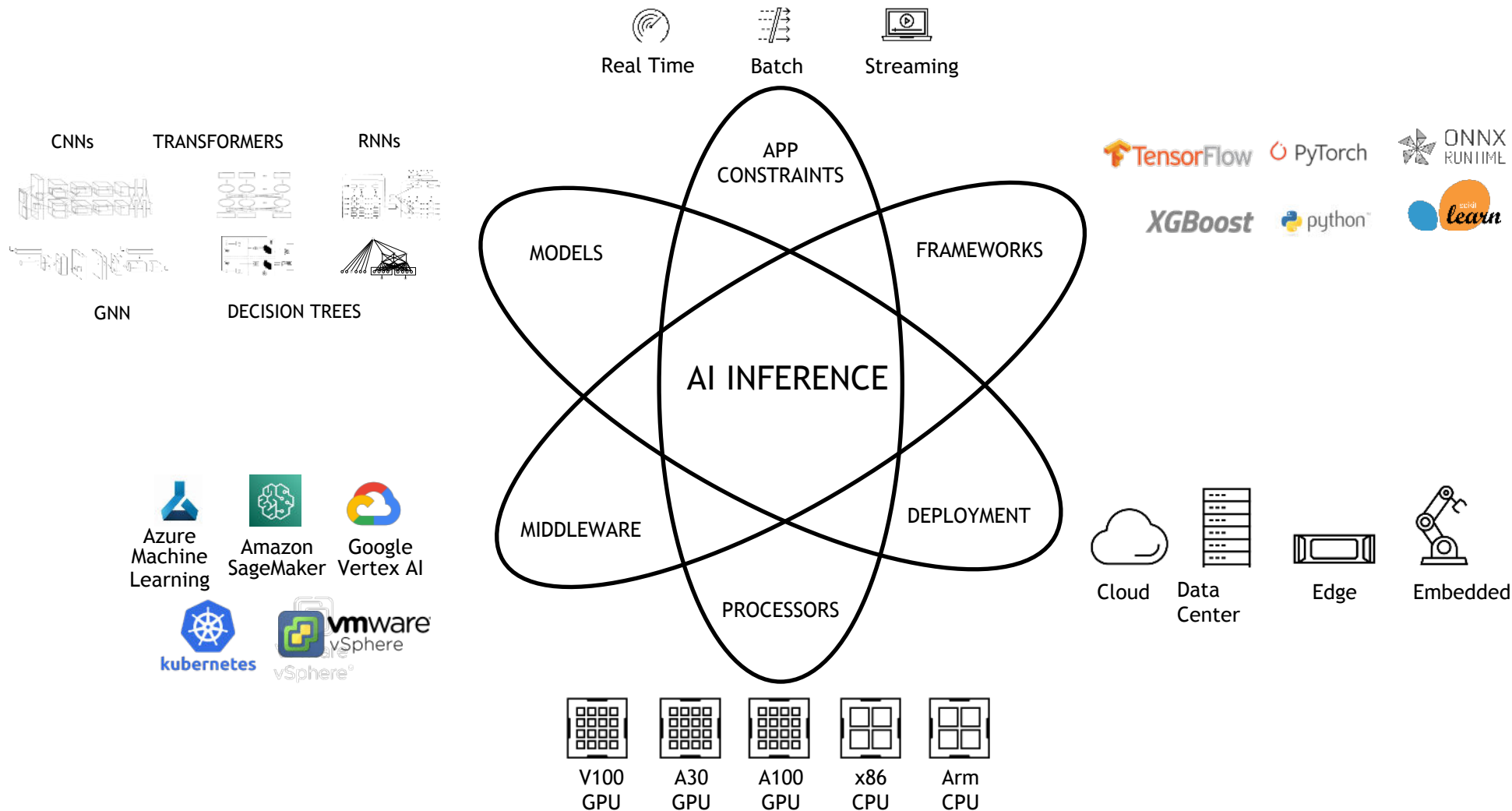


AI INFERENCE WORKFLOW

Two Part Process Implemented by Multiple Personas



AI INFERENCE IS HARD



TRITON INFERENCE SERVER

Open-Source Software For Fast, Scalable, Simplified Inference Serving

Any Framework



Supports Multiple Framework Backends Natively e.g., TensorFlow, PyTorch, TensorRT, XGBoost, ONNX, Python & More

Any Query Type



Optimized for Real Time, Batch, Streaming, Ensemble Inferencing

Any Platform



X86 CPU | Arm CPU | NVIDIA GPUs | MIG
Linux | Windows | Virtualization
Public Cloud, Data Center and Edge/Embedded (Jetson)

DevOps & MLOps



Integration With Kubernetes, KServe, Prometheus & Grafana
Available Across All Major Cloud AI Platforms

Performance & Utilization



Model Analyzer for Optimal Configuration
Optimized for High GPU/CPU Utilization, High Throughput & Low Latency

<https://developer.nvidia.com/nvidia-triton-inference-server>



NATIVELY SUPPORTED EXECUTION BACKENDS

TensorFlow 1.x/2.x

Any Model
SavedModel | GraphDef

PyTorch

Any model
JIT/Torchscript | Python

TensorRT

All TensorRT optimized
models

TF-TensorRT & TorchTRT

Any TensorFlow and PyTorch model

FIL

Tree based models (e.g. XgBoost,
Scikit-learn RandomForest, LightGBM)

ONNX RT

ONNX converted models

Python

Custom code in Python e.g.
pre/post processing, any Python
model.

