Move Enterprise AI Use Cases From Development to Production With Full-Stack AI Inferencing

Inferencing in the end-to-end AI workstream

The shift from AI training to AI inferencing

AI Training

AI Inference

DATA PROCESSING

Collect, cleanse and transform training data set

TRAIN

Select model and fine tune for desired accuracy

OPTIMIZE

Optimize model for target H/W accelerator

DEPLOY

Deploy on-premise, in the cloud or edge

RUN

Serve model for optimal user experience & TCO

AI inferencing requires a full stack approach

NVIDIA AI Inference Platform optimizes every layer of the stack

Production Runtimes

- Security & reliability
- Technical support
- Ongoing delivery

MLOPS Ecosystem

- Integration with cloud, AI tools, and K8
- Marketplace accessibility
- Retire cloud contractual spend commitments

Application Frameworks

- AI use cases & workflows
- Workload & infra management
- Data curation tools & pretrained models

Model Serving

- Batching of incoming requests
- Multi-instance model deployment
- Concurrent model execution

Compilers, Runtime-Libraries

- Model optimizations (quantization)
- Multi GPU/Node Communication
- Run time optimizations (memory mgmt.)

Hardware Accelerators

- Massive parallel processing
- AI specific H/W accelerators (transformer engines)
- Power efficiency and TCO

NVIDIA

AI Inference Platform

The broadest and deepest selection of GPUs for AI inference

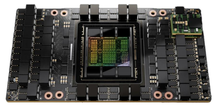
- Production Runtimes
- MLOPS Ecosystem
- Application Frameworks
- Model Serving
- Compilers, Runtime-Libraries
- **Hardware Accelerators**

