



INNOVATE

DATA AND AI/ML EDITION

22 February 2023

Deep learning on AWS with NVIDIA: From training to deployment

Michael Lang

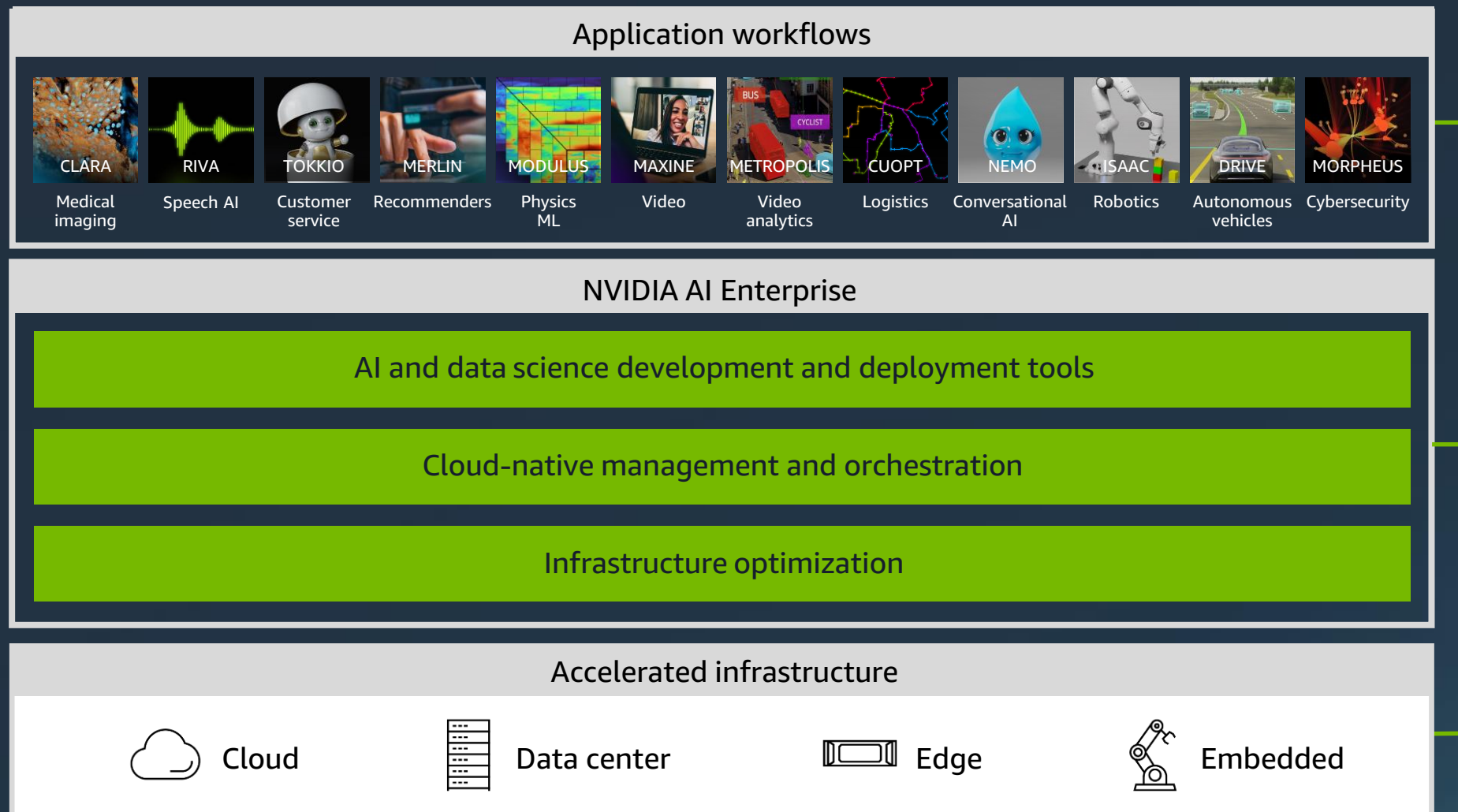
Solutions Architect Manager – APAC South
NVIDIA

Agenda

- NVIDIA and AWS relationship
- NVIDIA AI on AWS
- ML model training (at scale)
- ML model deployment and inference
- Conclusion
- Next steps

NVIDIA AI

End-to-end open platform for production AI



NVIDIA
LaunchPad



Hands-on
labs

NVIDIA and AWS relationship

GPU power from the cloud to the edge

Machine learning



ML training and cost-effective inference

Virtual workstations



Work from anywhere

High-performance compute



Solve large computational problems

Internet of things



Extend AI/ML to edge devices that act locally

Powerful | Cost-Effective | Flexible

<https://aws.amazon.com/nvidia/>

AWS and NVIDIA

GPU power from the cloud to the edge

[Apply for a Free Trial with NVIDIA on AWS »](#)

What's New

Get started with digital content creation on AWS
[Read the blog »](#)

AWS AND NVIDIA

Overview

Workloads

Machine Learning

Virtual Workstations

High Performance Compute

Internet of Things

Case Studies

Services

Additional Resources

AWS and NVIDIA have collaborated for over 10 years to continually deliver powerful, cost-effective, and flexible GPU-based solutions for customers. These innovations span from the cloud, with NVIDIA GPU-powered Amazon EC2 instances, to the edge, with services such as AWS IoT Greengrass deployed with NVIDIA Jetson Nano modules.

Customers around the world are using AWS and NVIDIA solutions for machine learning (ML), virtual workstations, high performance computing (HPC), and IoT services. Amazon EC2 instances powered by NVIDIA GPUs deliver the scalable performance needed for fast ML training, cost-effective ML inference, flexible remote virtual workstations, and powerful HPC computations. At the edge, customers can use AWS IoT Greengrass to extend a wide range of AWS cloud services to NVIDIA-based edge devices so the devices can act locally on the data they generate.



Introducing Amazon EC2 P4d instances (2:00)



GPU power from the cloud to the edge



The highest-performance instance for ML training and HPC applications powered by NVIDIA A100 GPUs



High-performance instances for graphics-intensive applications and ML inference powered by NVIDIA A10G GPUs



The best price performance in Amazon EC2 for graphics workloads powered by NVIDIA T4G GPUs



Deploy fast and scalable AI with NVIDIA Triton Inference Server in Amazon SageMaker



Improve your operations with computer vision at the edge powered by NVIDIA Jetson



Spot defects with automated quality inspection powered by NVIDIA Jetson



NVIDIA GPU-optimized software available for free on the NVIDIA NGC portal.

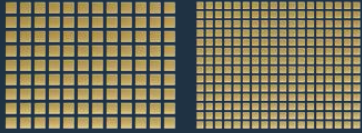
NVIDIA AI on AWS



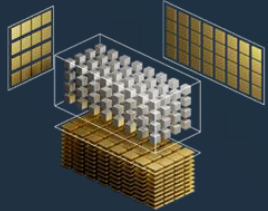
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

NVIDIA A100

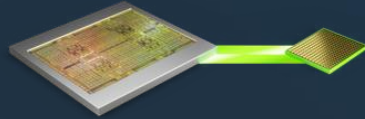
Supercharging High performing ai supercomputing gpu



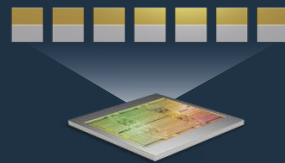
80 GB HBM2e
For largest datasets
and models



3rd-gen Tensor core



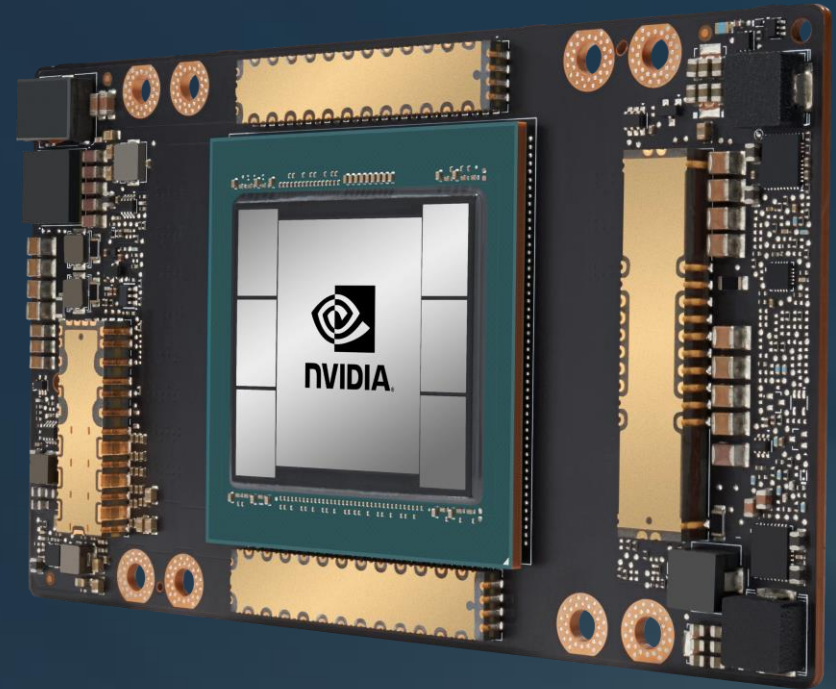
2 TB/s +
High-memory bandwidth
to feed extremely fast GPU