Faster Transformer Backend (Alpha)

Multi-GPU, multi-node inferencing for
large transformer models (GPT and T5)

OpenVINO

OpenVINO optimized models on Intel
architecture

Custom C++ Backend

Custom framework in C++

DALI

Pre-processing logic using DALI
operators

NVTabular

Feature engineering and preprocessing
library for tabular data.

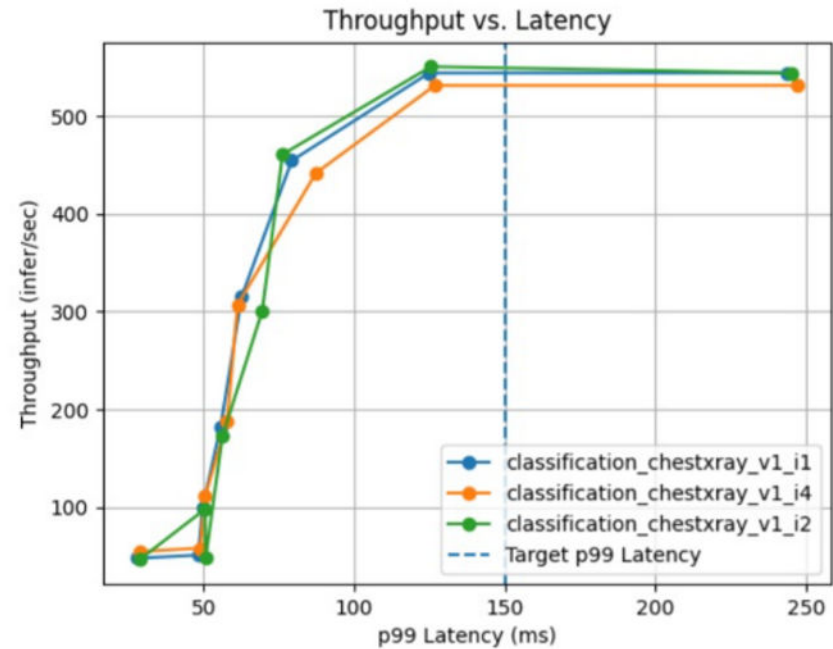
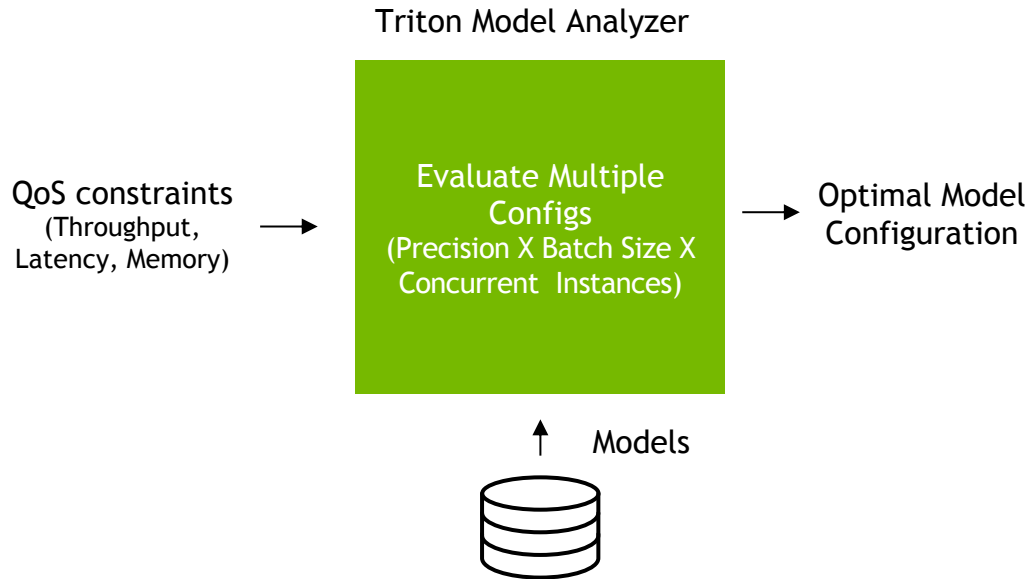
HugeCTR

