NVIDIA

STATE-OF-THE-ART AI

GPU Deep Learning with the
NVIDIA TensorRT Hyperscale Inference Platform

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 of performance. To get a complete picture about inference performance, there are seven factors to consider, ranging from programmability to rate of learning.

PROGRAMMABILITY

LOW LATENCY

ACCURACY

SIZE OF NETWORK

THROUGHPUT

EFFICIENCY

RATE OF LEARNING

The NVIDIA TensorRT Hyperscale Inference Platform delivers on all fronts. It delivers the best inference performance at scale with the versatility to handle the growing diversity of today's networks.

INSIDE THE NVIDIA TensorRT HYPERSCALE INFERENCE PLATFORM

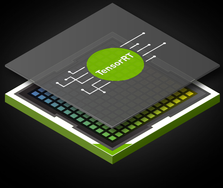
NVIDIA T4 POWERED BY TURING TENSOR CORES

Efficient, high-throughput inference depends on a world-class platform. The NVIDIA® Tesla® T4 GPU is the world's most advanced accelerator for all AI inference workloads. Powered by NVIDIA Turing™ Tensor Cores, T4 provides revolutionary multi-precision inference performance to accelerate the diverse applications of modern AI.

Multi-Precision
FP16 | Up to 65 TFLOPS
INT8 | Up to 130 TOPS
TFLOPS = trillion floating-point operations per second
TOPS = trillion operations per second

THE POWER OF NVIDIA TensorRT

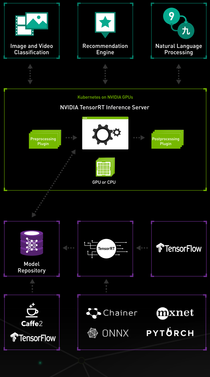
NVIDIA TensorRT™ is a high-performance inference platform that includes an optimizer, runtime engines, and inference server to deploy applications in production. TensorRT speeds apps up to 40X over CPU-only systems for video streaming, recommendation, and natural language processing.



- **Layer and Tensor Fusion**
- **Weight and Activation Precision Calbration**
- **Kernel Auto-Tuning**
- **Dynamic Tensor Memory**
- **Multi-Stream Execution**

PRODUCTION-READY DATA CENTER INFERENCE

The NVIDIA TensorRT inference server is a containerized microservice that enables applications to use AI models in data center production. It maximizes GPU utilization, supports all popular AI frameworks, and integrates with Kubernetes and Docker.



Application Developers
Avoid spending time writing inference capabilities from scratch and focus on creating **innovative solutions with AI.**

DevOps Engineers
Easily deploy inference services for **multiple applications** and take advantage of **orchestration, load balancing, and autoscaling.**

Data Scientists and Researchers
Focus on designing and training models using **any of the top AI frameworks** without worrying about inference implementation.

THE BEST AI PLATFORM.

www.nvidia.com/data-center-inference

NVIDIA