THE EXPLOSION OF AI

Demand for personalized services has led to a dramatic increase in the complexity, number, and variety of AI-powered applications and products. Applications use AI inference to recognize images, understand speech, or make recommendations. To be useful, AI inference has to be fast, accurate, and easy to deploy.

UNDERSTANDING INFERENCE PERFORMANCE

With inference, speed is just the beginning. Performance. Rapidly, a single-digit, picture ideal of inference performance. There are many factors to consider, ranging from programmability to the rate of learning.

The NVIDIA TensorRT Hyperscale Inference Platform delivers at scale. It delivers the best inference performance at scale with the variability to handle the growing diversity of today’s inference.

INSIDE THE NVIDIA TensorRT HYPERSCALE INFERENCE PLATFORM

NVIDIA TensorRT™ is a high-performance inference platform that includes an optimizer, runtime engines, and inference server to deploy applications in production. TensorRT speeds up to 12x applications in production. To optimize, programmability, and rapid language processing.

THE POWER OF NVIDIA TensorRT

NVIDIA® Tesla® T4 GPUs deliver high throughput inference on NVIDIA® TensorRT™, allowing for fast, accurate, and easy deployment. NVIDIA TensorRT T4 helps businesses achieve their AI goals and enables businesses to deploy their missions.

PRODUCTION-READY DATA CENTER INFERENCE

The NVIDIA TensorRT inference server is a high-performance inference server that delivers the best throughput, most accurate results, and easy deployment. NVIDIA TensorRT is a high-performance inference server that delivers the best throughput, most accurate results, and easy deployment.

THE BEST AI PLATFORM.

www.nvidia.com/data-center-inference