Recommender model with large
embeddings



MODEL ANALYZER

Optimal Triton Configuration

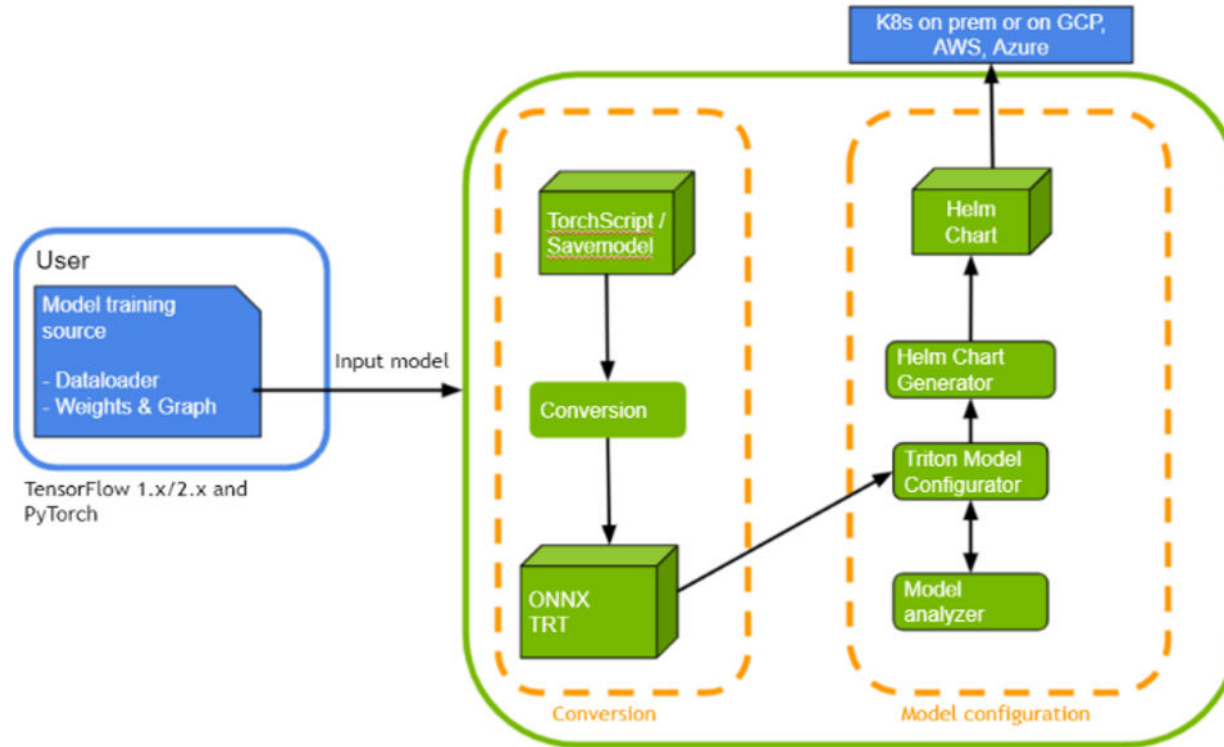


Throughput vs. Latency curves for 3 configurations of classification_chestxray_v1



MODEL NAVIGATOR

Accelerate optimized model deployment



https://github.com/triton-inference-server/model_navigator



TRITON FOREST INFERENCE LIBRARY (FIL)

Model Explainability With Shapley Values

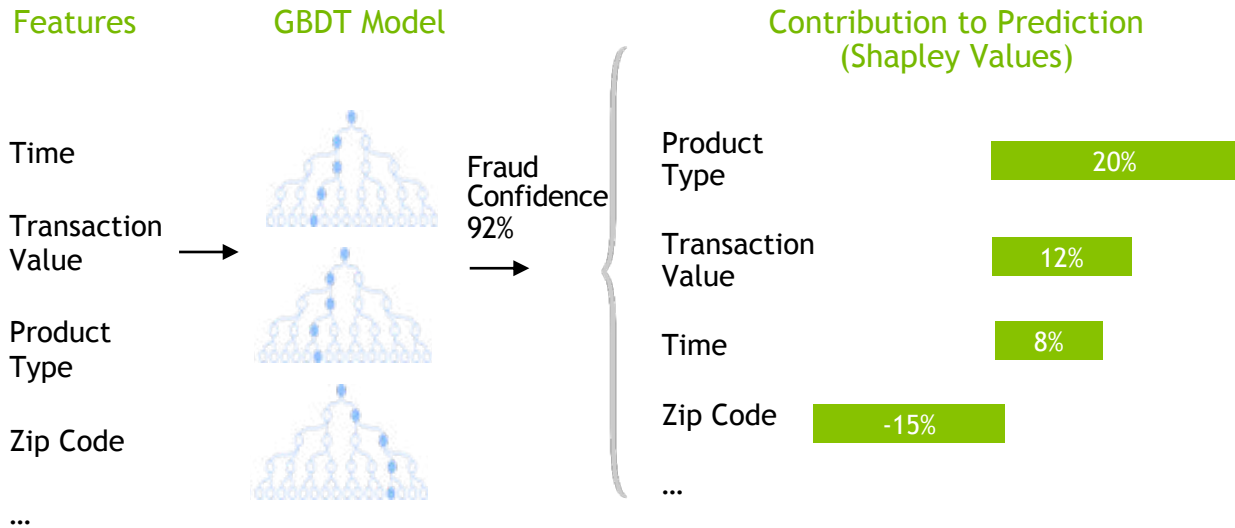
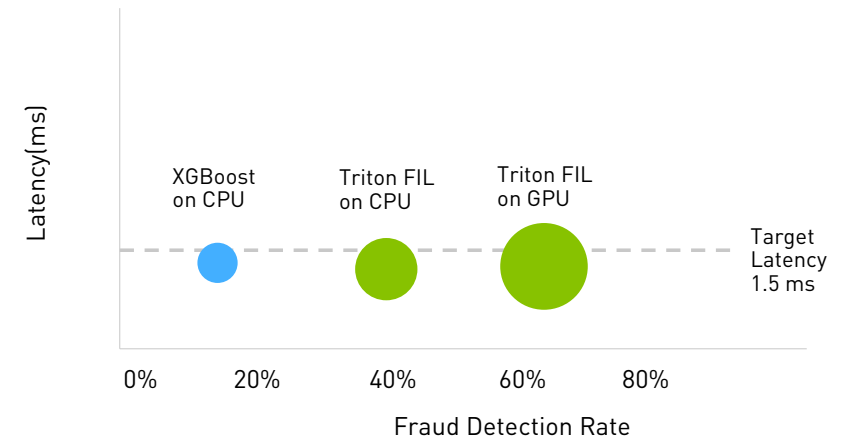


Illustration of Fraud detection with Triton FIL



LATEST TRITON FEATURES

MLFlow Integration Plugin (POC)

