



TensorRT Inference Server

Production Inference for the Data Center

Maximize data center utilization with concurrent execution of multiple frameworks and models

OVERVIEW

The NVIDIA TensorRT inference server is a production-ready inference microservice that maximizes GPU utilization. Many inference solutions are one-off designs that lack the performance and flexibility to be seamlessly deployed in modern production data center environments.

NVIDIA TensorRT Inference Server simplifies the deployment of inference applications in data centers by providing a microservice which enables applications to use AI models in data center production. It supports the top AI frameworks as well as custom backends. It maximizes utilization by running multiple models concurrently, per GPU, as well as across multiple GPUs.

TensorRT Inference Server also seamlessly supports Kubernetes with health and latency metrics and integrates with Kubeflow for simplified deployment.

KEY POINTS

- Maximizes GPU utilization, supporting multiple AI models on a single GPU and multiple GPUs, with intelligent request batching to optimize performance.
- Increases utility by supporting all of the top AI frameworks. Simply drop any combination of models into the model repository and perform inference on them with an API call.
- Provides metrics to seamlessly integrate into DevOps deployments and scale with Kubernetes on NVIDIA GPUs, other autoscalers, load balancers, schedules, etc.