284x Faster Inference than x86

NVIDIA Grace Hopper

1.9x Faster Inference than H100

NVIDIA HGX H200

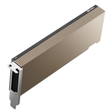


30x Faster Inference than A100

NVIDIA H100

1.7x Faster Inference than HGX A100

NVIDIA L40S



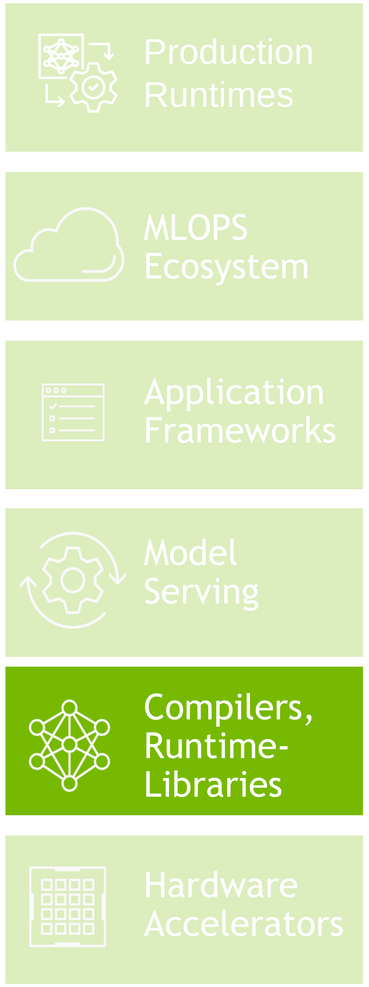
120x Faster AI Video than CPUs

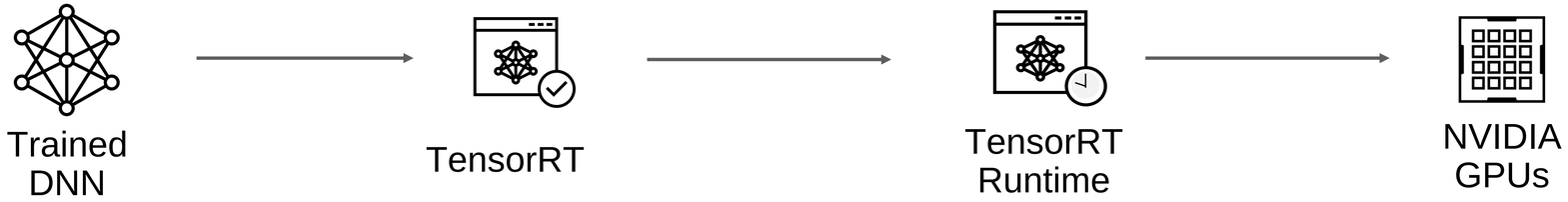
NVIDIA L4

- GH200: 65B LLM, Simulation x86, DGX H100 and Grace Hopper systems
- *H200: Inference on Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 1x GPU BS 8 | H200 1x GPU BS 3*
- *H100: Inference on Megatron 530B parameter model chatbot for input sequence length=128, output sequence length=20 | A100 cluster: HDR IB network | H100 cluster: NDR IB network for 32 A100 vs 16 H100 for 1 sec latency*
- *L40S: ResNet-50 V1.5, INT8, BS=32, 1x A100 SXM, 1x L40S*
- *L4: End-to-end Computer Vision pipeline using CV-CUDA with pre-processing, decode, inference (SegFormer, ResNet-101), encode, post-processing using 8x L4 server, TRT 8.6.0 vs 2S Xeon 8362 Server with OpenCV 4.7*

NVIDIA TensorRT

Accelerate and optimize every network, including CNNs, RNNs, and Transformers





- Reduced mixed precision: FP32, TF32, FP16, and INT8.
- Layer and tensor fusion: Optimizes use of GPU memory bandwidth.
- Kernel auto-tuning: Select best algorithm on target GPU.
- Dynamic tensor memory: Deploy memory-efficient apps.
- Multi-stream execution: Scalable design to process multiple streams.
- Time fusion: Optimizes RNN over time steps.

NVIDIA TensorRT-LLM

Accelerate and optimize inference performance for the latest LLMs

Production Runtimes

MLOPS Ecosystem

Application Frameworks

Model Serving

Compilers, Runtime-Libraries

Hardware Accelerators

State of the Art LLM optimization

Multi GPU multi node inference
Memory bandwidth optimizations
In-flight batching
Paged attention
Windows beta release

Built for NVIDIA GPUs

Hopper
Ada-Lovelace
Ampere
Volta
Turning

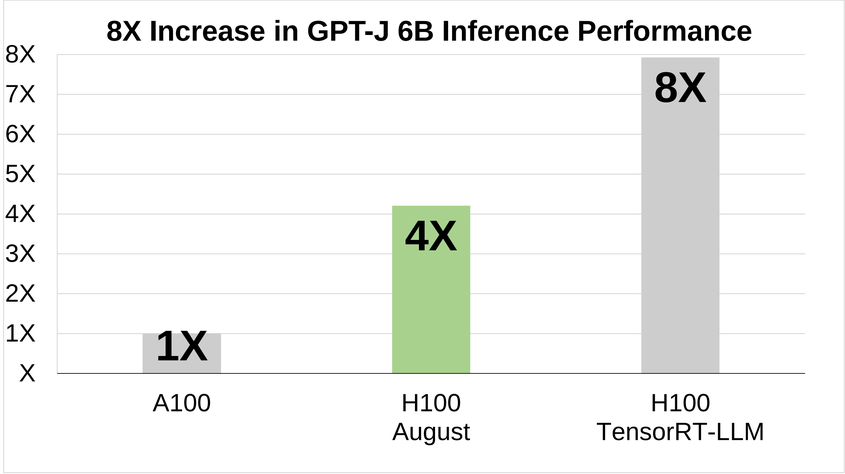
Supports latest LLMs

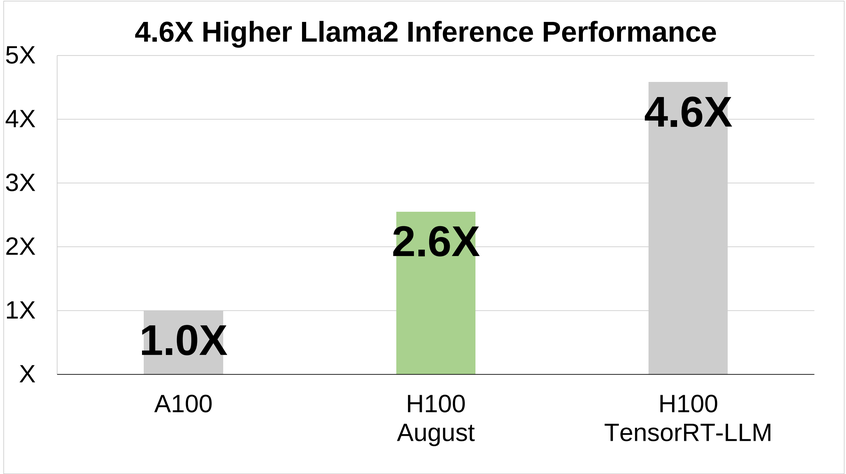
LLaMA-v2
BLOOM
Falcon
GPT-J
GPT-Nemo
Mistral
MPI
and more…