Plugin to convert MLFlow Model Repo to Triton model repo format

Implicit State Management for TensorRT Backend

Native support for stateful models (e.g. ASR). Also improves performance for autoregressive models. Support for TF/PyT/ORT backends on roadmap

Business Logic Scripting

Enables users to programmatically run “business logic” (control flow, conditionals, etc) in ensemble pipelines

Inferentia Support

Run models on AWS Inferentia

Java HTTP Client (Alpha)

HTTP Java client available in Triton

Fleet Command Integration

Helm chart to deploy Triton in Fleet Command

Embedded Devices

Triton for Jetpack for Xavier, Nano, and TX2
Now supports PyTorch and Python backends

FIL Backend- CPU Performance Optimizations

Enable multi-threading on CPU and other optimizations, resulting in 10x throughput speedup

Rate Limiter

Enables cross model prioritization. Controls how many requests are fed into each model simultaneously

Container Composition Utility Tool

Tooling to automate process of customizing Triton container and reducing size by adding/removing backends



Triton FasterTransformer Backend

Inference on Giant Multi-GPU, Multi-Node NLP Models with Billions of Parameters

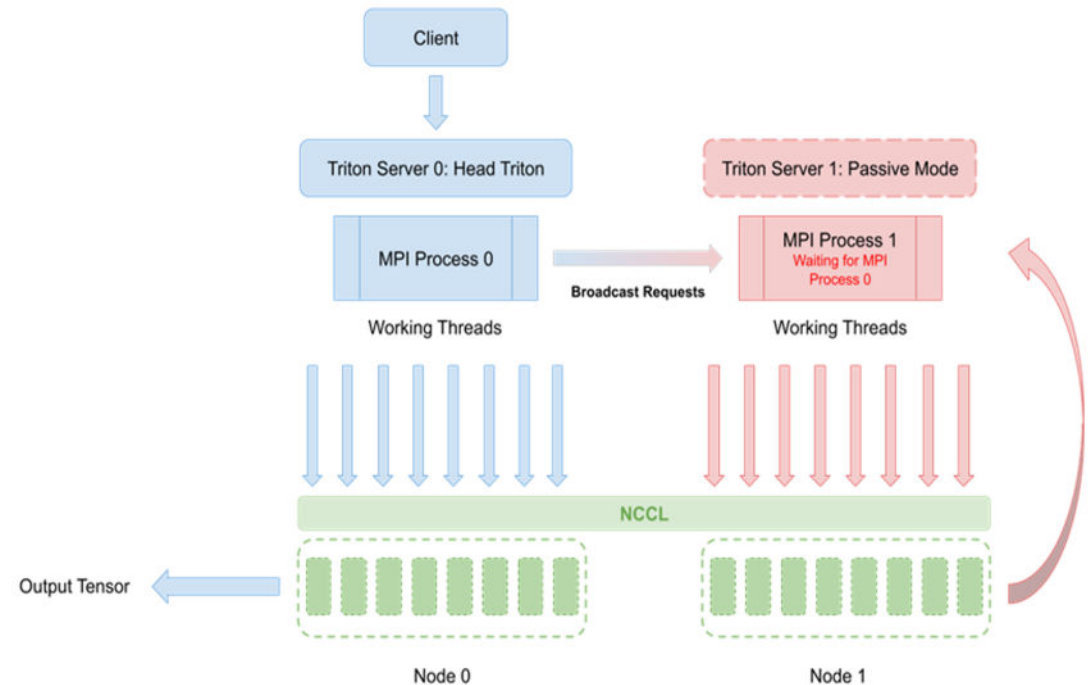
Goal: Serve giant transformer models and accelerate inference performance

Current Capabilities:

- Written in C++/CUDA and relies on cuBLAS, cuBLASLt, cuSPARSELt
- Optimize kernels to accelerate inference for encoder/decoder layers of transformer models
- Integrated as a backend in Triton Inference Server
- Uses tensor/pipeline parallelism for multi-GPU, multi-node inference
- FP16, FP32 supported
- POC of Post-training weight-only INT8 quantization for GPT
 - Only for BS 1-2
- Supports sparsity for BERT
- Uses MPI and NCCL to enable inter/intra node communication

Exceptions/Limitations:

- Supports only GPT and T5 style models currently for multi-node
- Model must be converted to FasterTransformer format
- Megatron and HuggingFace converters provided
 - POC of Tensorflow/ONNX converters
- Currently beta release



Nemo-Megatron EA Program:

<https://developer.nvidia.com/nemo-megatron-early-access>

Triton FasterTransformer Open Source Github Repo:

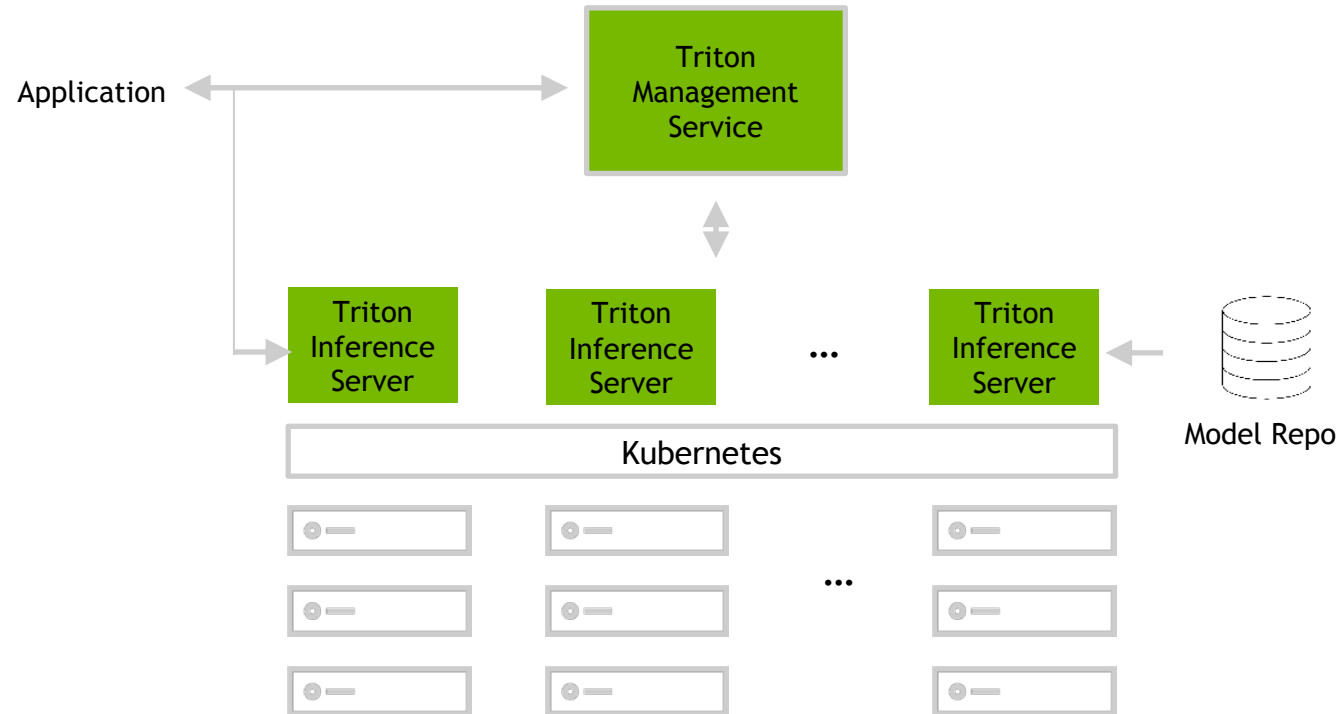
https://github.com/triton-inference-server/fastertransformer_backend



UPCOMING FEATURE – TRITON MANAGEMENT SERVICE

Triton Management Service (TMS)

- Deploys Triton with requested models
- Load models on demand, unloads models when not in use
- Helps group models from different frameworks together to ensure they co-exist efficiently
- Spins up new Triton pods on increasing inference load
- Health check w/ restarts of Triton instances



ECOSYSTEM INTEGRATIONS



Amazon SageMaker



Azure Machine Learning



Google Vertex AI
(*New* Native Integration)



Alibaba Cloud
PAI-EAS



Amazon Elastic
Kubernetes Service (EKS)



Amazon Elastic
Container Service
(ECS)



Azure
Kubernetes
Service (AKS)



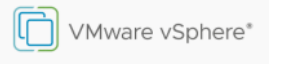
Google Kubernetes
Engine (GKE)



Kubeflow/KServe



Prometheus



TRITON ADOPTION ACROSS USE CASES



Search & Ads



Fraud Detection



Fraud Detection



Fraud Detection



Serving platform



Grammar Check



Meeting Transcription



Document Translation



Image
Classification &
Recommendation



Image
Segmentation



Preventive maintenance



Defect Detection



Video Content
Audit



Product Identification



Package Analytics



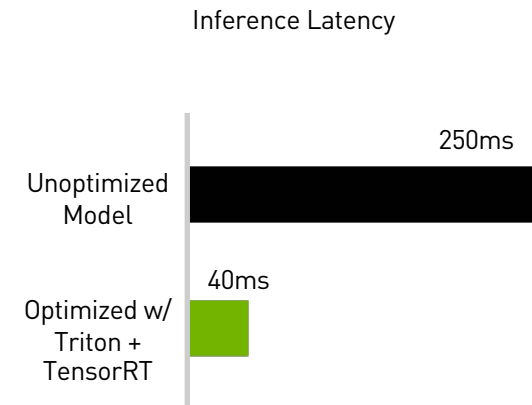
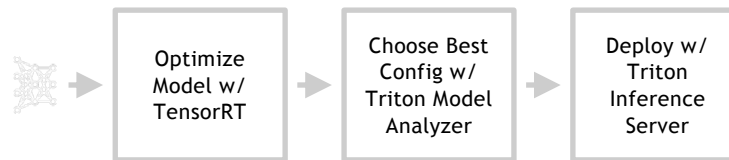
Clinical Notes
Analytics



AMAZON ADOPTS NVIDIA AI FOR REAL TIME SPELL CHECK FOR PRODUCT SEARCH

Real Time Spelling Correction Of Search Text
Triton + TensorRT Meets Latency Target While
Optimizing for Throughput

Triton Model Analyzer Reduced Time to Find
Optimal Configuration from Weeks to Hours



MICROSOFT ADOPTS TRITON FOR DOCUMENT TRANSLATION IN TRANSLATOR SERVICE

30 language pairs w/ 1s latency on GPU

100X throughput vs CPU (154 sentences/s vs 1 sentence/s)

