AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

PRT219

SPONSORED BY NVIDIA

Deep learning on AWS with NVIDIA: From training to deployment

Jiahong Liu
Solutions Architect
NVIDIA

Edwin Weill
Senior Solutions Architect
NVIDIA

Speakers

Jiahong Liu

Edwin Weill

Agenda

NVIDIA and AWS relationship

NVIDIA AI on AWS

Training (at scale)

Deployment and inference

Call to action

Conclusion

NVIDIA AI

END-TO-END OPEN PLATFORM FOR PRODUCTION AI

Application workflows

CLARA	RIVA	TOKKIO	MERLIN	MODULUS	MAXINE	METROPOLIS	CUOPT	NEMO	ISAAC	DRIVE	MORPHEUS
Medical imaging	Speech AI	Customer service	Recommenders	Physics ML	Video	Video analytics	Logistics	Conversational AI	Robotics	Autonomous vehicles	Cybersecurity

NVIDIA AI Enterprise

- AI and data science development and deployment tools
- Cloud-native management and orchestration
- Infrastructure optimization

Accelerated infrastructure

Cloud | Data center | Edge | Embedded

NVIDIA LaunchPad

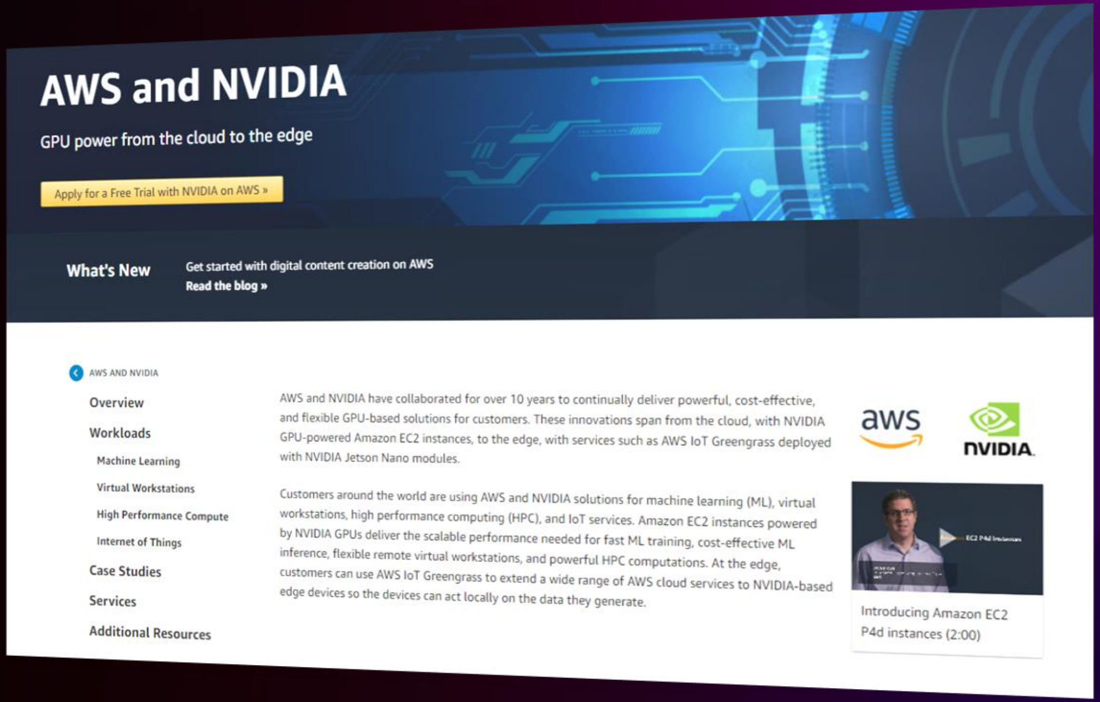
Hands-on labs

NVIDIA and AWS relationship

GPU power from the cloud to the edge

Machine learning	Virtual workstations	High-performance compute	Internet of things
ML training and cost-effective inference	Work from anywhere	Solve large computational problems	Extend to edge devices that act locally