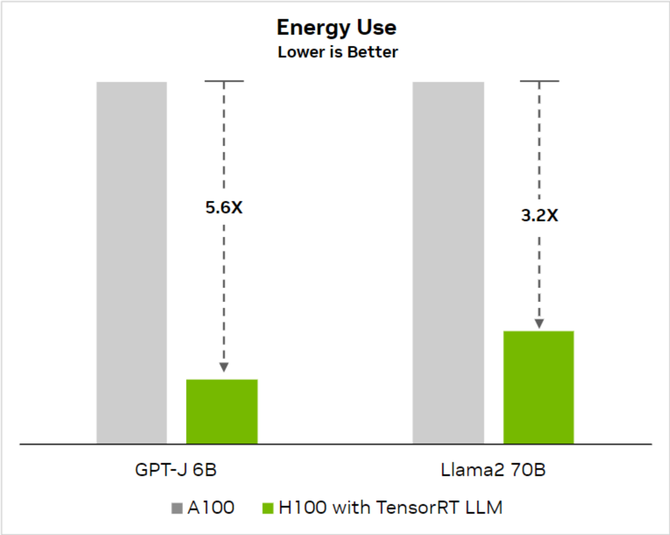
NVIDIA TensorRT-LLM supercharges performance and lowers TCO

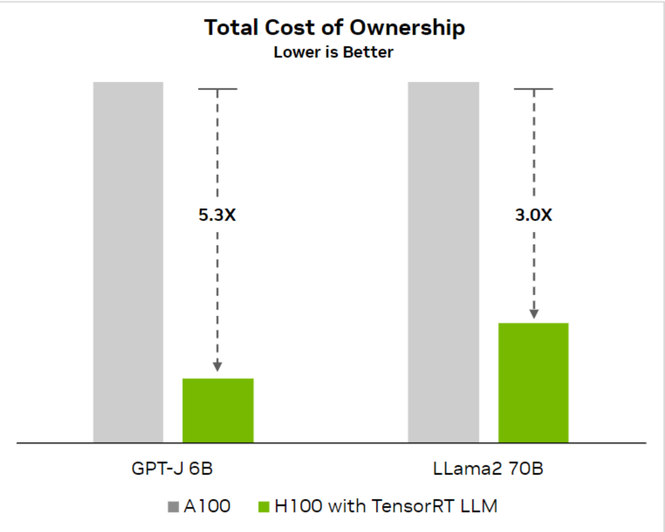
Accelerate Inference and Reduce Energy Usage

- Enterprise Support
- MLOPS Ecosystem
- Application Frameworks
- Model Serving
- **Compilers, Runtime-Libraries**
- Hardware Accelerators









H100 TensorRT-LLM results for September 2023
Text summarization, variable input/output length, CNN / DailyMail dataset |
A100 FP16 PyTorch eager mode | H100 FP8 | H100 FP8, TensorRT-LLM, in-flight batching

NVIDIA Triton Inference Server

The most versatile cross-platform and fully-featured inferencing server

- Enterprise Support
- MLOPS Ecosystem
- Application Frameworks
- **Model Serving**
- Compilers, Runtime-Libraries
- Hardware Accelerators