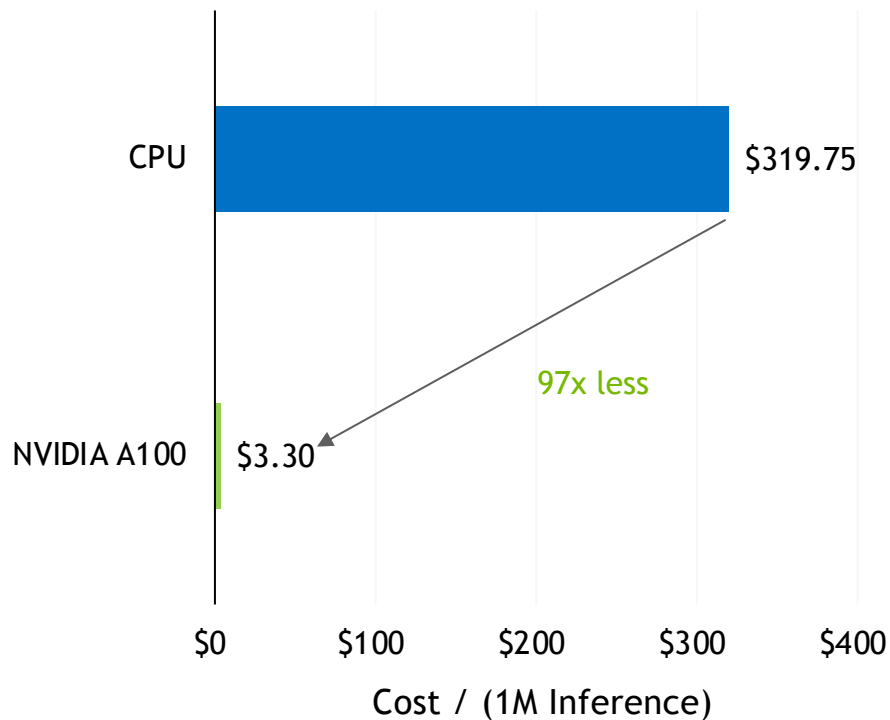
Cost effectively scales to thousands/millions of users (Few GPU servers vs. 100's of CPU servers)