Powerful | Cost-Effective | Flexible



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 in the AWS Marketplace

aws

NVIDIA

NVIDIA AI on AWS

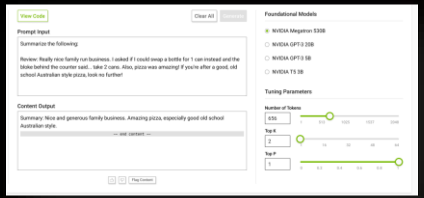
NGC

PORTAL TO AI SERVICES, SOFTWARE, SUPPORT

NGC catalog

Cloud services

End-to-end AI development

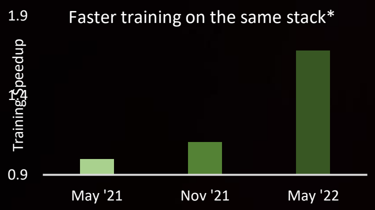


AI services for NLP, biology, speech

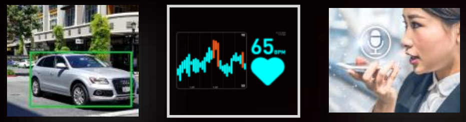
AI workflow management & support

Performance optimized

Tested across
GPU-accelerated platforms



Monthly SW container updates



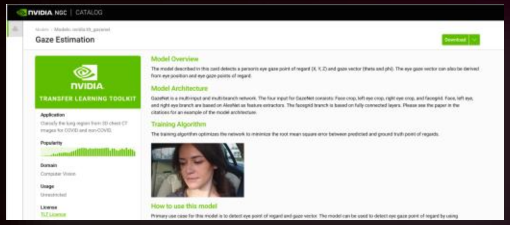
SOTA models

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

jupyter

One-click deploy from NGC

Multiple cloud
providers

Develop once; deploy
anywhere with NVIDIA VMI

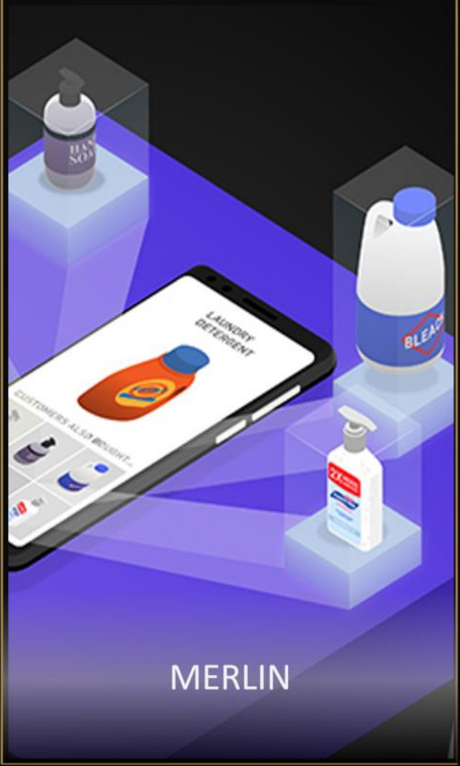
ngc.nvidia.com

Accelerating the next wave of AI

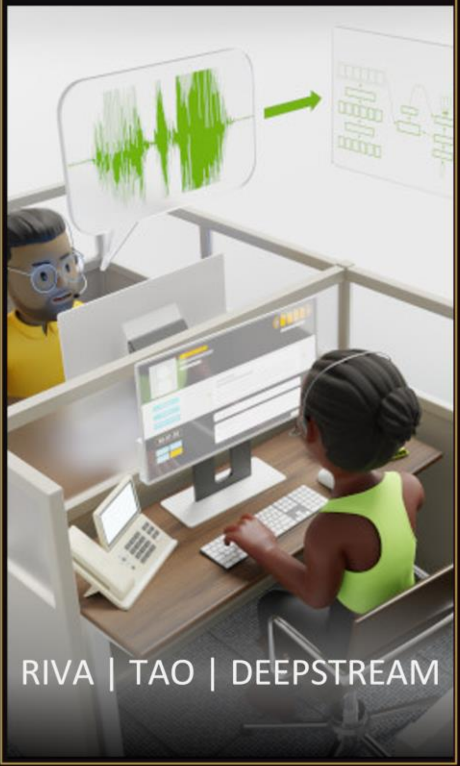
AI PLATFORM UPDATES

DATA SCIENCE
Analytics, ML, Visualization

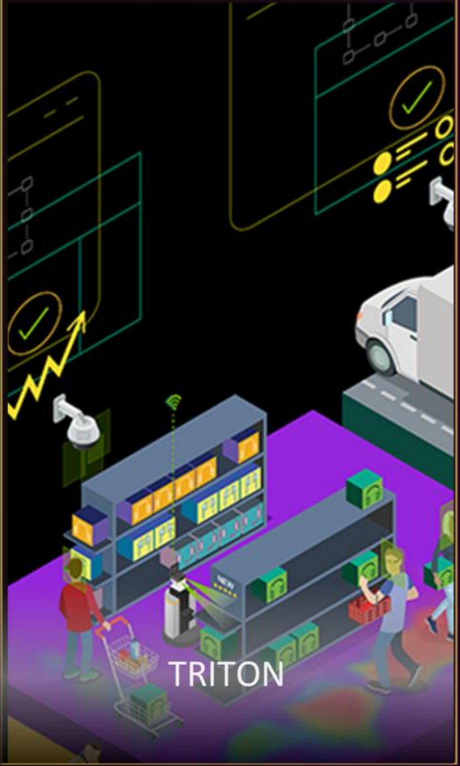
RECOMMENDER
Personalization Engine, Simplified



SPEECH & VISION
Conv AI, Video Analytics



INFERENCE
Fast, Scalable Predictions

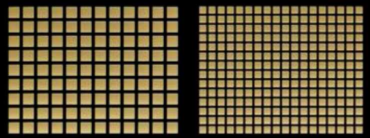


NLP
Large Language Model

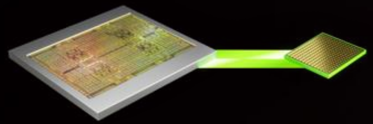
NVIDIA.

NVIDIA A100

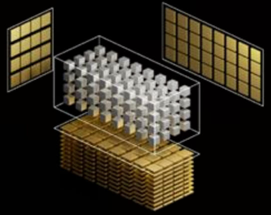
SUPERCHARGING HIGH PERFORMING AI SUPERCOMPUTING GPU



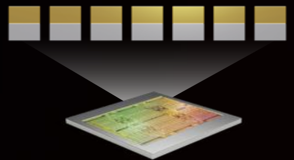
80 GB HBM2e
For largest datasets
and models



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

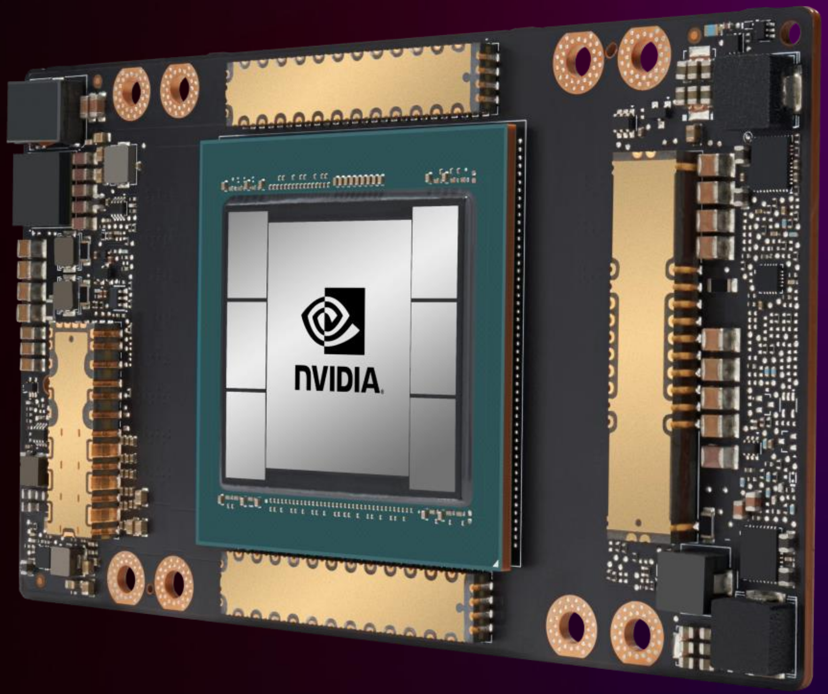


3rd-gen Tensor core



Multi-instance GPU

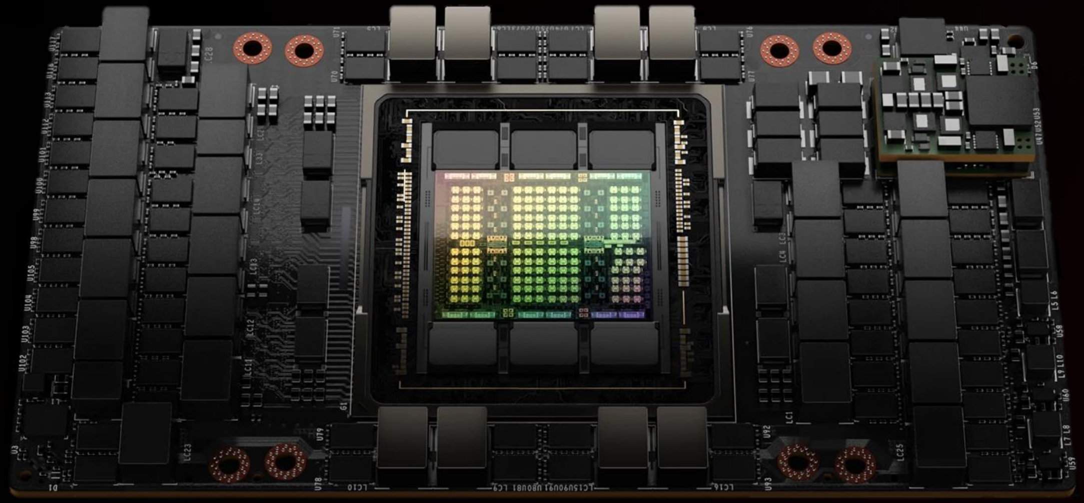
3rd-gen NVLink



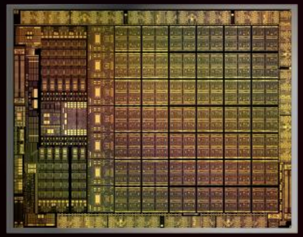
Powering Amazon EC2 P4d/P4de instances

NVIDIA H100 – Coming soon to AWS

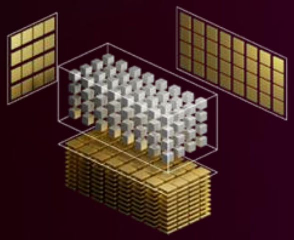
THE NEW ENGINE OF THE WORLD'S AI INFRASTRUCTURE



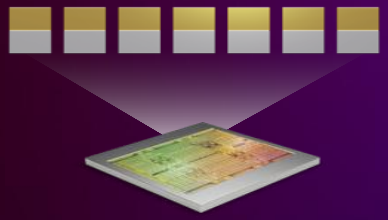
Powering the next generation of GPU systems on AWS