Multi-instance GPU



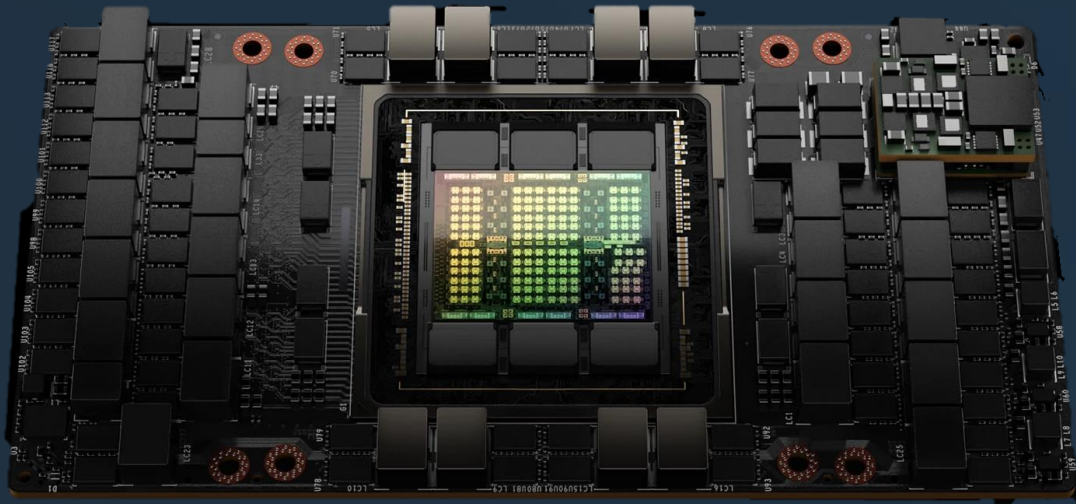
3rd-gen NVLink



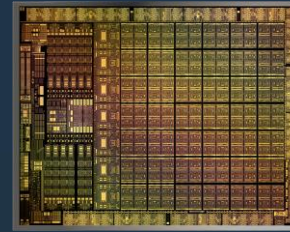
Powering Amazon EC2 P4d/P4de instances

NVIDIA H100 – Coming soon to AWS

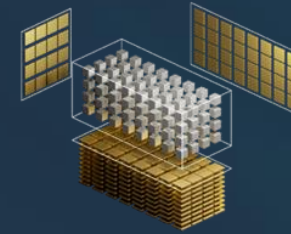
The new engine of the world's AI infrastructure



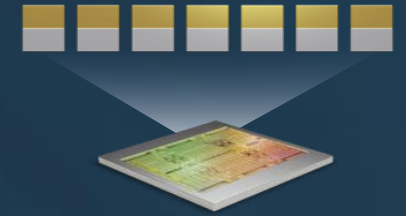
Powering the next generation of GPU systems on AWS



Advanced
chip



Transformer
engine



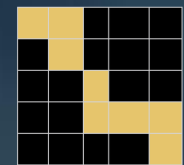
2nd-gen MIG



Confidential
computing



4th-gen
NVLink

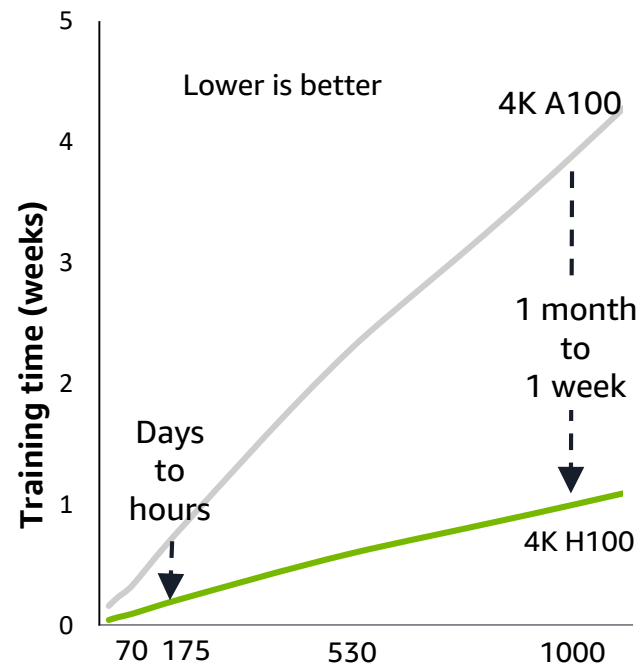


DPX instructions

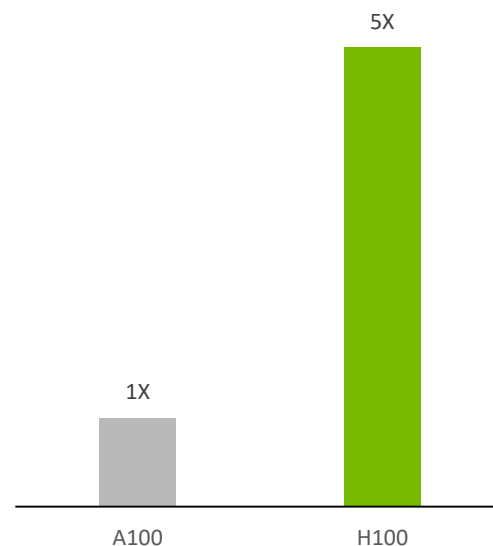
NVIDIA H100 supercharges large language models

Hopper architecture addresses LLM needs at scale

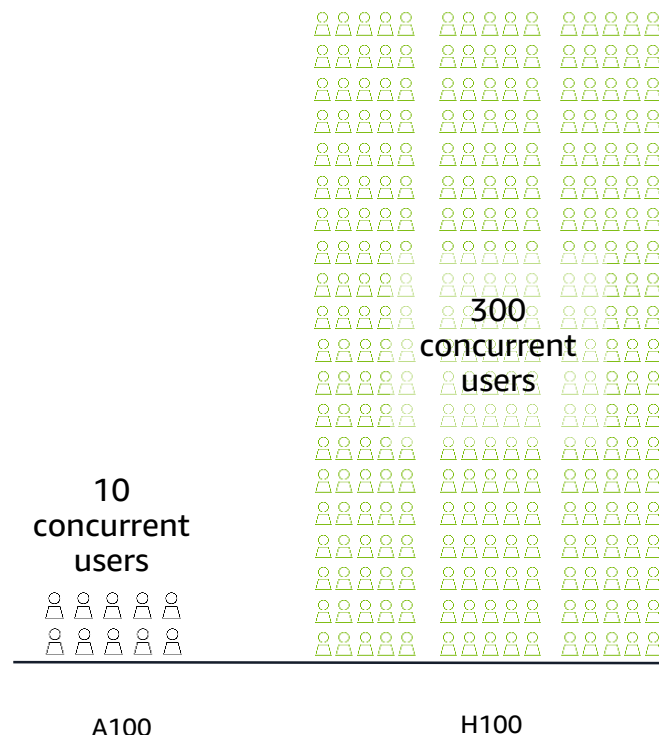
Supercharged LLM training



High-performance prompt learning



30x real-time inference throughput



LLM Training | 4,096 GPUs | H100 NDR IB | A100 HDR IB | 300 billion tokens

P-tuning | DGX H100 | DGX A100 | 530B Q&A tuning using SQuAD dataset

Inference | Chatbot | 10 DGX H100 NDR IB | 10 DGX A100 HDR IB | <1 second latency | 1 inference/second/user

H100 data center projected workload performance, subject to change



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

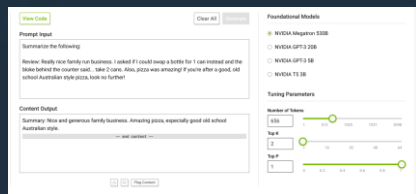


NGC

Portal to AI services, software, support

Cloud services

End-to-end AI development



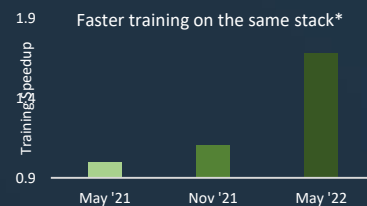
AI services for NLP, biology, speech



AI workflow management & support

Performance optimized

Tested across
GPU-accelerated platforms



Monthly SW container updates



SOTA models

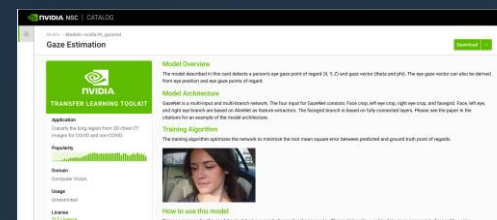
NGC catalog

Fully transparent

Quickly find and
deploy the right SW

vulnerabilities	OS package	Medium	(CVE-2021-3995) libmount1
vulnerabilities	OS package	Medium	(CVE-2021-3995) fdisk
vulnerabilities	OS package	Medium	(CVE-2019-9152) hdf5-helpers
vulnerabilities	OS package	Medium	(CVE-2018-17233) hdf5-helpers