New GPU model replaces 20 separate CPU models

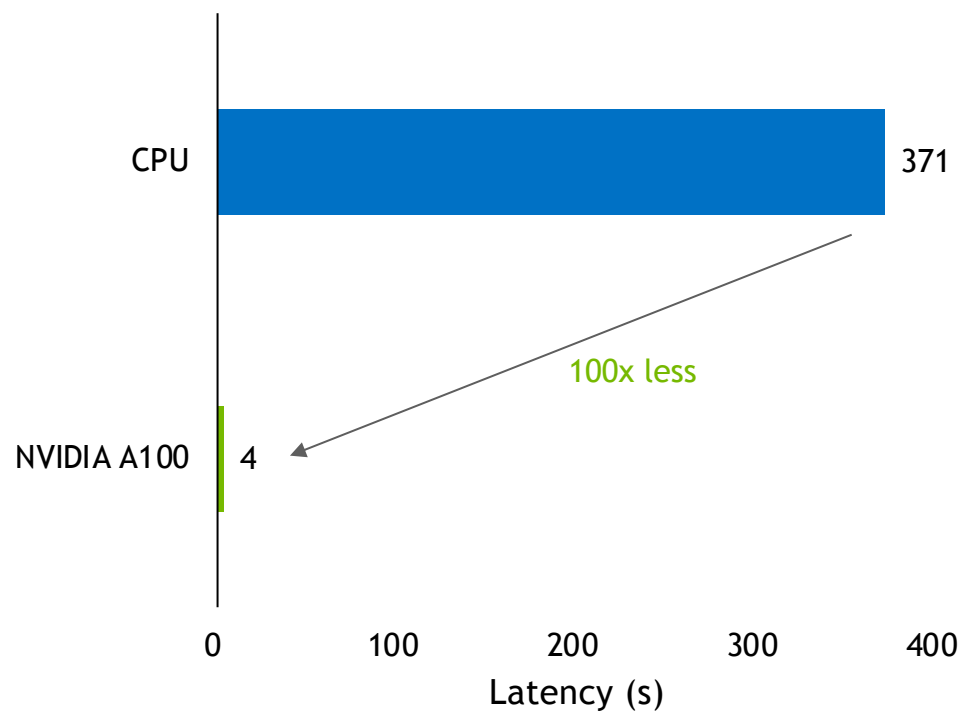


EASILY GET TCO AND LATENCY ADVANTAGES WITH TRITON

BERT-Large on A100 vs. CPU



NVIDIA A100 97x better TCO compared to CPU

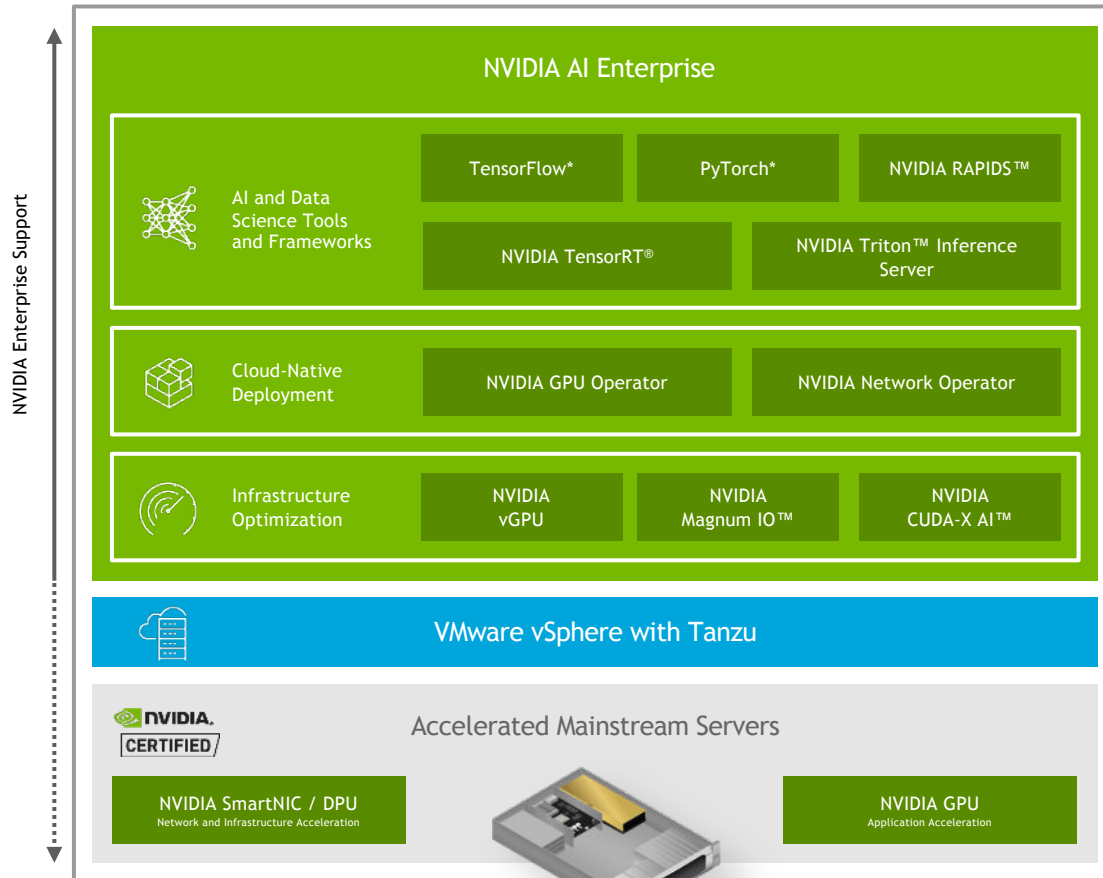


NVIDIA A100 100x lower latency meeting real-time threshold



NVIDIA AI ENTERPRISE SOFTWARE SUITE

Enabling AI and Data Analytics on VMware vSphere



Optimized for Performance

Bare-metal performance across multiple nodes to power large, complex training and machine learning workloads virtualized



Certified for VMware vSphere

Reduce deployment risks with a complete suite of NVIDIA AI software certified for the VMware data center