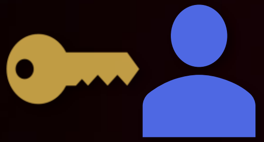
Advanced chip



Transformer engine

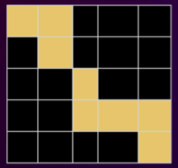


2nd-gen MIG



Confidential computing

4th-gen NVLink

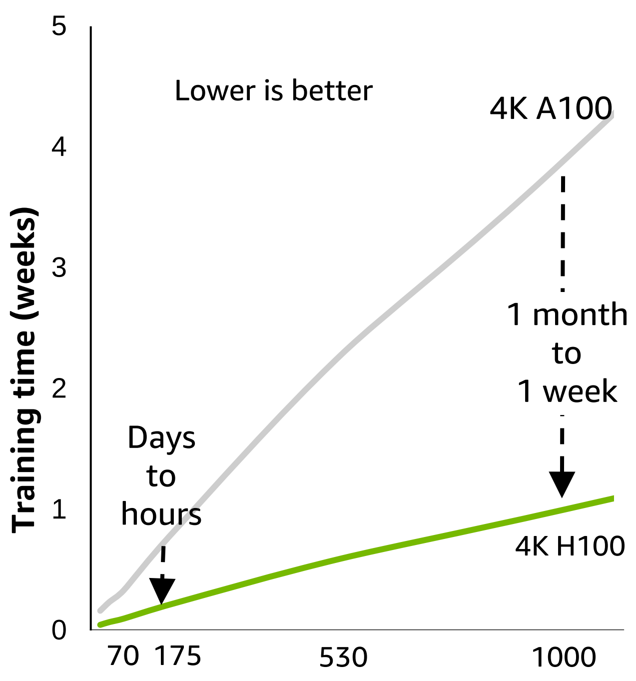


DPX instructions

NVIDIA H100 supercharges LLMs

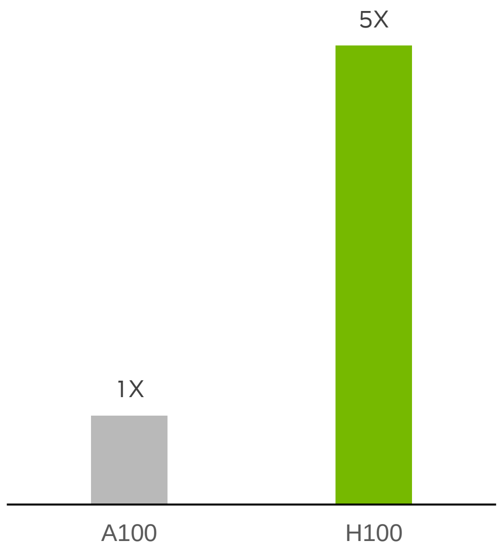
HOPPER ARCHITECTURE ADDRESSES LLM NEEDS AT SCALE

Supercharged LLM training



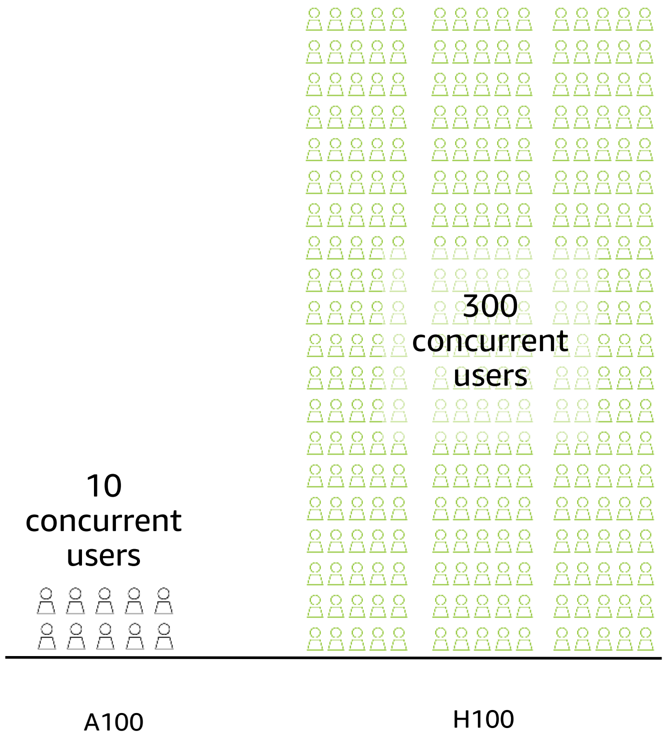
Time-to-train by LLM size (billion parameters)

High-performance prompt learning



530B P-tuning time-to-train

30x real-time inference throughput



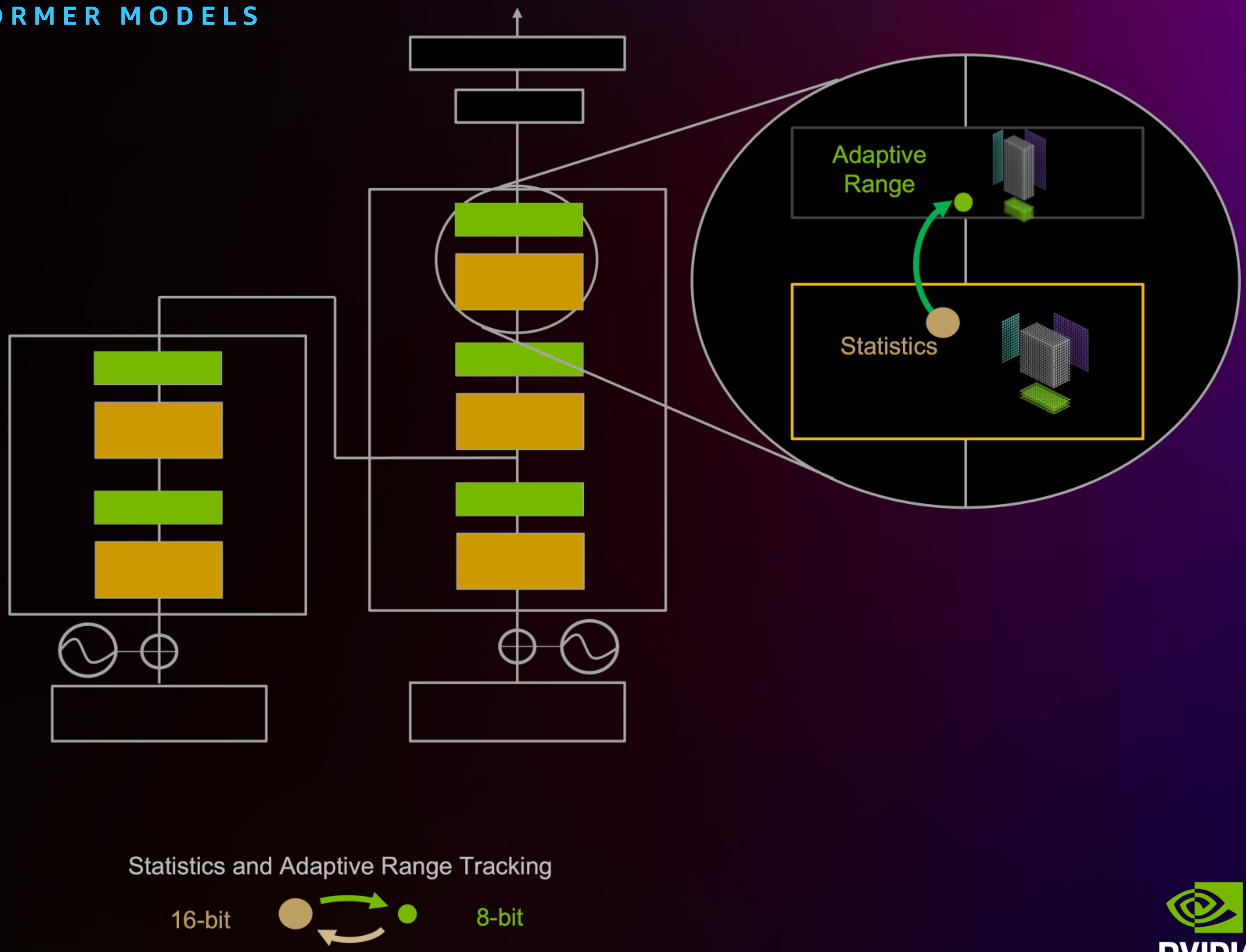
530B inference on 10 DGX systems

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

Transformer engine

TENSOR CORE OPTIMIZED FOR TRANSFORMER MODELS

- 6x faster training and inference of transformer models
- NVIDIA tuned adaptive range optimization across 16-bit and 8-bit math
- Configurable macro blocks deliver performance without accuracy loss



Why is this important for your model training?