Detailed security scan reports



Model resumes

Accelerates development

Focus on building, not setup



One-click deploy from NGC

Multiple cloud providers

Develop once; deploy
anywhere with NVIDIA VMI

ngc.nvidia.com



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Amazon EC2 instances powered by NVIDIA GPUs

Accessible via AWS, AWS Marketplace, and AWS services

NVIDIA GPU	AWS instance	GA	Use case recommendations	Regions	GPU memory	GPUs	On-demand price/hour
T4g	G5g	11/2021	Graphic workloads such as Android game streaming, ML inference, graphics rendering, and AV simulation	5	16 GB	1, 2	\$0.42
A10G	G5	11/2021	Best performance for graphics, HPC, and cost-effective ML inference	3	24 GB	1, 4, 8	\$1.00
A100	P4d, P4de	11/2020	Best performance, ML training, HPC across industries	8	40, 80 GB	8	\$32.77
V100	P3, P3dn	10/2017	ML training, HPC across industries	14+	16, 32 GB	1, 4, 8	\$3.06–\$31.21
T4	G4	9/2019	The universal GPU, ML inference, training, remote visualization workstations, rendering, video transcoding <i>Includes Quadro Virtual Workstation</i>	20+	16 GB	1, 4, 8	\$0.52–\$7.82

EC2 G5g is now available in US East (N. Virginia), US West (Oregon), and Asia Pacific (Tokyo, Seoul, and Singapore) Regions; On-Demand, Reserved, and Spot pricing available

EC2 G5 is now available in US East (N. Virginia), US West (Oregon), and Europe (Ireland) Regions; On-Demand, Reserved, Spot, or as part of Savings Plans

EC2 P4d is now available in US East (N. Virginia and Ohio), US West (Oregon), Europe (Ireland and Frankfurt), and Asia Pacific (Tokyo and Seoul) Regions; On-Demand, Reserved, Spot, Dedicated Hosts, or Savings Plans availability

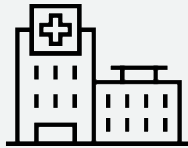


Training computer vision and conversational AI

Proliferation of use cases

Healthcare

Patient monitoring
Smart hospitals
Robot-assisted surgery



Industrial manufacturing

Automated optical inspection
Worker safety
Process automation



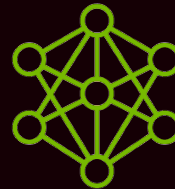
Retail

Detecting people movement
Analyzing action
Warehouse logistics

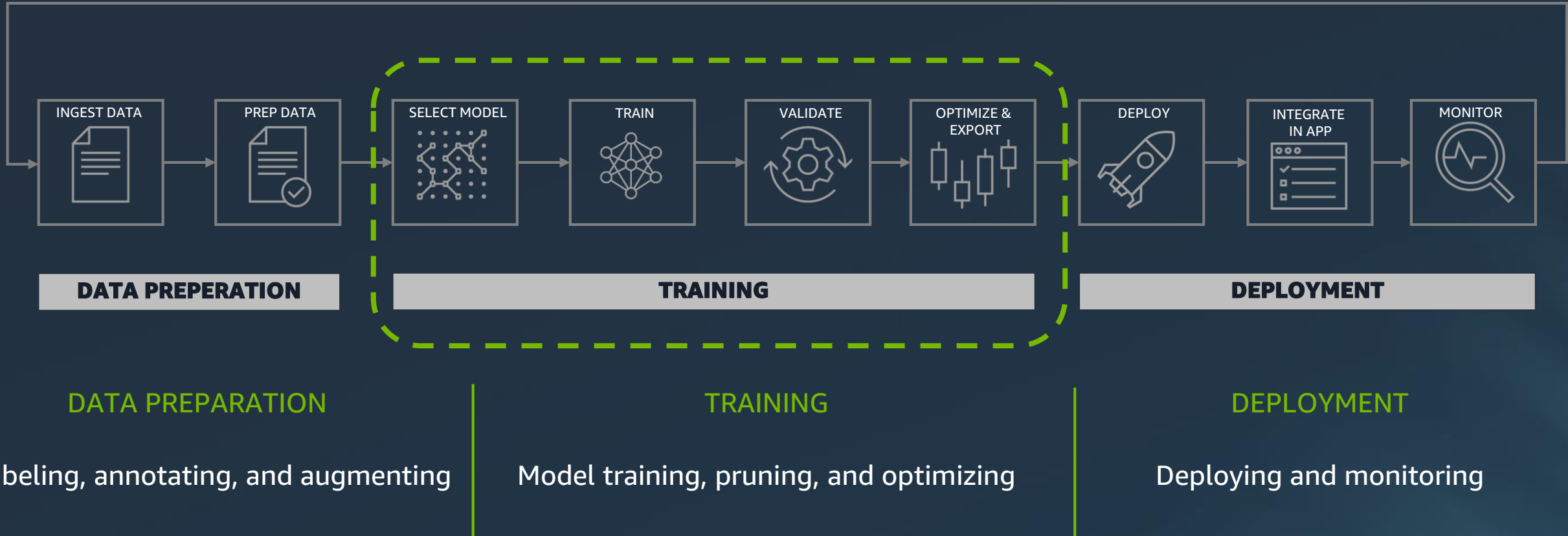


Smart infrastructure

Pedestrian safety
Traffic management
Waste management



Creating an AI application is hard and complex



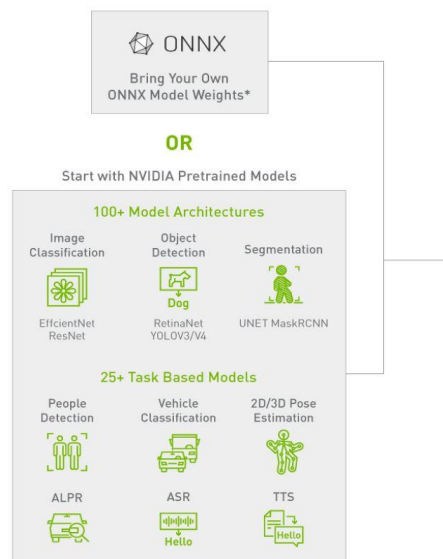
Get started today with the TAO Toolkit: <https://developer.nvidia.com/tao-toolkit-get-started>

NVIDIA TAO Toolkit

Train, adapt, optimize

Create custom, production-ready AI models in hours rather than months

- 1 Bring your own model weights or choose from NVIDIA's library of model architectures or task-based models



How can I run this?

- Containerized on Amazon EC2
- Containerized with Amazon EC2
- Bring-your-own-container on Amazon SageMaker

