NVIDIA TRITON INFERENCE SERVER

Quick Reference Guide

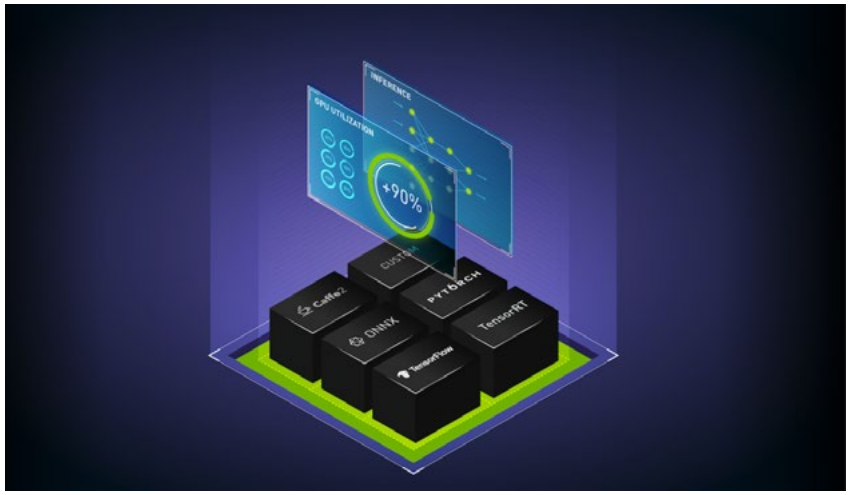
DEPLOY, RUN, AND SCALE AI MODELS IN PRODUCTION—WITH EASE

NVIDIA Triton™ Inference Server is an open-source inference serving software that helps standardize model deployment and execution to deliver fast, scalable AI in production from any framework on any GPU- or CPU-based infrastructure.

Triton executes multiple models concurrently on a single GPU or CPU to deliver high throughput and utilization.

It also optimizes serving for real-time inference under strict latency constraints. Models can be updated live in production without restarting Triton or the application. Triton enables multi-GPU, multi-node inference on very large models that cannot fit in a single GPU's memory.

Triton Model Analyzer is an offline tool to help select the optimal deployment configuration such as batch size, precision, and concurrent execution instances on the target processor to meet application's latency, throughput, and memory requirements.



DEPLOY MODELS WITH FLEXIBILITY

Any Model

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, graph neural networks (GNNs), decision trees, and more

Any Framework

TensorFlow, PyTorch, NVIDIA® TensorRT™, ONNX, XGBoost, OpenVINO, and more

Any Query Type

Real time, batch, audio and video streaming, ensembles

Any Processor

NVIDIA GPUs, Multi-Instance GPU (MIG), x86, and Arm® CPUs

Any Deployment Platform

Bare metal or virtualized, Kubernetes, MLOps platforms, and more

Any Deployment Location

Public cloud, On-prem Data Center, Enterprise Edge, Embedded devices