Amazon EC2 instance 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 *Includes Quadro Virtual Workstation*	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

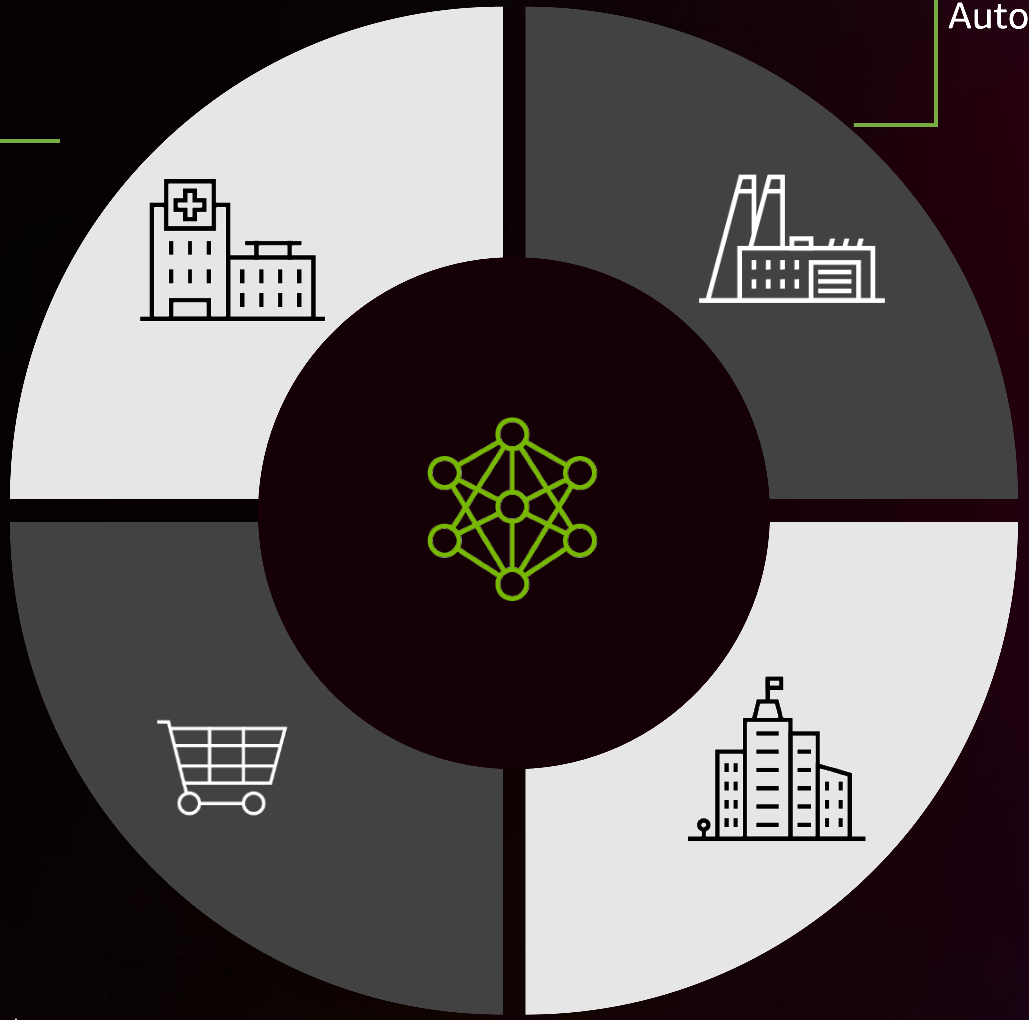
Proliferation of use cases

Healthcare

Patient monitoring
Smart hospitals
Robot-assisted surgery

Industrial manufacturing

Automated optical inspection
Worker safety
Process automation



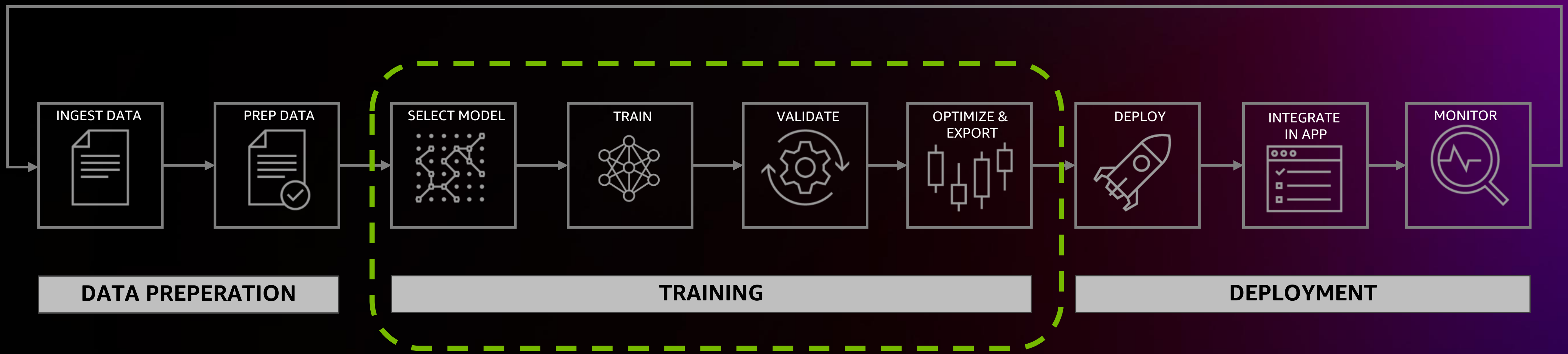
Retail

Detecting people movement
Analyzing action
Warehouse logistics

Smart infrastructure

Pedestrian safety
Traffic management
Waste management

Creating an AI application is hard and complex



DATA PREPERATION	TRAINING	DEPLOYMENT
Labeling, annotating, and augmenting	Model training, pruning, and optimizing	Deploying and monitoring

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

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

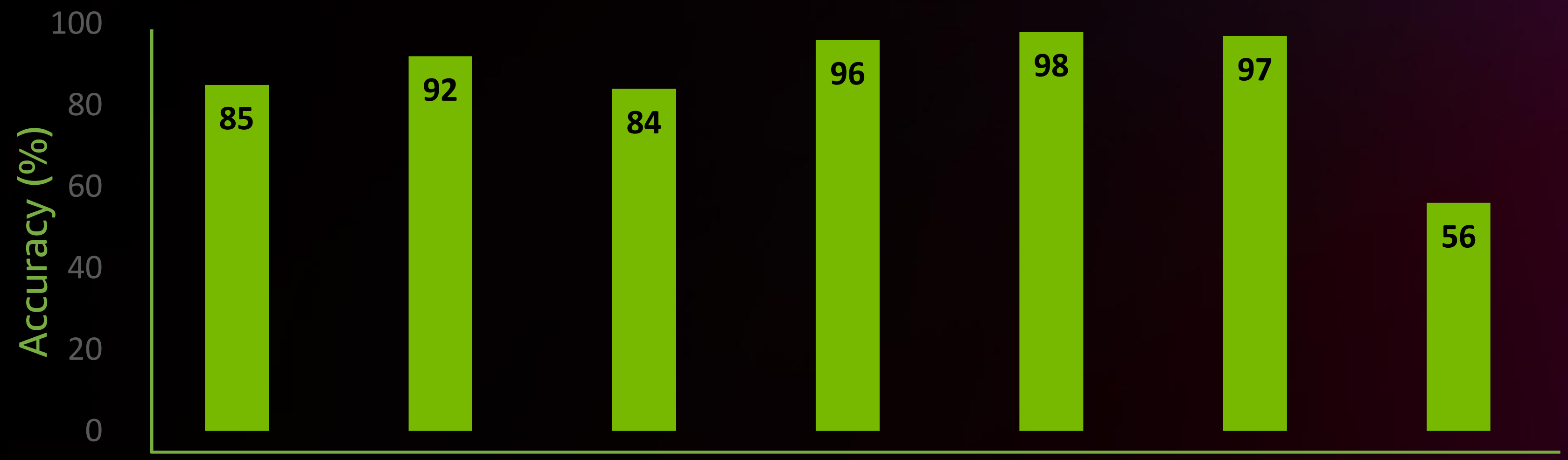
OPTIMIZE FOR DEPLOYMENT

Optimize for inference and integrate with Riva or DeepStream

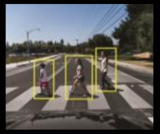
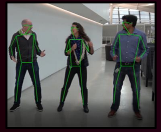
SUPPORTED BY EXPERTS