All available from the NGC catalog

TRAIN EASILY

Fine-tune NVIDIA pretrained models with a fraction of the data

CUSTOMIZE FASTER

Built on TensorFlow and PyTorch that abstract away the AI framework complexity

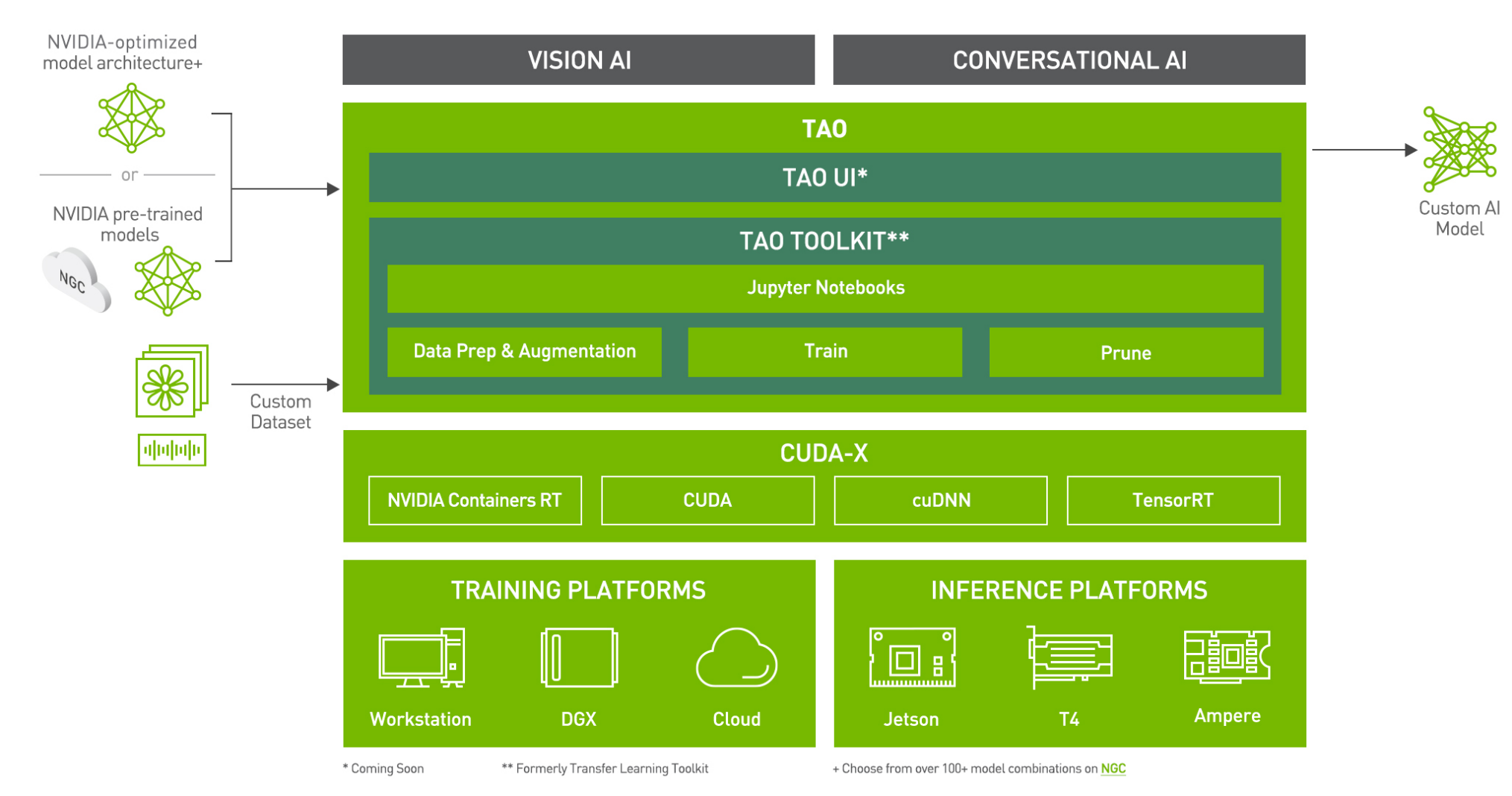
OPTIMIZE FOR DEPLOYMENT

Optimize for inference and integrate with Riva or DeepStream

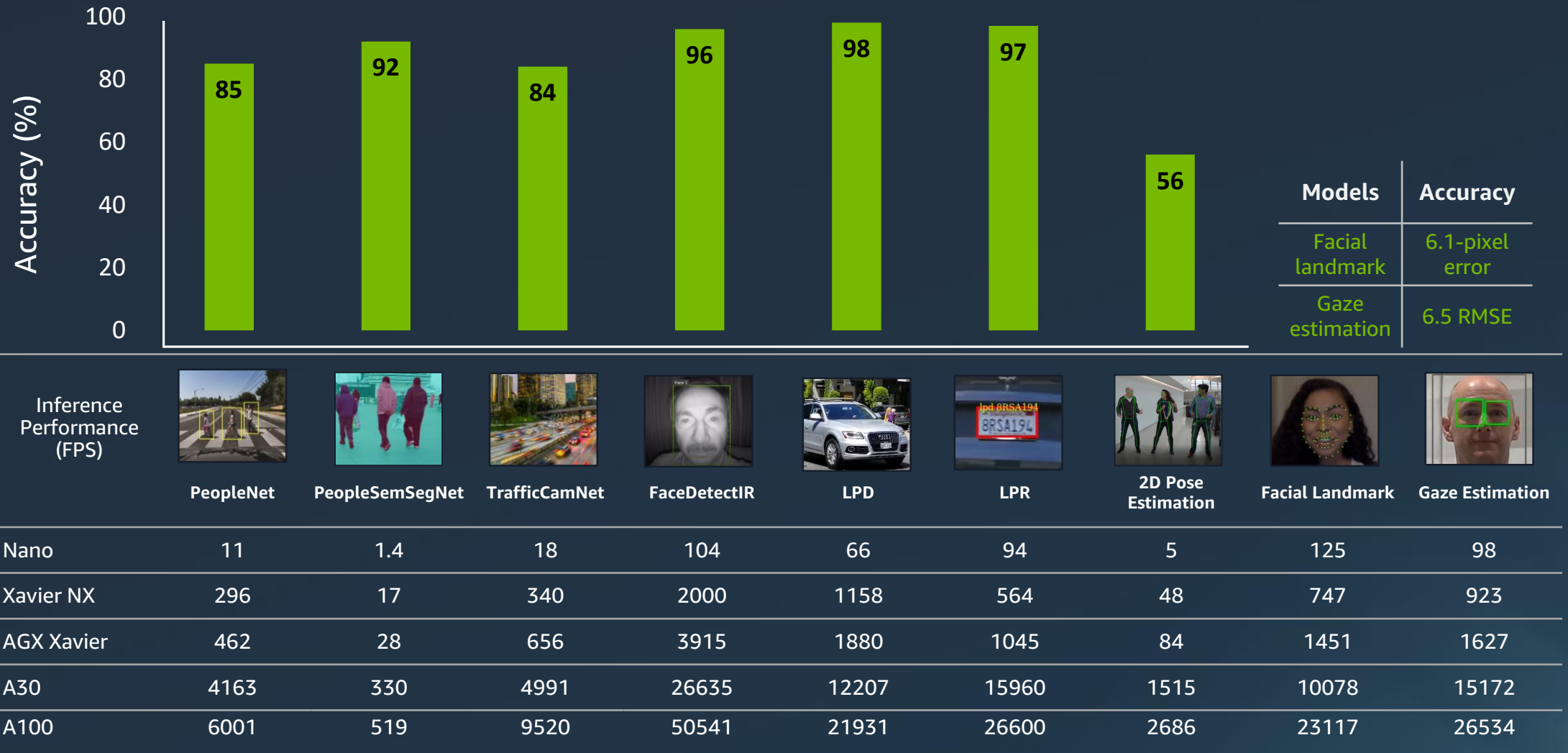
SUPPORTED BY EXPERTS

Supported by NVIDIA experts to help resolve issues from development to deployment

The NVIDIA TAO stack



High-performance pretrained vision AI models



15+ pretrained models – download for free from [NGC](#)

Pretrained conversational AI models



Automatic Speech Recognition

Jasper

QuartzNet

CitriNet

N-Gram

Support for models that are used in the conversational AI pipeline



Natural Language Processing

BERT Punctuation

BERT NER

BERT Text Classification

BERT Intent & Slot

Domain-Specific NER

BERT & Megatron QA

Adapt with your dataset using
NVIDIA TAO Toolkit



Text to Speech

FastPitch

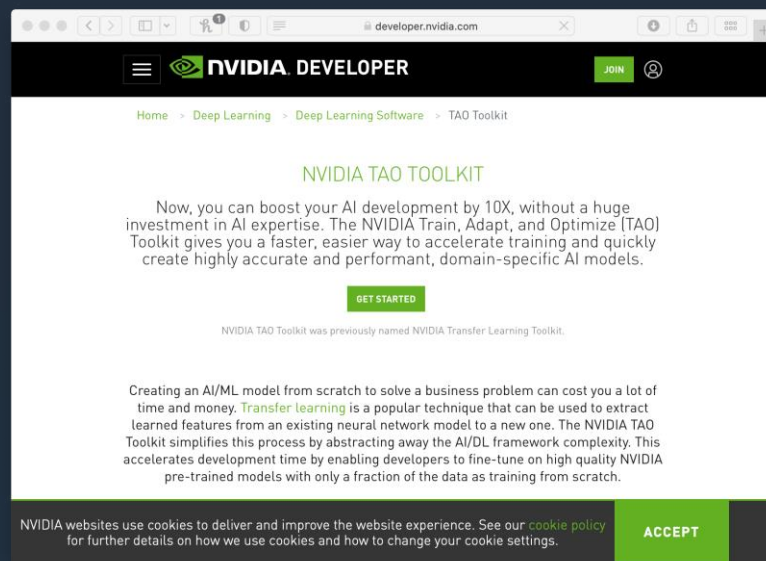
HiFi-GAN

Deploy with turnkey inference applications in **NVIDIA Riva**

<https://developer.nvidia.com/blog/building-and-deploying-conversational-ai-models-using-nvidia-tao-toolkit/>

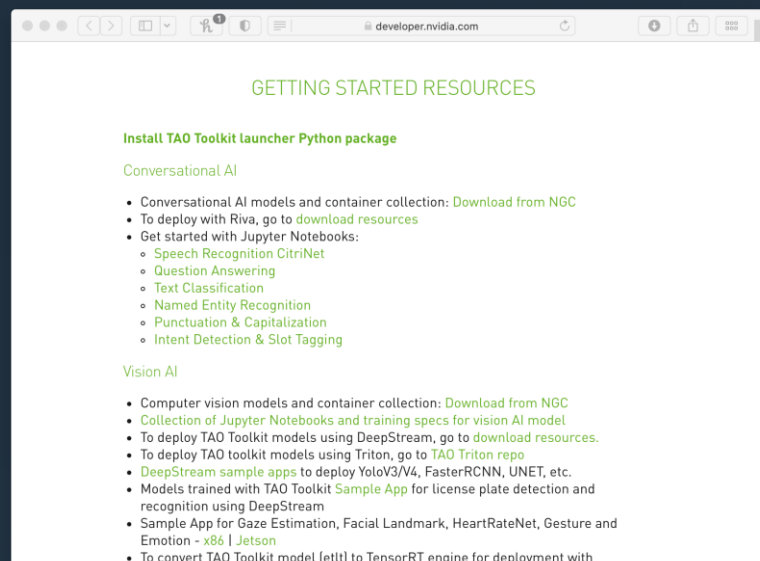
Resources

Getting Started with the TAO Toolkit



TAO Toolkit product page

All information related to product features and developer blogs



TAO Toolkit getting started page

Detailed information on how to get started with the TAO Toolkit



TAO Toolkit whitepaper

Includes examples on data augmentation, adding new classes

Developer resources



2D Pose Estimation
Model with NVIDIA TAO
Toolkit [Part 1](#) | [Part 2](#)