Any Framework

TensorFlow, PyTorch, ONNX, XGBoost, OpenVINO, Python, Custom

Any Model

Deep learning, tree-based (XGBoost, scikit-learn etc.), model ensembles

Any Query Type

Real time, Batched, Audio and Video streaming

Any HW or Location

CPU or GPU on Cloud, on-prem, edge and embedded

Any Platform

ML platform, Kubernetes, Virtualized | CPU metrics

High Performance

Concurrent model execution, Model analyzer, ++

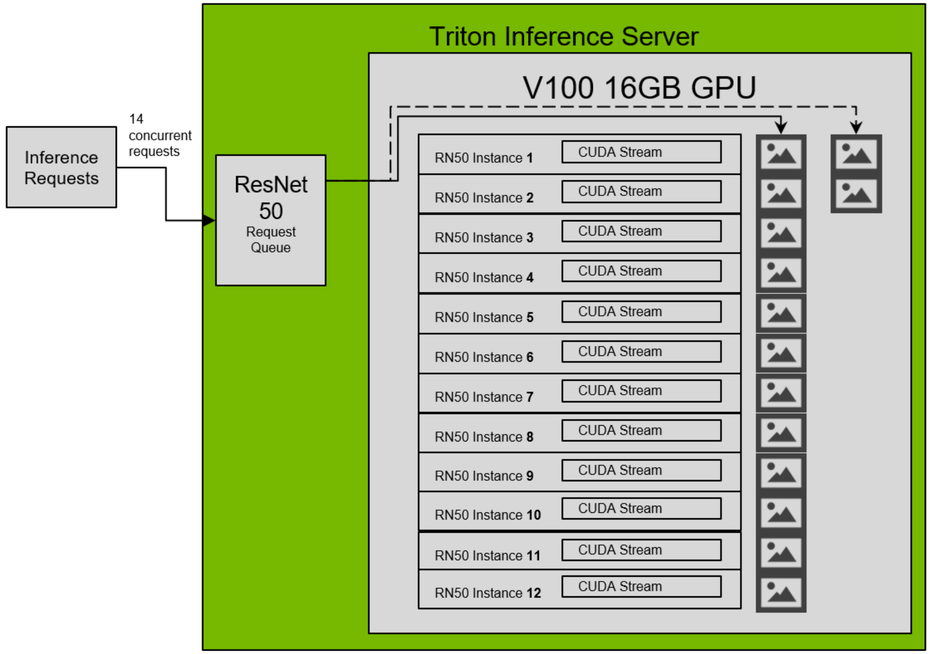
NVIDIA Triton Inference Server

The most versatile cross-platform and fully-featured inferencing server

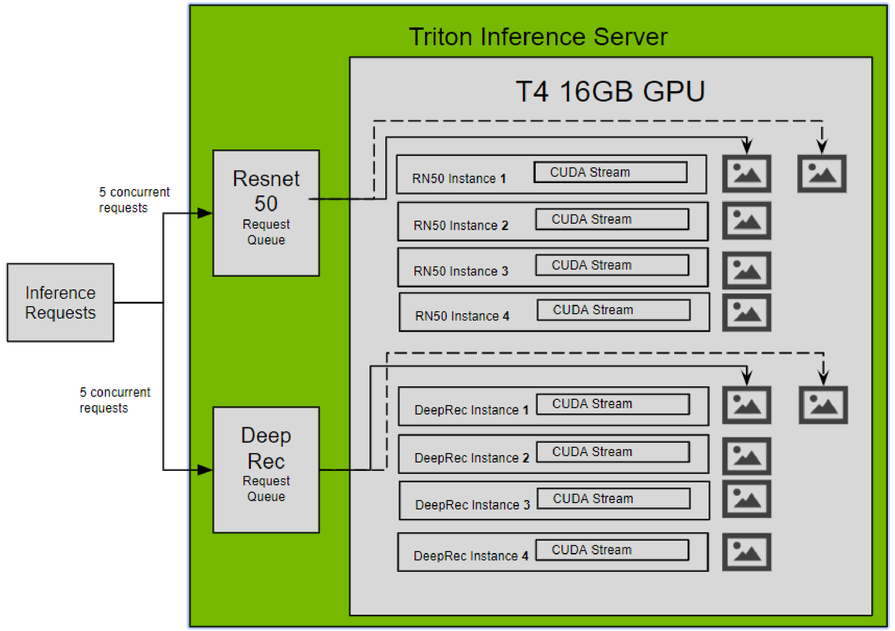
- Enterprise Support
- MLOPS Ecosystem
- Application Frameworks
- **Model Serving**
- Compilers, Runtime-Libraries
- Hardware Accelerators