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

High-performance pretrained vision AI models



Models	Accuracy
Facial landmark	6.1-pixel error
Gaze estimation	6.5 RMSE

	PeopleNet	PeopleSemSegNet	TrafficCamNet	FaceDetectIR	LPD	LPR	2D Pose Estimation	Facial Landmark	Gaze Estimation
Nano	11	1.4	18	104	66	94	5	125	98
Xavier NX	296	17	340	2000	1158	564	48	747	923
AGX Xavier	462	28	656	3915	1880	1045	84	1451	1627
A30	4163	330	4991	26635	12207	15960	1515	10078	15172
A100	6001	519	9520	50541	21931	26600	2686	23117	26534

15+ Pretrained models – download for free from NGC

Pretrained conversational AI models

Automatic Speech Recognition

- Jasper
- QuartzNet
- CitriNet
- N-Gram

Natural Language Processing

- BERT Punctuation
- BERT NER
- BERT Text Classification
- BERT Intent & Slot
- Domain-Specific NER
- BERT & Megatron QA

Text to Speech

- FastPitch
- HiFi-GAN

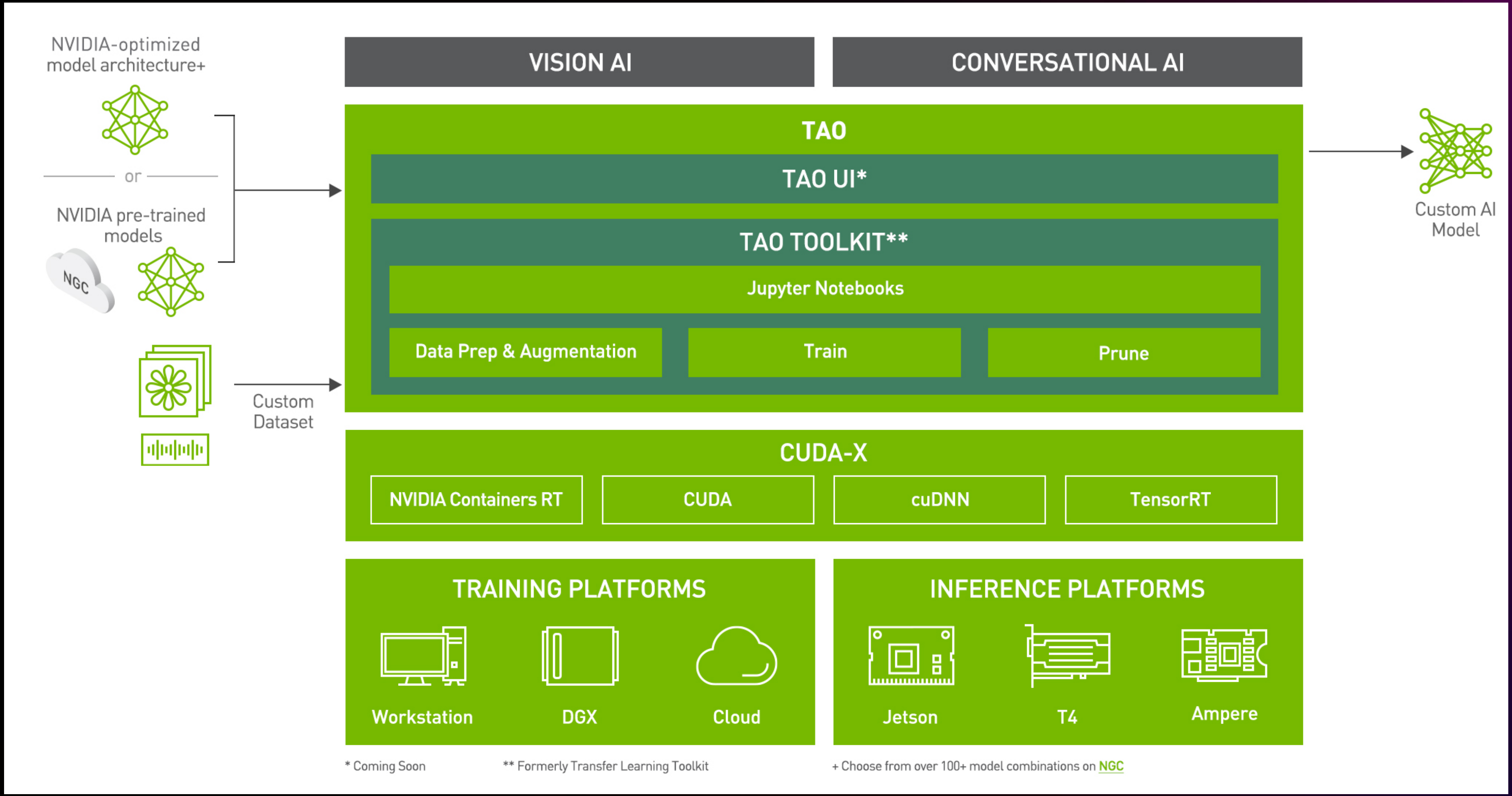
Support for models that are used in the conversational AI pipeline

Adapt with your dataset using NVIDIA TAO Toolkit

Deploy with turnkey inference applications in NVIDIA Riva

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

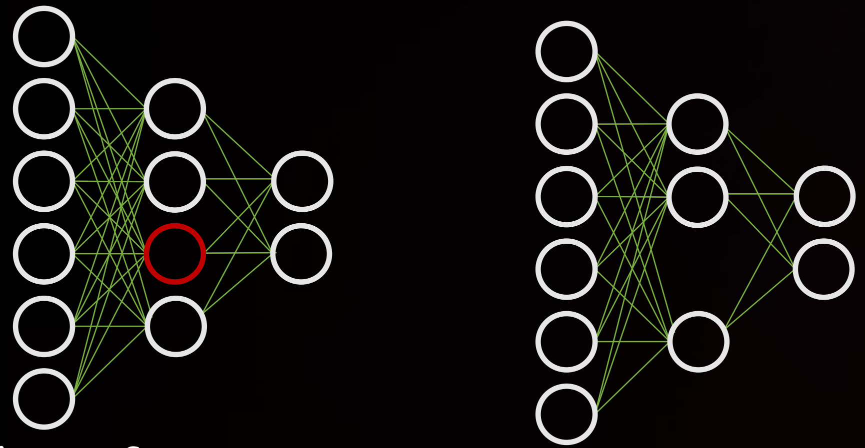
The NVIDIA TAO stack



TAO – Features

MODEL PRUNING AND QUANTIZATION

Model pruning



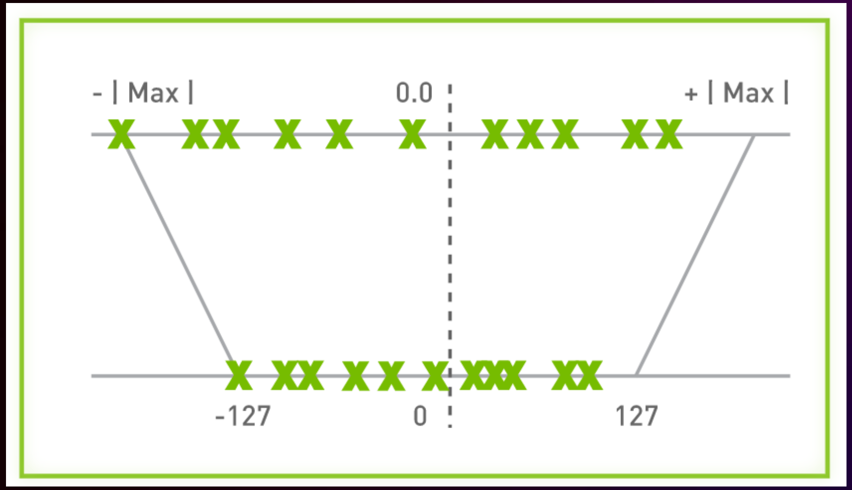
6 inputs, 6 neurons,
32 connections

6 inputs, 5 neurons,
24 connections

TrafficCamNet

Quantization

- Post Training Quantization (PTQ) for quantization after training is done
- Quantization Aware Training (QAT) for quantization error from weights and tensors during training