[Supercharge your AI workflow
with TAO Toolkit whitepaper](#)



[Train and deploy action
recognition model](#)



[Building conversational AI models
using the NVIDIA TAO Toolkit](#)

Computer vision

- TAO Toolkit computer vision models and container collection: [Download from NGC](#)
- To deploy TAO Toolkit models using DeepStream, go to [download resources](#)
- [Collection of Jupyter Notebooks and training specs for vision AI models](#)

Conversational AI

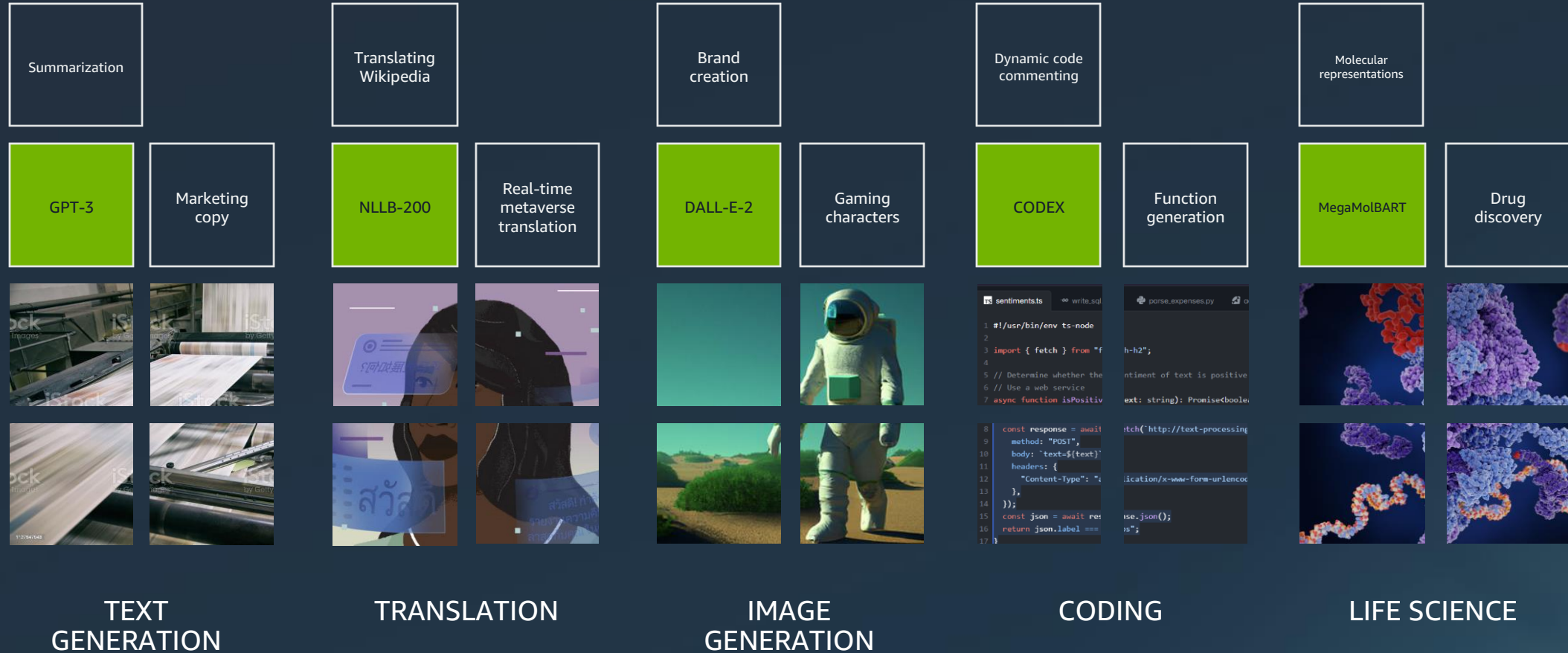
- TAO Toolkit conversational AI models and container collection: [Download from NGC](#)
- To deploy with Riva, go to [download resources](#)
- Get started with Jupyter Notebooks: [Speech Recognition](#) | [Question Answering](#) | [Text Classification](#) | [Named Entity Recognition](#) | [Punctuation & Capitalization](#) | [Intent Detection & Slot Tagging](#)

[TAO TOOLKIT GETTING STARTED PAGE](#)

Training (at scale) large language models

LLMs unlock new opportunities

LLMs transcend language and pattern matching



When large language models make sense

	Traditional NLP approach	Large language models
Requires labeled data	Yes	No
Parameters	100s of millions	Billions to trillions
Desired model capability	Specific (one model per task)	General (model can do many tasks)
Training frequency	Retrain frequently with task-specific training data	Never retrain or retrain minimally

Zero-shot (or few-shot learning)

Painful and impractical to get a large corpus of labeled data

Models can learn new tasks

If you want models with “common sense” and can generalize well to new tasks

A single model can serve all use cases

At scale, you avoid costs and complexity of many models, saving cost in data curation, training, and managing deployment

Training and deploying LLMs is not for the faint of heart

LLMs are challenging to build & Deploy

UNMET NEEDS

Large-scale data processing

Multilingual data processing & training

Finding optimal hyperparameters

Convergence of models

Scaling on clouds

Deploying for inference

Deployment at scale

Evaluating models in industry standard benchmarks

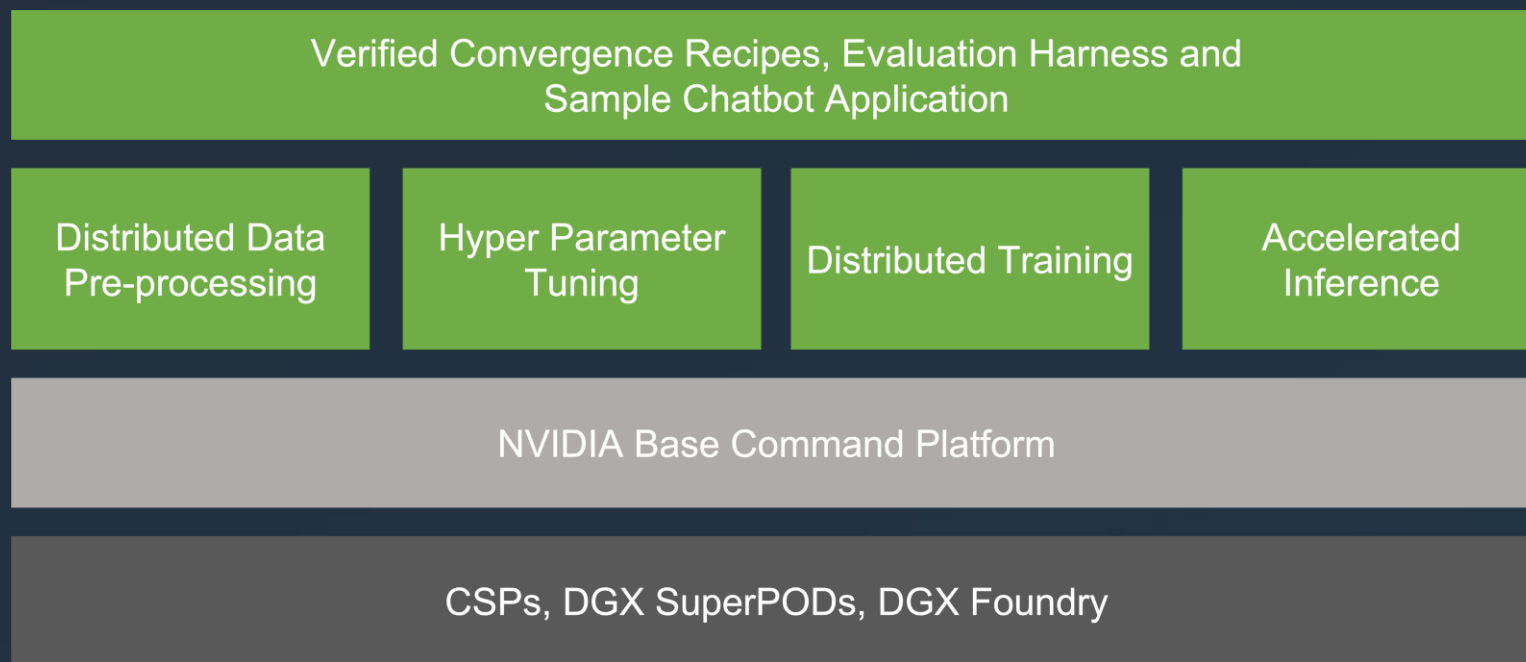
Differing infrastructure setups

Lack of knowledge

- Training and deploying models take months to years
- Requires deep technical expertise
- Extensive compute resources in the scale of 1,000s GPUs for training a 530B model over several months
- Tools to scale to 1,000s of GPUs are limited
- All leading to high financial investments, in the order of tens of millions of dollars for 175B+ models

NeMo Megatron

End-to-end framework for training and deploying large-scale language models with trillions of parameters