Concurrent Model Execution

Example 1
Concurrently run max number of instances of a single model on one GPU



Example 2
Concurrently run max number of instances of multiple models on one GPU



NVIDIA Triton Inference Server

The most versatile cross-platform and fully-featured inferencing server

Enterprise Support

MLOPS Ecosystem

Application Frameworks

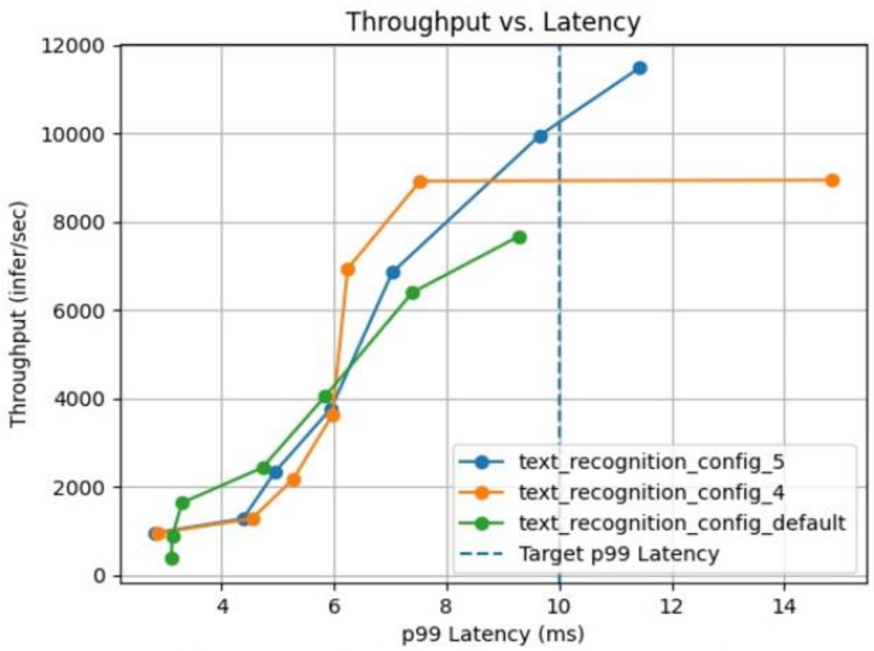
Model Serving

Compilers, Runtime-Libraries

Hardware Accelerators

Model Analyzer

- Run customizable configuration sweeps to meet SLAs
- Identify best configuration for optimal performance under constraints
- Analyze results through reports and charts



Throughput vs. Latency curves for 3 best configurations.

NVIDIA Triton Inference Server

The most versatile cross-platform and fully-featured inferencing server

Enterprise Support

MLOPS Ecosystem

Application Frameworks

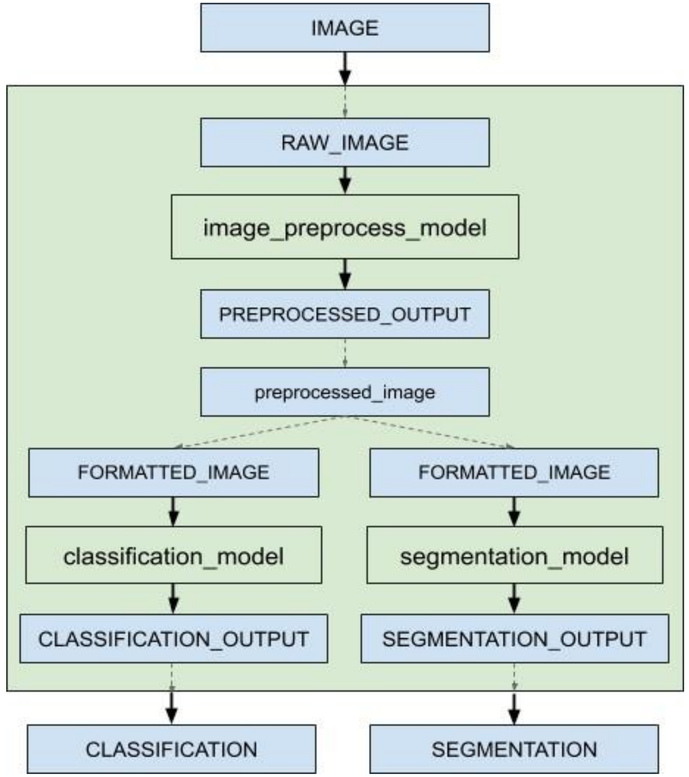
Model Serving

Compilers, Runtime-Libraries

Hardware Accelerators

Model Ensembles

- Serve pipelines of one or more models
- Shares GPU memory to optimize performance
- Each model in pipeline can run on different framework and H/W