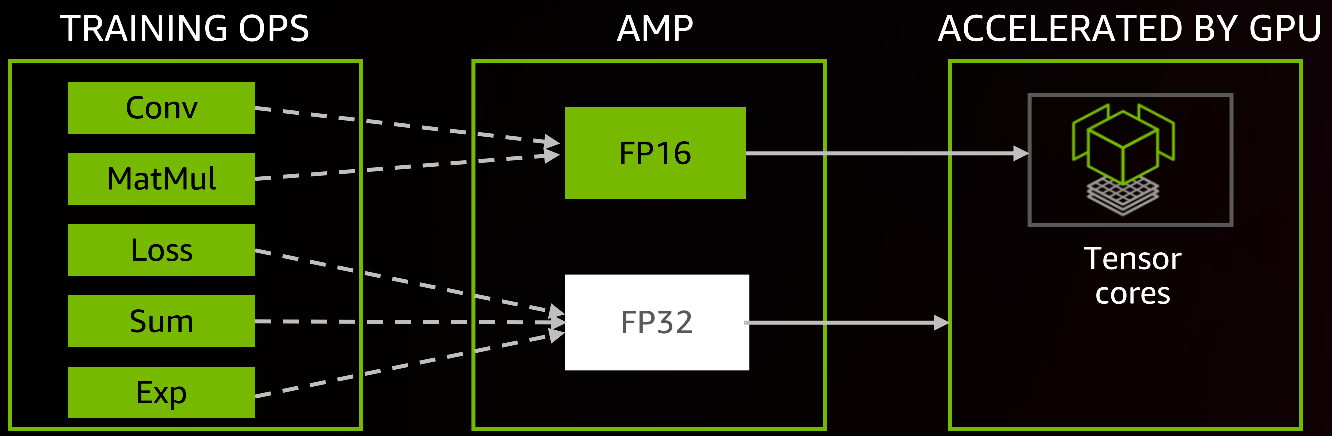


Transformation of floating-point weights to integer

TAO – Features

AUTOMATED MIXED PRECISION AND DATA AUGMENTATION

Automated mixed precision (AMP)

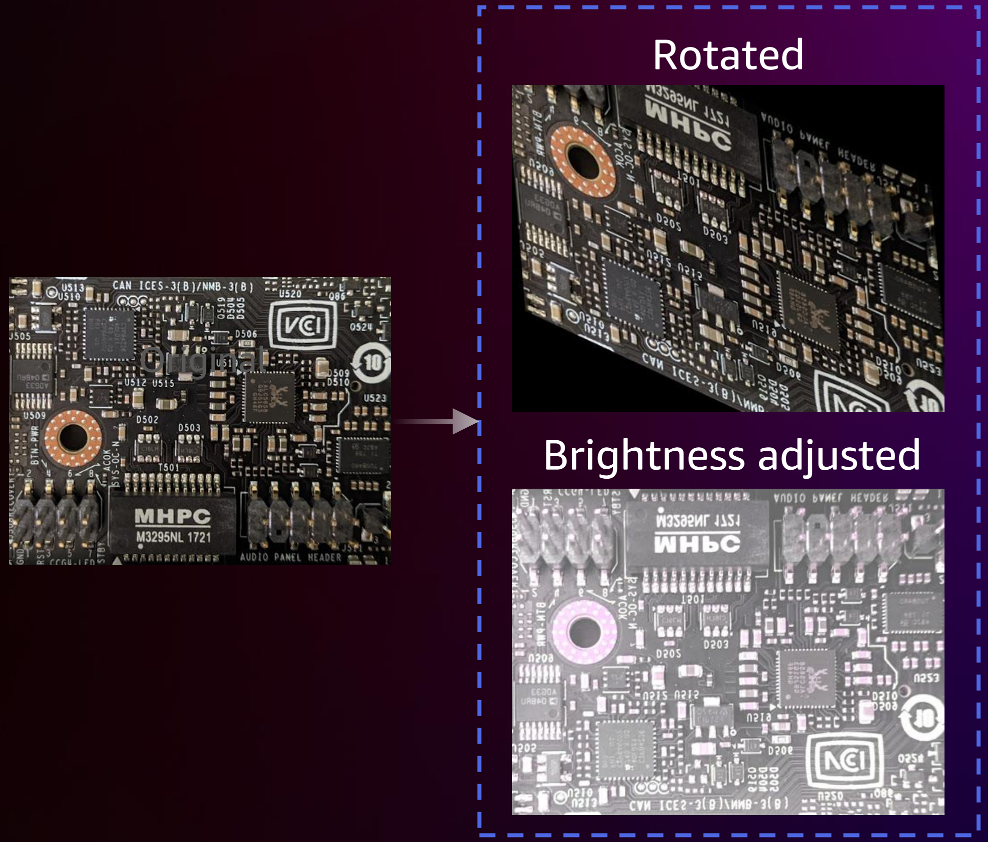


Accelerate AI training

Reduce memory bandwidth

Train larger models or larger batches

Data augmentation



As simple as it can get

Configure model parameter

```
model_config:
  model_type: rgb
  input_type: "2d"
  # input_type: "3d"
  backbone: resnet18
  rgb_seq_length: 32
  rgb_pretrained_model_path: resnet18_2d_rgb_hmdb5_32.tlt
  # rgb_pretrained_model_path: resnet18_3d_rgb_hmdb5_32.tlt
  rgb_pretrained_num_classes: 5
  sample_strategy: consecutive
  sample_rate: 1
```

Configure training parameter

```
train_config:
   optim:
      lr: 0.01
      momentum: 0.9
      weight_decay: 0.0001
      lr_scheduler: MultiStep
      lr_steps: [5, 15, 25]
      lr_decay: 0.1
   epochs: 50
   checkpoint_interval: 1
```

Configure dataset

```
dataset_config:
   train_dataset_dir: /data/train
   val_dataset_dir: /data/test
   label_map:
      pushup: 0
      pullup: 1
      situp: 2
   output_shape:
   - 224
   - 224
   batch_size: 32
   workers: 8
   clips_per_video: 15
```