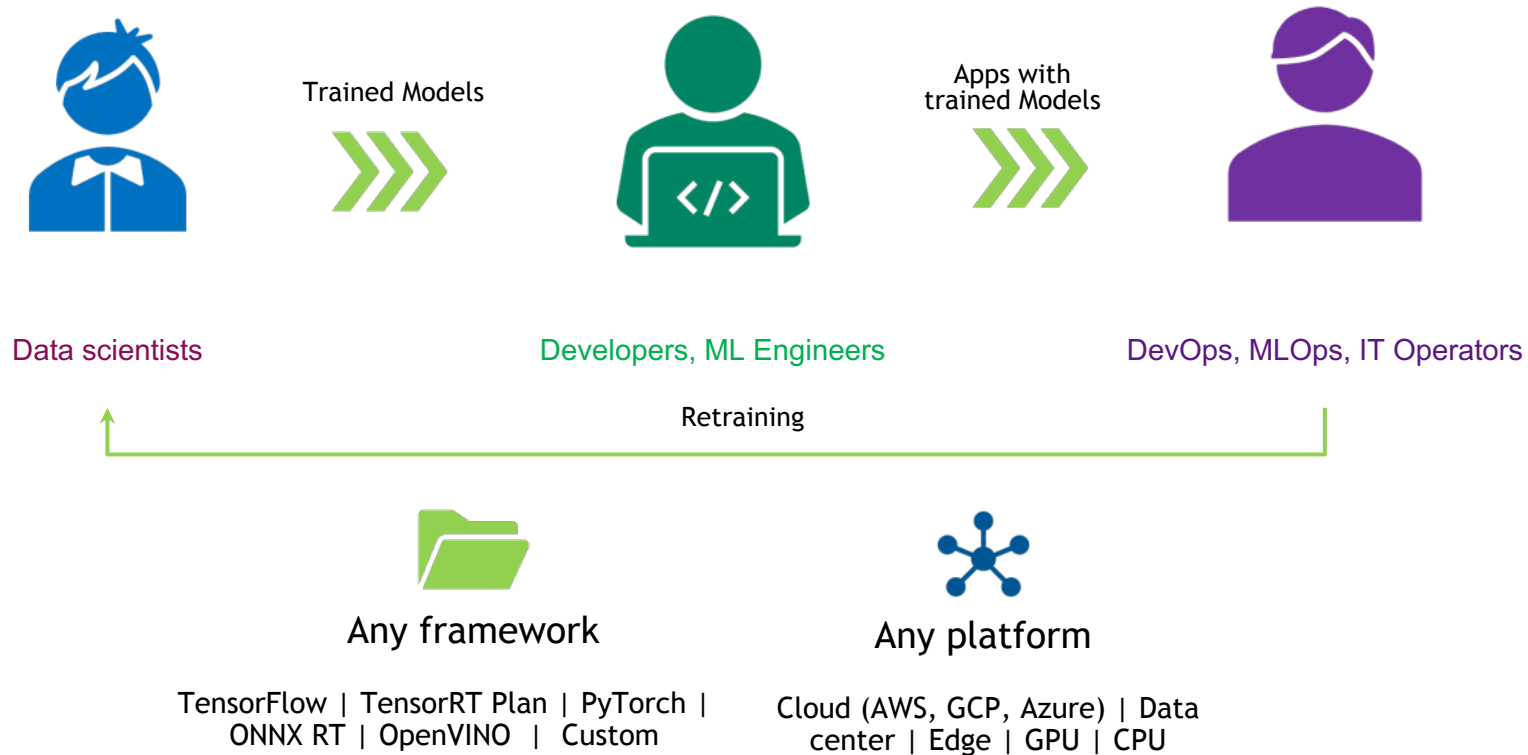
NVIDIA Enterprise Support

Ensure mission-critical AI projects stay on track with access to NVIDIA experts



TRITON: BRINGING THE 3 TEAMS TOGETHER

Inference Serving: Simplified, Highly Performant & Flexible



LEARN MORE AND DOWNLOAD

For more information

<https://developer.nvidia.com/nvidia-triton-inference-server>

Get the ready-to-deploy container with monthly updates from the NGC catalog:

<https://catalog.ngc.nvidia.com/orgs/nvidia/containers/tritonserver>

Open-source GitHub repository:

<https://github.com/NVIDIA/triton-inference-server>

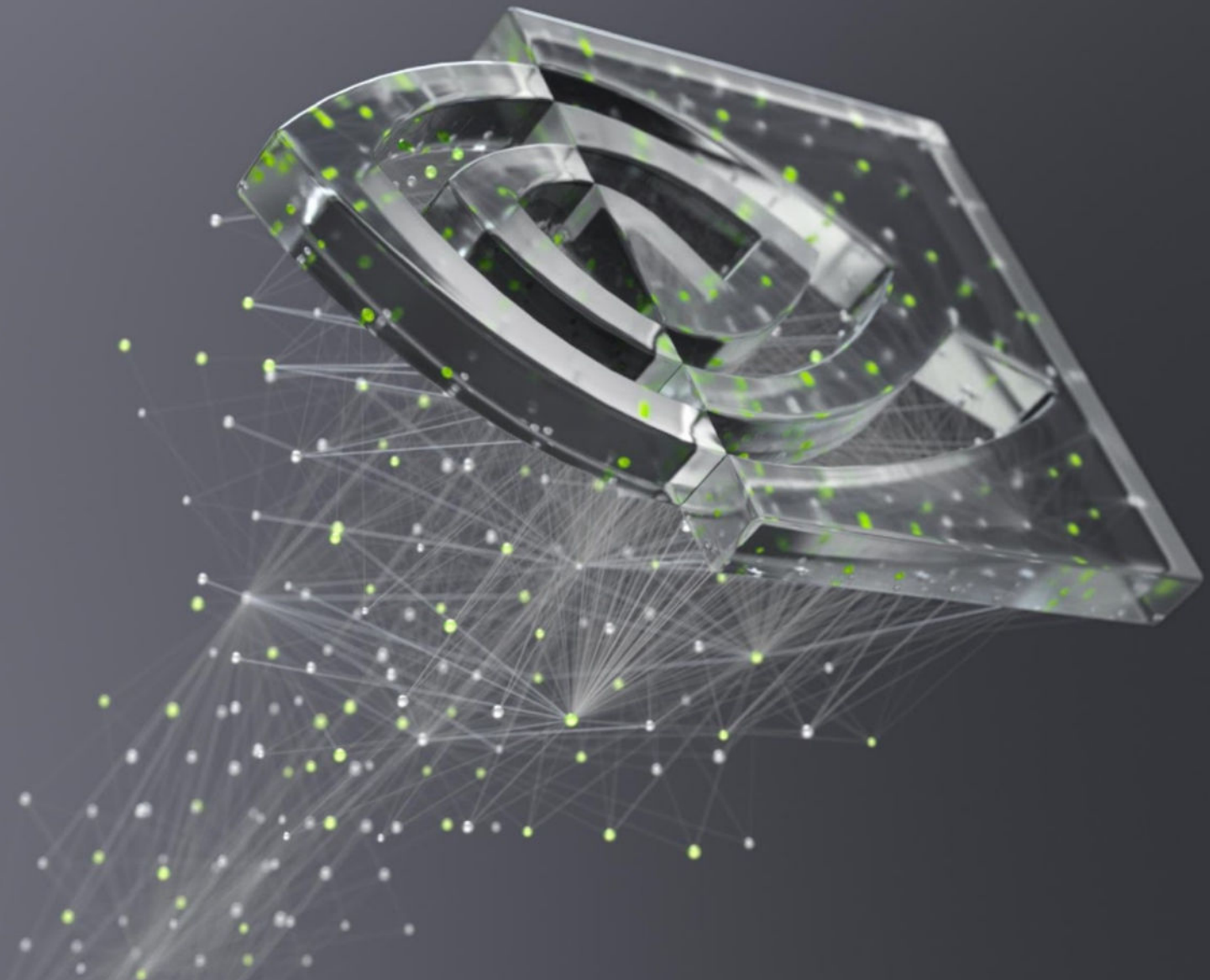
Latest release information

<https://github.com/triton-inference-server/server/releases>

Quick start guide

<https://github.com/triton-inference-server/server/blob/main/docs/quickstart.md>





nVIDIA