NVIDIA Triton Inference Server

The most versatile cross-platform and fully-featured inferencing server

Enterprise Support

MLOPS Ecosystem

Application Frameworks

Model Serving

Compilers, Runtime-Libraries

Hardware Accelerators

Deploy Triton and serve models in 3 easy steps

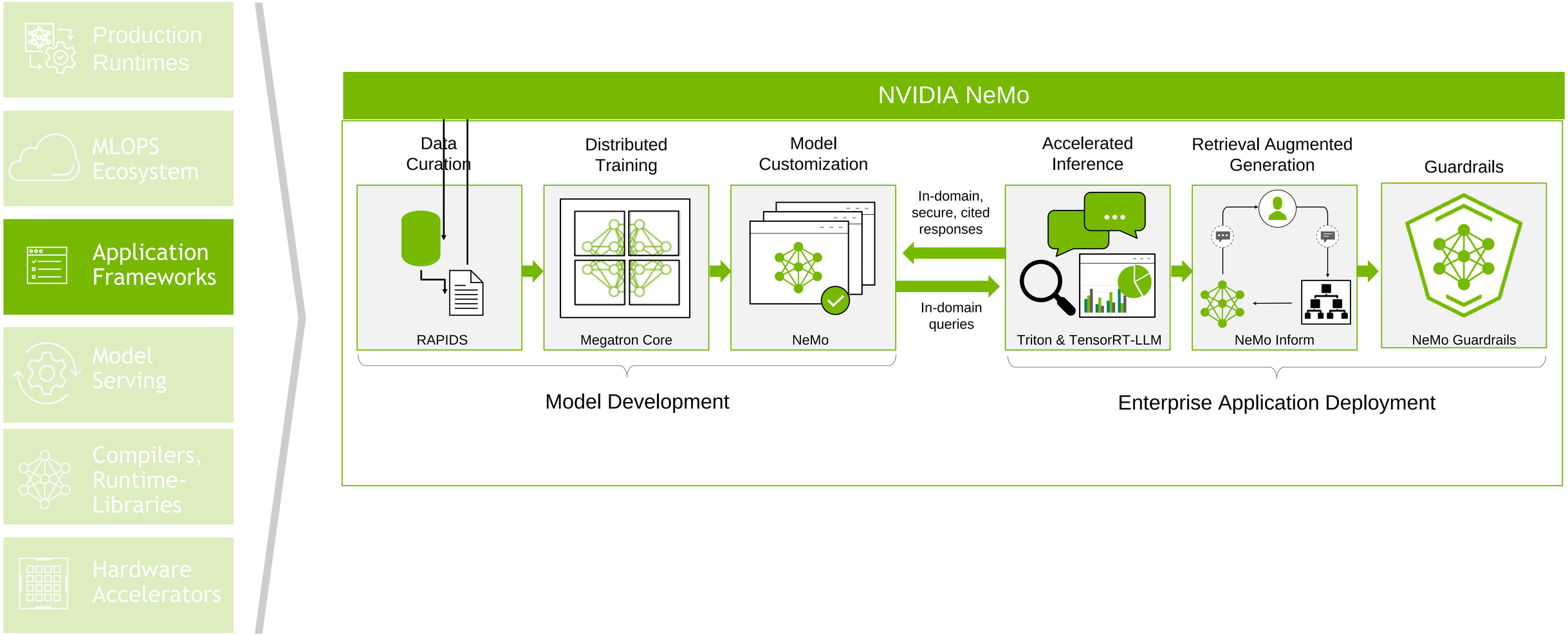
```
# Step 1: Create the example model repository
git clone -b r23.10 https://github.com/triton-inference-server/server.git
cd server/docs/examples
./fetch_models.sh

# Step 2: Launch triton from the NGC Triton container
docker run --gpus=1 --rm --net=host -v ${PWD}/model_repository:/models
nvcr.io/nvidia/tritonserver:23.10-py3 tritonserver --model-repository=/models

# Step 3: Sending an Inference Request
docker run -it --rm --net=host nvcr.io/nvidia/tritonserver:23.10-py3-sdk
/workspace/install/bin/image_client -m densenet_onnx -c 3 -s INCEPTION
/workspace/images/mug.jpg

# Inference should return the following
Image '/workspace/images/mug.jpg':
    15.346230 (504) = COFFEE MUG
    13.224326 (968) = CUP
    10.422965 (505) = COFFEEPOT
```