Simplifying **>10K** lines of code into a few options

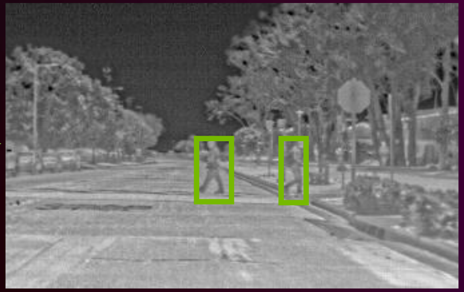
Art of the possible

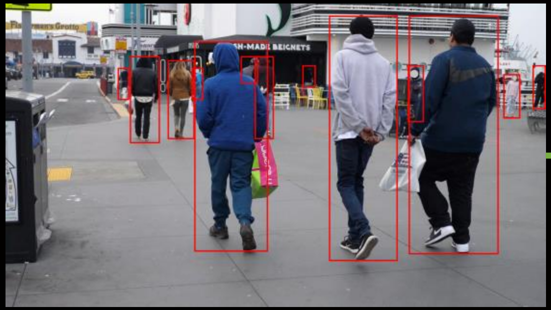
ADAPTING TO DIFFERENT CAMERA TYPES

Thermal IR dataset

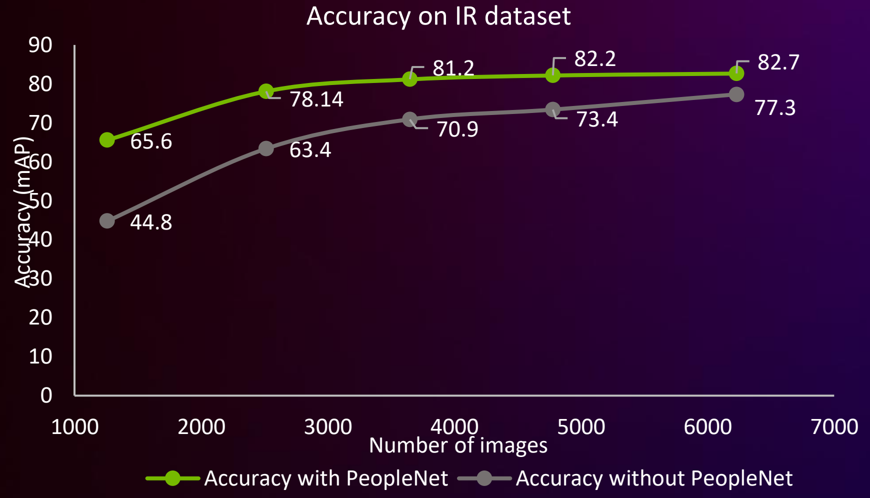
TAO TOOLKIT

Train Adapt Optimize





PeopleNet model



TAO Toolkit whitepaper

GitHub project repository

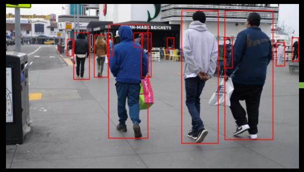
Higher accuracy with 50% less data

Art of the possible

ADDING NEW CLASSES

Goal: Add “Helmet” class to existing people detect model

Labeled helmet dataset



PeopleNet model

TAO TOOLKIT

Inference

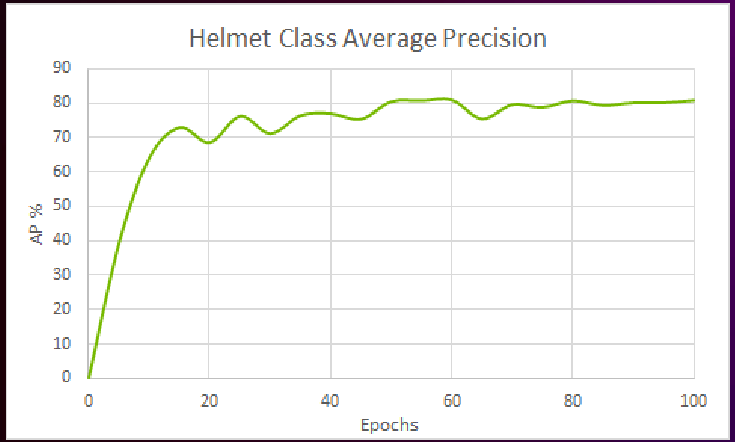
People and face annotation

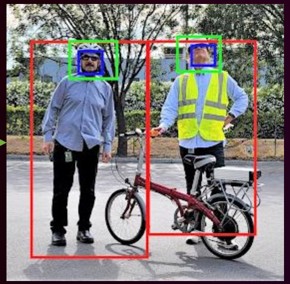
Complete annotated dataset

TAO TOOLKIT

Train | Adapt | Optimize

80% AP over 100 epochs





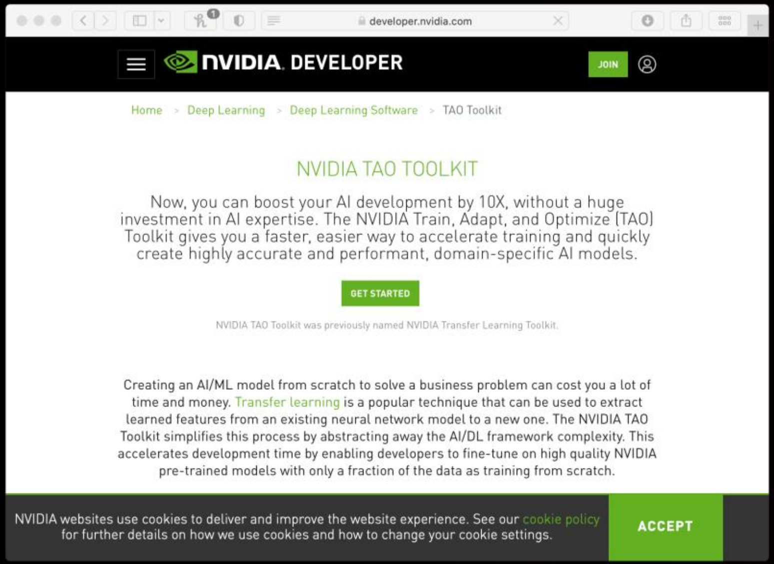
Model with helmet, people, and face detection

[TAO Toolkit whitepaper]

[GitHub project repository]

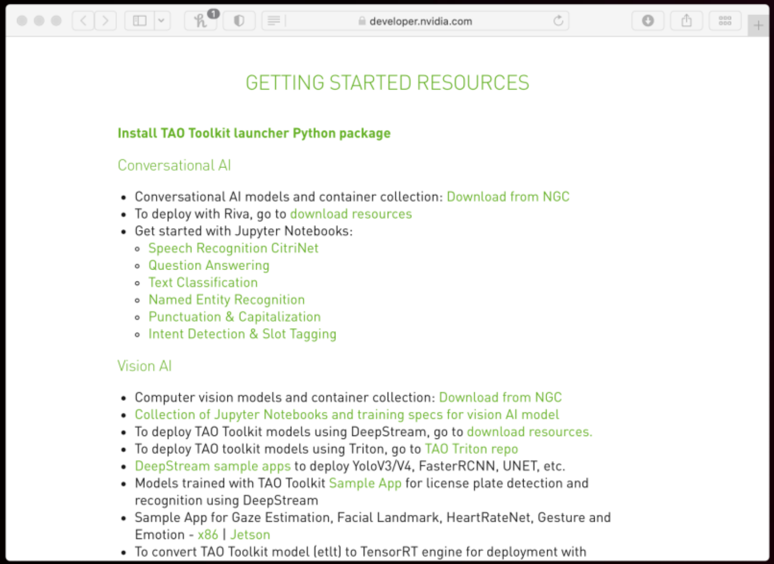
Resources

GETTING STARTED WITH THE TAO TOOLKIT



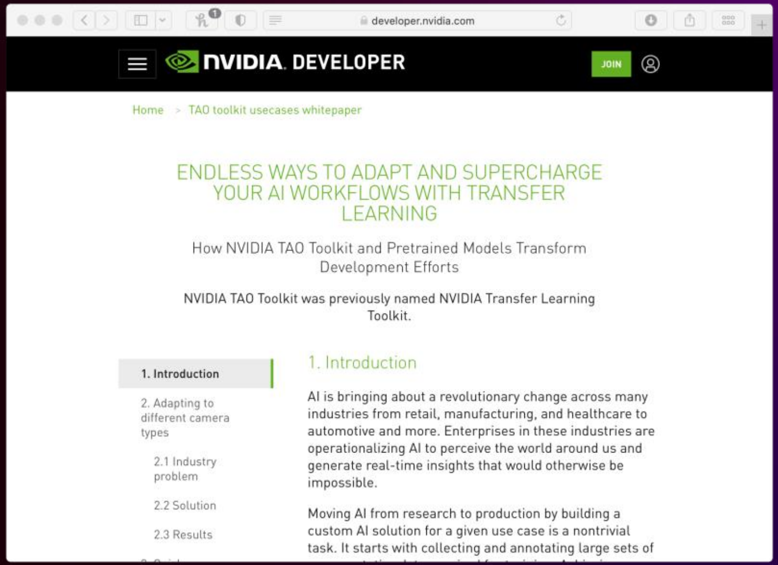
[TAO Toolkit product page]