- Rapidly create and tune state-of-the-art custom language models
- Linear scaling to 1,000s of GPUs for up to a trillion parameter language models
- 30% speed-up in training using new sequence parallelism and selective activation recomputation techniques
- Distributed inference using Triton Inference Server
- Prompt learning capabilities with P-tuning and prompt tuning

Model availability

Models NVIDIA verified training recipes

GPT-3: 126M, 5B, 20B, 40B, 175B

T5: 220M, 3B, 11B, 23B, 41B

mT5: 170M, 390M, 3B, 11B, 23B

NVIDIA publicly available model checkpoints

T5: 3B

GPT-3: 5B, 20B

Training and inference support for popular community pretrained models (coming in Q4 2022)

Now in open beta

Find out more:

[NVIDIA NeMo Megatron](#)

Solving pain points across the stack

NeMo Megatron simplifies the path to an LLM

Unmet needs

Large-scale data processing

Multilingual data processing and training

Finding optimal hyperparameters

Convergence of models

Scaling on clouds

Deploying for inference

Deployment at scale

Evaluating models in industry-standard benchmarks

Differing infrastructure setups

Lack of knowledge



How we are helping

Data curation and preprocessing tools

Relative positional embedding (RPE) – multilingual support

Hyperparameter tool

Verified recipes for large GPT and T5-style models

Scripts/configs to run on AWS

Model navigator + export to FT functionalities

Quantization to accelerate inferencing

Productization evaluation harness

Full-stack support with FP8 and Hopper support

Documentation

NeMo Megatron

Value Proposition

End-to-end

Bring your own data,
train and deploy LLM



Performance at scale

SOTA training techniques



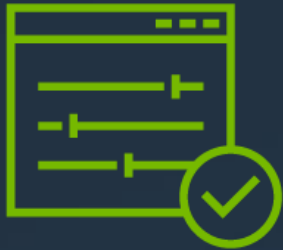
Easy to use

Containerized
framework



Fastest time to solution

Tools and SOTA performance



Customization

Source-open approach



Availability

Train on your choice
of infrastructure



Battle-hardened

Enterprise-grade framework with
verified recipes to work OOTB

- NeMo Megatron is an end-to-end application framework for training and deploying LLMs with billions and trillions of parameters
- Turnkey containerized framework with recipes for training and deploying GPT-3 (up to 1T parameters), T5, and mT5 (up to 50B parameters) style models

Training container

Inference container

Resources

GETTING STARTED

[Register here for open beta](#)

[NVIDIA NeMo Megatron](#)

[NVIDIA brings large language AI Models to enterprises worldwide | NVIDIA newsroom](#)

DEV BLOGS

[Adapting P-Tuning to solve non-english downstream tasks](#)

[NVIDIA AI platform delivers big gains for large language models](#)

[Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, the world's largest and most powerful generative language model | NVIDIA developer blog](#)

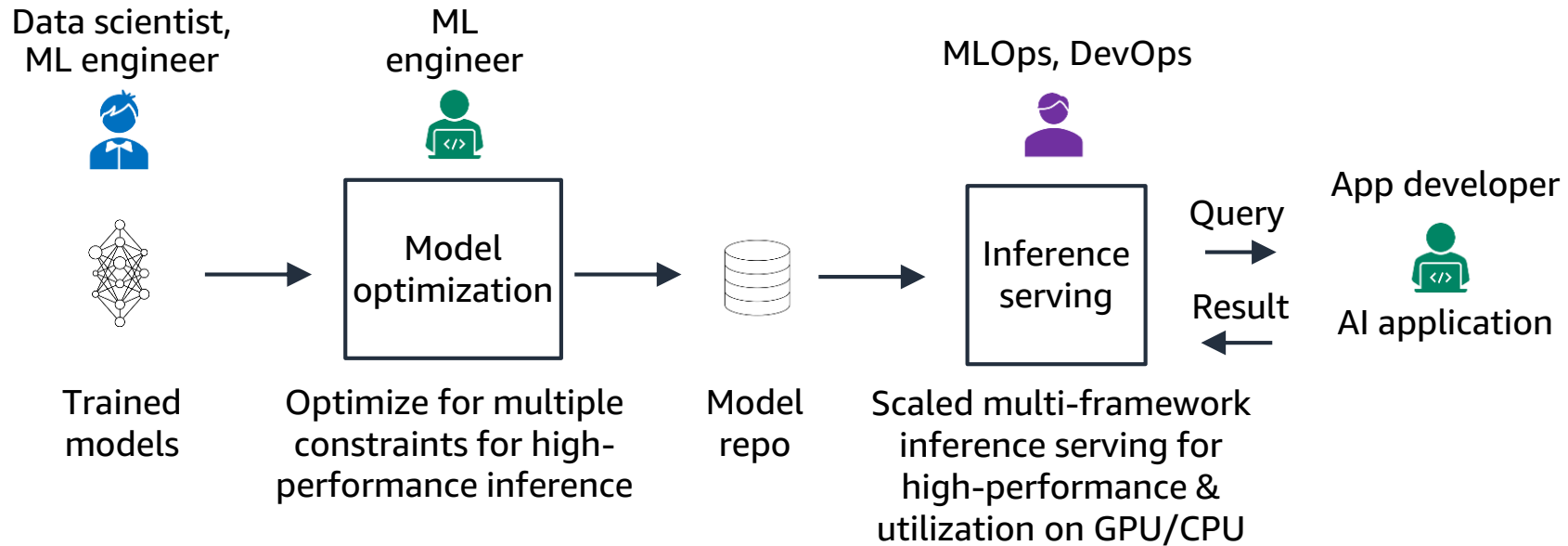
CUSTOMER STORIES

[The King's Swedish: AI rewrites the book in Scandinavia](#)