Image classification example with Densenet

COFFEE MUG

NVIDIA NeMo for building generative AI applications

Build, customize and deploy generative AI models



NVIDIA AI Inference Platform in the cloud

Broad and deep support of the NVIDIA Inference Platform in the cloud

- Production Runtimes
- **MLOPS Ecosystem**
- Application Frameworks
- Model Serving
- Compilers, Runtime-Libraries
- Hardware Accelerators

	aws	Google Cloud	Microsoft Azure	ORACLE CLOUD Infrastructure
NVIDIA Optimized Cloud Machine Images	• NVIDIA GPU optimized AMI • Free with option to purchase enterprise support through NVIDIA AI Enterprise	• NVIDIA Cloud Native Stack VMI • Free with option to purchase enterprise support through NVIDIA AI Enterprise	• NVIDIA GPU-Optimized VMI • Free with option to purchase enterprise support through NVIDIA AI Enterprise	• NVIDIA GPU Cloud Machine Image • Free with option to purchase enterprise support through NVIDIA AI Enterprise
NVIDIA Triton Inference Server	• Available on AWS Deep Learning Containers • Deploy with SageMaker Python SDK, AWS CLI ,Boto3 • Supports AWS SageMaker Multi-Model Endpoint (MME) API Contract	• Vertex AI Prediction supports deploying models on Triton running on custom NGC container • Deploy Triton as a containerized microservice on GKE managed cluster using One-Click Triton Inference Server App for GKE	• Deploy using Azure CLI, Python SDK v2, and Azure ML studio (specify Type as “triton_model” in YAML deployment file) • Supports No-code deployment in managed online endpoints and Kubernetes online endpoints	• Seamless integration with OCI Data Science’s model deployment (pass Triton as env variable) • Push Triton image to OCI Container Registry and save the model to the model catalog • Use with OCI software developer kits (SDKs), APIs, or the Oracle Cloud Console
NVIDIA GPU Powered Cloud Instances	• NVIDIA H100 | P5 • NVIDIA A100 | P4D • NVIDIA V100 | P3 • NVIDIA A10G | G5 • NVIDIA T4(G) | G4, G5g • NVIDIA H200 | Coming Soon • Just announced GH200, L40S, L4	• NVIDIA H100 | A3 • NVIDIA L4 | G2 • NVIDIA A100 | A2 • NVIDIA V100 | N1 • NVIDIA T4 | N1 • NVIDIA H200 | Coming Soon	• NVIDIA H100 | ND H100 v5, NCads H100 v5 • NVIDIA A100 | ND, NC v4 • NVIDIA V100 | NC v3, ND v2 • NVIDIA A10 | NVadsA10 v5 • NVIDIA T4 | NCasT4_v3 • NVIDIA H200 | Coming Soon	• NVIDIA H100 | BM.GPU.H100.8 • NVIDIA A100 | GPU.A100 • NVIDIA A10 | VM.GPU.A10 • NVIDIA V100 | VM.GPU3 • NVIDIA L40S | Coming soon • NVIDIA H200 | Coming Soon • NVIDIA GH200 | Coming Soon

1. Cloud service providers listed in alphabetical order. Please visit CSP’s respective website for latest NVIDIA GPU powered compute Instances at www.aws.com , www.cloud.google.com, www.azure.Microsoft.com, and www.oracle.com

NVIDIA AI Enterprise

Enterprise-grade software platform for uninterrupted AI runtimes

- **Production Runtimes**
- MLOPS Ecosystem
- Application Frameworks
- Model Serving
- Compilers, Runtime-Libraries
- Hardware Accelerators

MLOps

AI Applications

NVIDIA AI Enterprise for Inference

Infrastructure Management

Cloud Native Management and Orchestration
GPU Operator, Network Operator

Cluster Management
Base Command Manager Essentials

Infra Acceleration Libraries
Magnum IO, vGPU, CUDA

Application Frameworks

LLM	Speech AI	Cybersecurity	Medical Imaging	More
NeMo	Riva	Morpheus	Clara	•••

AI Development

Data Science / Prep
RAPIDS, RAPIDS Accelerator for Apache Spark

Model Training and Customization
NeMo, TAO, PyTorch, TensorFlow

Optimize for Inference
TensorRT, TensorRT-LLM

Deploy at Scale
Triton Inference Server

Cloud | Data Center | Workstations | Edge

THANK YOU