Deployment and inference

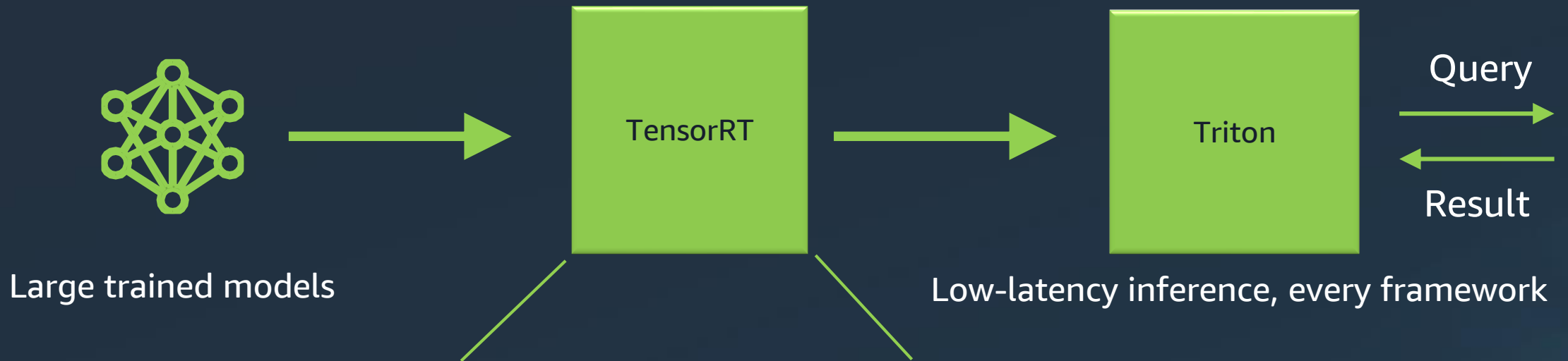
AI inference workflow

Two-part process implemented by multiple personas

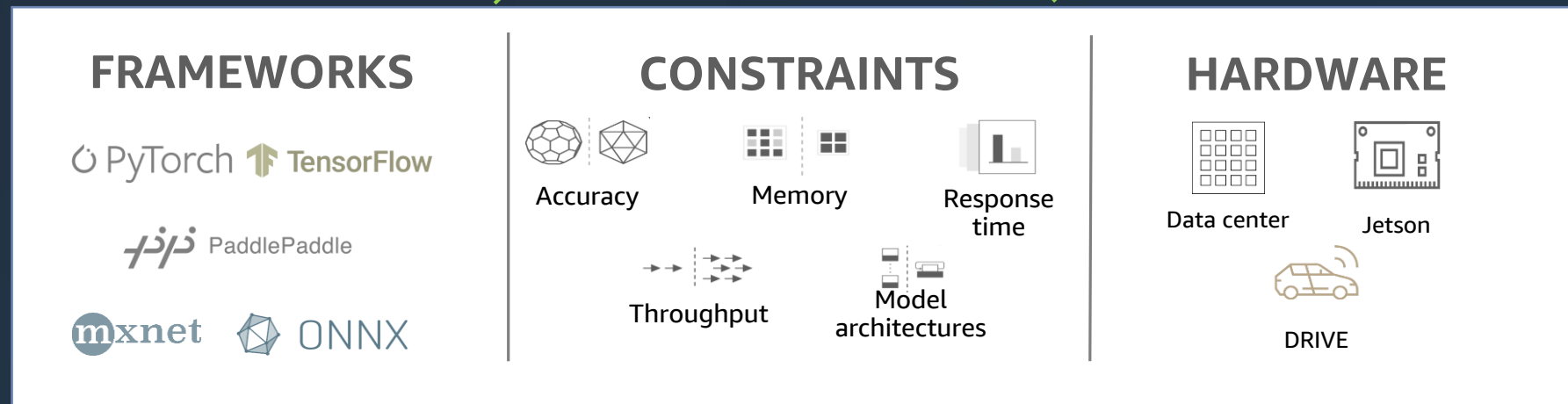


Inference is complex

REAL TIME | COMPETING CONSTRAINTS | RAPID UPDATES




Low-latency inference, every framework

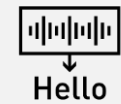


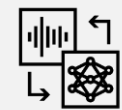
A world-leading inference performance


TensorRT accelerates every workload


BEST-IN-CLASS RESPONSE TIME AND THROUGHPUT vs. CPUs

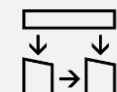
 **36x**
Computer vision
< 7 ms

 **583x**
Speech recognition
< 100 ms

 **21x**
NLP
< 50 ms

 **10x**
Reinforcement
learning

 **178x**
Text-to-speech
< 100 ms

 **12x**
Recommenders
< 1 sec

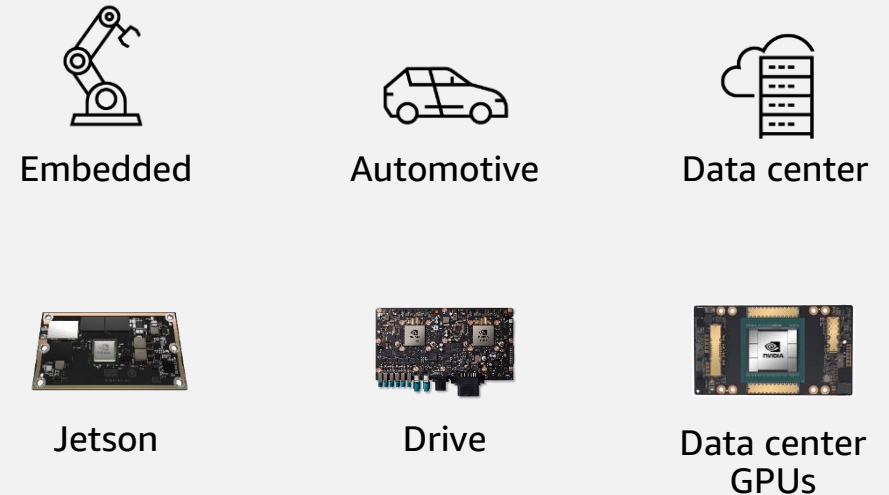
NVIDIA TensorRT

SDK for High-Performance Deep Learning Inference

Optimize and deploy neural networks in production

Maximize throughput for latency-critical applications with compiler and runtime; optimize every network, including CNNs, RNNs, and transformers

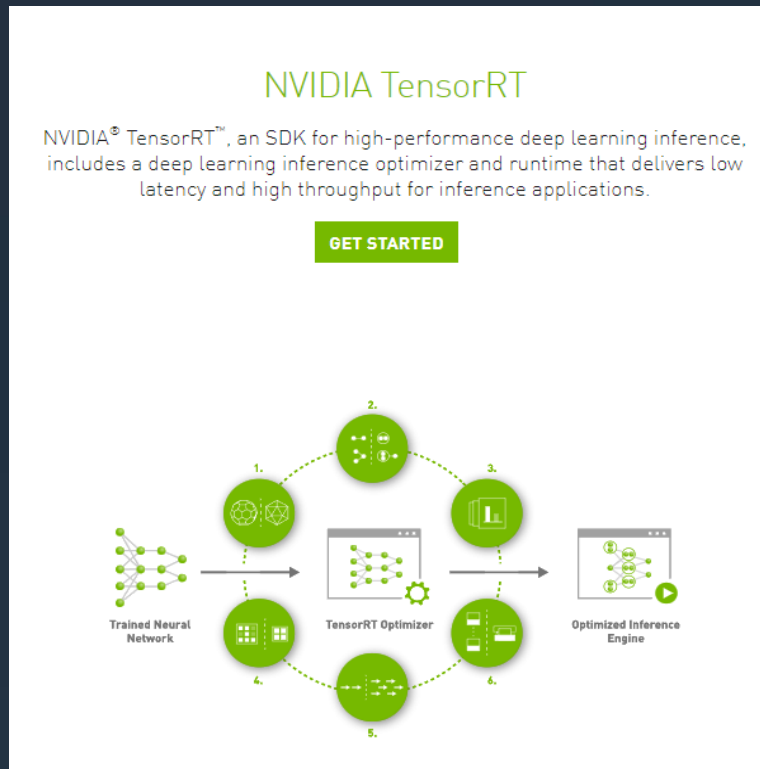
1. Reduced mixed precision: FP32, TF32, FP16, and INT8
2. Layer and tensor fusion: Optimizes use of GPU memory bandwidth
3. Kernel auto-tuning: Select best algorithm on target GPU
4. Dynamic tensor memory: Deploy memory-efficient applications
5. Multi-stream execution: Scalable design to process multiple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



Download TensorRT today

Tensorflow with Tensorrt

TensorRT



Torch-TensorRT

NVIDIA NGC | CATALOG

Catalog > Containers > PyTorch

PyTorch

Accelerated with NVIDIA

Description

PyTorch is a GPU accelerated tensor computational framework. Functionality can be extended with common Python libraries such as NumPy and SciPy. Automatic differentiation is done with a tape-based system at the functional and neural network layer levels.

Publisher
Facebook

Latest Tag
22.02-py3

Modified
March 10, 2022

Compressed Size
6.51 GB

Overview Tags Layers Security Scans

PyTorch

PyTorch is an optimized tensor library for deep learning. Automatic differentiation is done with a tape-based system at the functional and neural network layer levels. This functionality provides speed and accuracy as a deep learning framework and provides a place to start developing common applications, such as natural language processing (NLP), recommenders, and computer vision.

The PyTorch NGC Container is optimized for GPU acceleration and contains a validated set of libraries that enable and optimize the container also contains software for accelerating ETL (cuDNN, NCCL), and Inference (TensorRT) workload.

Prerequisites

Using the PyTorch NGC Container requires the host system to have the following installed:

- Docker Engine
- NVIDIA GPU Drivers
- NVIDIA Container Toolkit

TensorFlow-TensorRT

NVIDIA NGC | CATALOG

Catalog > Containers > TensorFlow

TensorFlow

Accelerated with NVIDIA

Description

TensorFlow is an open source platform for machine learning. It provides comprehensive tools and libraries in a flexible architecture allowing easy deployment across a variety of platforms and devices.

Publisher
Google Brain Team

Latest Tag
22.02-tf2-py3

Modified

Overview Tags Layers

TensorFlow

TensorFlow is an open source platform for machine learning. It provides comprehensive tools and libraries in a flexible architecture allowing easy deployment across a variety of platforms and devices. The TensorFlow NGC Container contains a validated set of libraries that enable and optimize the container also contains software for accelerating ETL (DALI, RAPIDS), Training (cuDNN, NCCL), and Inference (TensorRT) workload.

Prerequisites

Using the TensorFlow NGC Container requires the host system to have the following installed:

- Docker Engine
- NVIDIA GPU Drivers
- NVIDIA Container Toolkit

TensorRT 8.4 GA is available for free to the members of the NVIDIA Developer Program: developer.nvidia.com/tensorrt

NVIDIA Triton Inference Server

Open-source software for fast, scalable, simplified inference serving

Any framework



Supports multiple framework backends natively; e.g., TensorFlow, PyTorch, TensorRT, XGBoost, ONNX, Python & more

Any query type



Optimized for real time, batch, streaming, ensemble inferencing

Any platform



X86 CPU | Arm CPU | NVIDIA GPUs | MIG
Linux | Windows | virtualization
Public cloud, data center, and edge/embedded (Jetson)

DevOps & MLOps



Integration with Kubernetes, KServe, Prometheus & Grafana
Available across all major cloud AI platforms

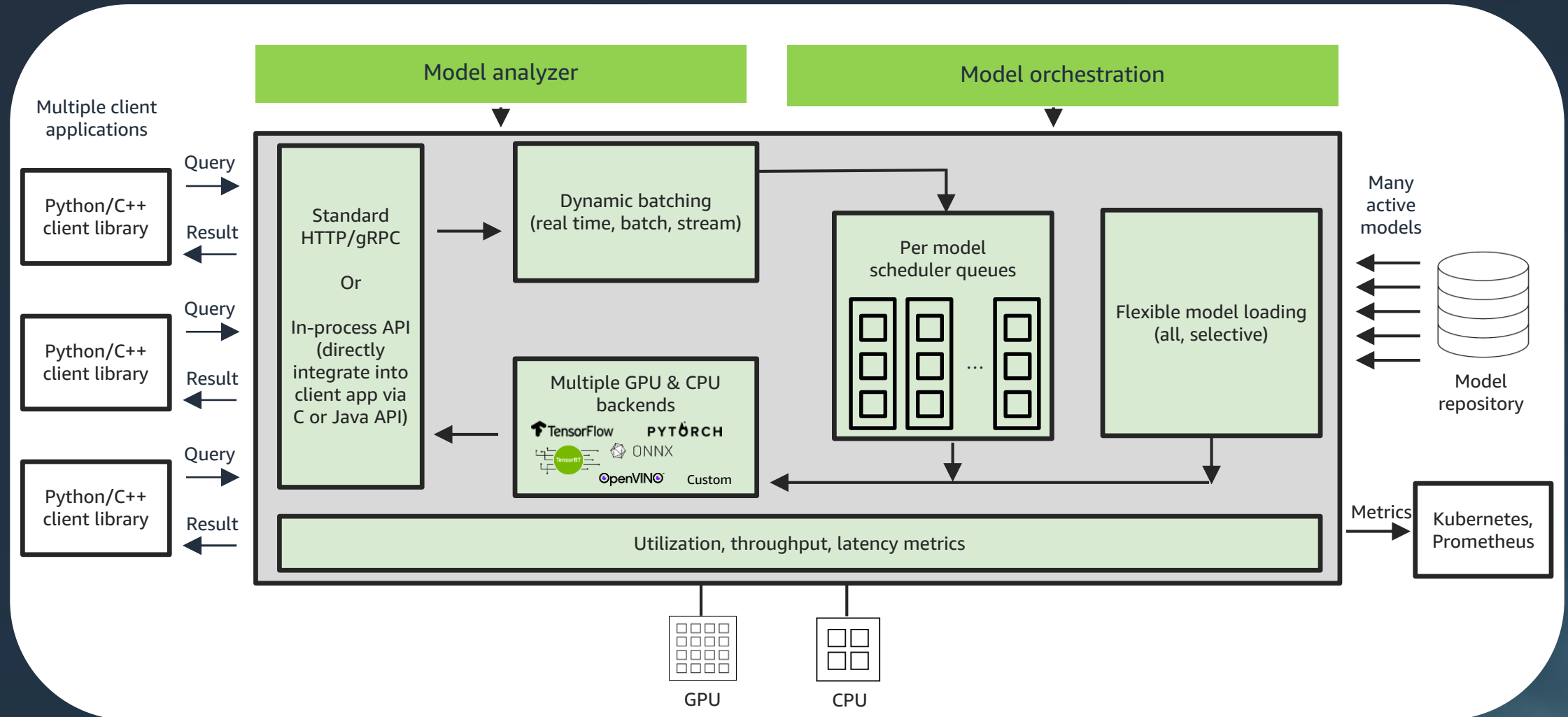
Performance & utilization



Model Analyzer for optimal configuration
Optimized for high GPU/CPU utilization, high throughput & low latency

Tritons architecture

Delivering high performance across frameworks



Concurrent model execution

INCREASE THROUGHPUT AND UTILIZATION

Dynamic batching scheduler

GROUP REQUESTS TO FORM LARGER BATCHES, INCREASE GPU UTILIZATION

Optimal model configuration

USING THE MODEL ANALYZER CAPABILITY

Large language model inference

USING TRITON'S FASTERTRANSFORMER BACKEND

Model pipelines with business logic scripting

CONTROL FLOW AND LOOPS IN MODEL ENSEMBLES

Decoupled models

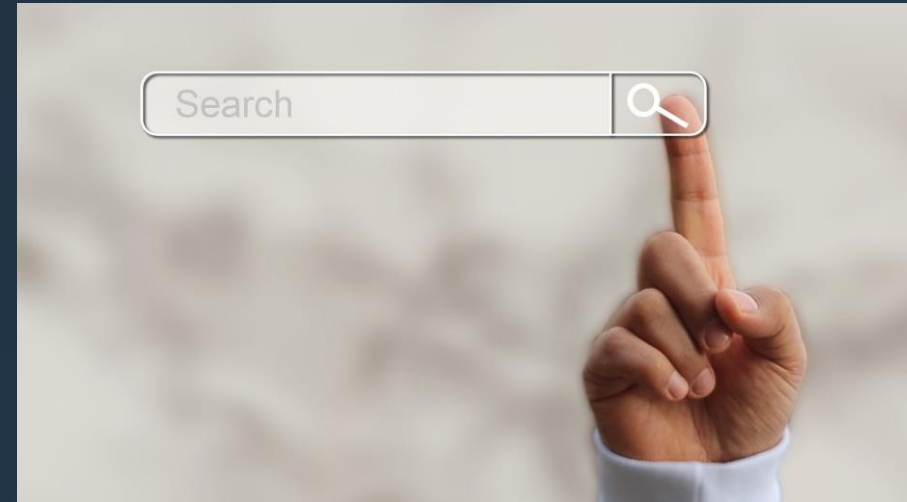
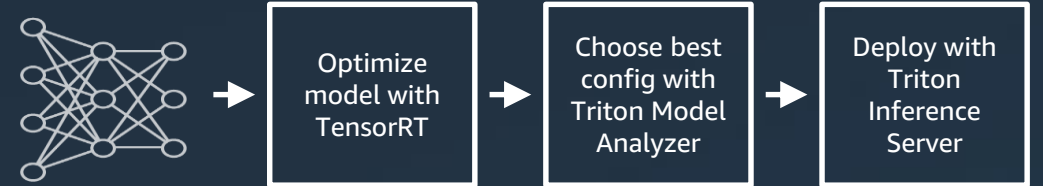
ALLOWS 0, 1, OR 1+ RESPONSES PER REQUEST



Real-time spell check for product search

Amazon Search

- One of the most visited ecommerce websites
- Deep learning (DL) AI model for automatic spell correction to search effortlessly
- Triton + TensorRT meets sub-50 ms latency target and delivers 5x throughput for DL model on GPUs on AWS
- Triton Model Analyzer reduced time to find optimal configuration from weeks to hours



<https://aws.amazon.com/blogs/machine-learning/how-amazon-search-achieves-low-latency-high-throughput-t5-inference-with-nvidia-triton-on-aws/>

Learn more and download

For more information

<https://developer.nvidia.com/nvidia-triton-inference-server>

Get the ready-to-deploy container with monthly updates from the NGC catalog

<https://catalog.ngc.nvidia.com/orgs/nvidia/containers/tritonserver>

Open-source GitHub repository

<https://github.com/NVIDIA/triton-inference-server>

Latest release information

<https://github.com/triton-inference-server/server/releases>

Quick start guide

https://github.com/triton-inference-server/server/blob/main/docs/getting_started/quickstart.md

Triton Inference Server on Amazon SageMaker



A Triton Inference Server container developed with NVIDIA – includes NVIDIA Triton Inference Server along with useful environment variables to tune performance (e.g., set thread count) on SageMaker



Use with SageMaker Python SDK to deploy your models on scalable, cost-effective SageMaker endpoints without worrying about Docker



Code examples to find readily usable code samples using Triton Inference Server with popular machine learning frameworks on Amazon SageMaker

Amazon SageMaker & Triton technical resources

Triton on Amazon SageMaker

[Achieve hyperscale performance for model serving using NVIDIA Triton Inference Server on Amazon SageMaker](#)

[Amazon announces new NVIDIA Triton Inference Server on Amazon SageMaker](#)

[Deploy fast and scalable AI with NVIDIA Triton Inference Server in Amazon SageMaker](#)

[Use Triton Inference Server with Amazon SageMaker](#)

[How Amazon Search achieves low-latency, high-throughput T5 inference with NVIDIA Triton on AWS](#)

[Getting the most out of NVIDIA T4 on AWS G4 Instances](#)

[Deploying the Nvidia Triton Inference Server on Amazon ECS](#)

AWS AI/ML Heroes collaboration

[NVIDIA Triton spam detection engine of C-suite labs](#)

[Blurry faces: Training, optimizing and deploying a segmentation model on Amazon SageMaker with NVIDIA TensorRT and NVIDIA Triton](#)

Sign up for NVIDIA and AWS free ML Course

In this course, you will gain hands-on experience on building, training, and deploying scalable machine learning models with Amazon SageMaker and Amazon EC2 instances powered by NVIDIA GPUs



Hands-on Machine Learning with AWS/NVIDIA | Coursera
<https://www.coursera.org/learn/machine-learning-aws-nvidia>



Free e-book: Dive into deep learning
<https://d2l.ai>

Recap and next steps

Recap and key takeaways

What did we learn today?

NVIDIA GPUs power the most compute-intensive workloads from computer vision to speech to language and many more

NVIDIA TAO is a toolkit for training CV and speech models efficiently

NVIDIA NeMo Megatron is a open-source toolkit for large language model training and deployment

NVIDIA TensorRT is an SDK for optimizing deep learning models

NVIDIA Triton is an inference server for deploying your models

Join the NVIDIA Inception program for startups

Accelerate your startup's growth and build your solutions faster with engineering guidance, free technical training, preferred pricing on NVIDIA products, opportunities for customer introductions and co-marketing, and exposure to the VC community



APPLY TO INCEPTION TODAY

<https://www.nvidia.com/en-us/startups>



GET THE LATEST NEWS, UPDATES, AND MORE

<https://www.nvidia.com/en-us/preferences/email-signup/>

Thank you!

Michael Lang

MiLang@NVIDIA.com