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]

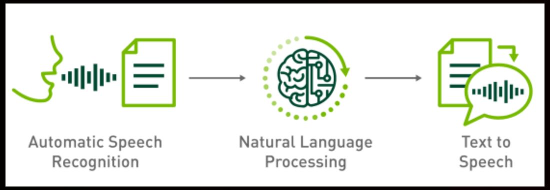
Includes examples on data augmentation, adding new classes

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

Summarization

GPT-3

Marketing copy

TEXT GENERATION

Translating Wikipedia

NLLB-200

Real-time metaverse translation

TRANSLATION

Brand creation

DALL-E-2

Gaming characters

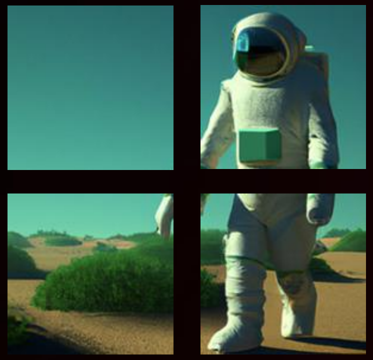
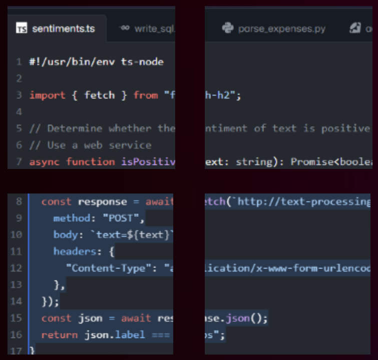


IMAGE GENERATION

Dynamic code commenting

CODEX

Function generation

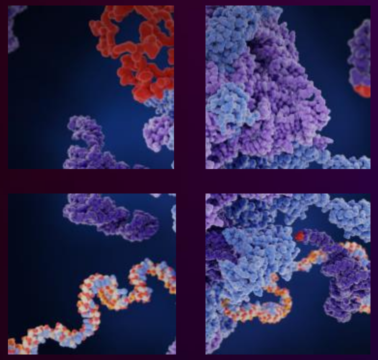


CODING

Molecular representations

MegaMolBART

Drug discovery

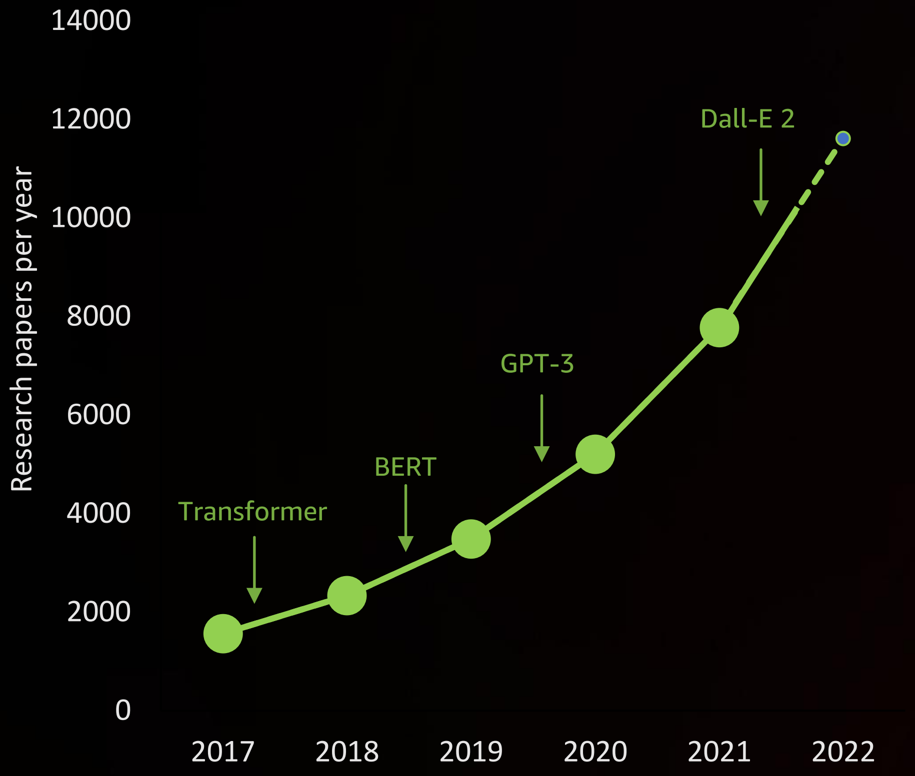


LIFE SCIENCE

Large language models codifying intelligence

LLM RESEARCH ACCELERATING INNOVATION AND ABILITIES

Transformer and LLM research papers per year

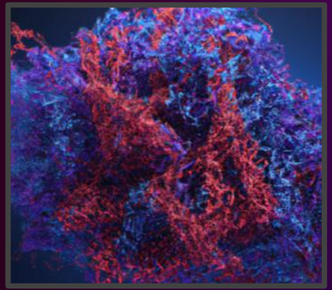


Explosion of use cases

IMAGE GENERATION
Brand creation
Gaming characters

RECOMMENDATIONS
Ecommerce
Personalized content

LIFE SCIENCE RESEARCH
Molecular representations
Drug discovery



TRANSLATION
Translating Wikipedia
Real-time metaverse translation

TEXT GENERATION
Summarization
Marketing copy

CODING
Dynamic code comments
Function generation

*Research paper published on Arxiv.org related to transformers and LLMs in computer science subject area, with projected count for rest of 2022

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

NVIDIA

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

Verified Convergence Recipes, Evaluation Harness and Sample Chatbot Application			
Distributed Data Pre-processing	Hyper Parameter Tuning	Distributed Training	Accelerated Inference
NVIDIA Base Command Platform			
CSPs, DGX SuperPODs, DGX Foundry			

- 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)

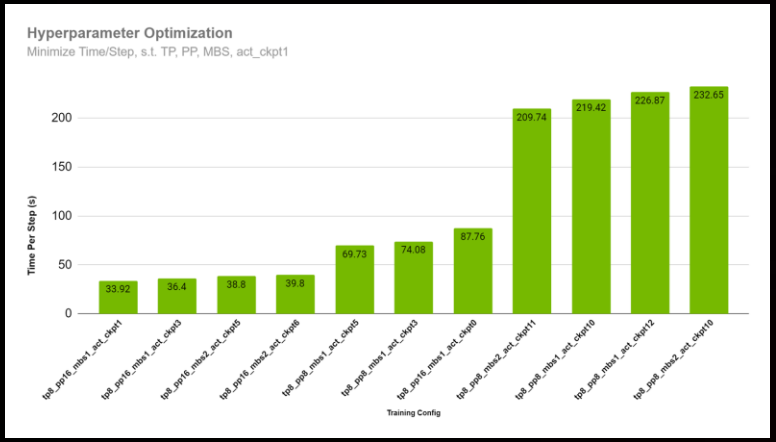
Now in open beta

Find out more:
NVIDIA NeMo Megatron

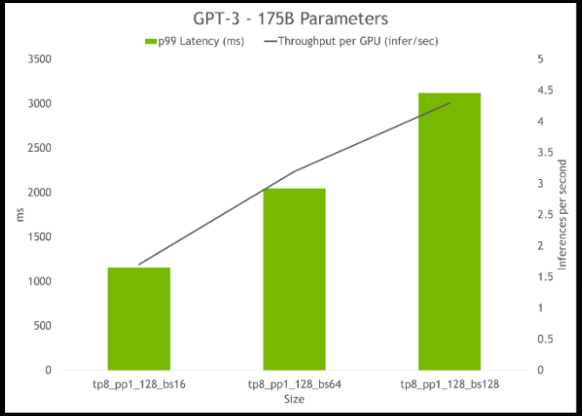
NeMo Megatron – Features

HYPERPARAMETER TOOL AND DATA CURATION & PREPROCESSING

Hyperparameter tool

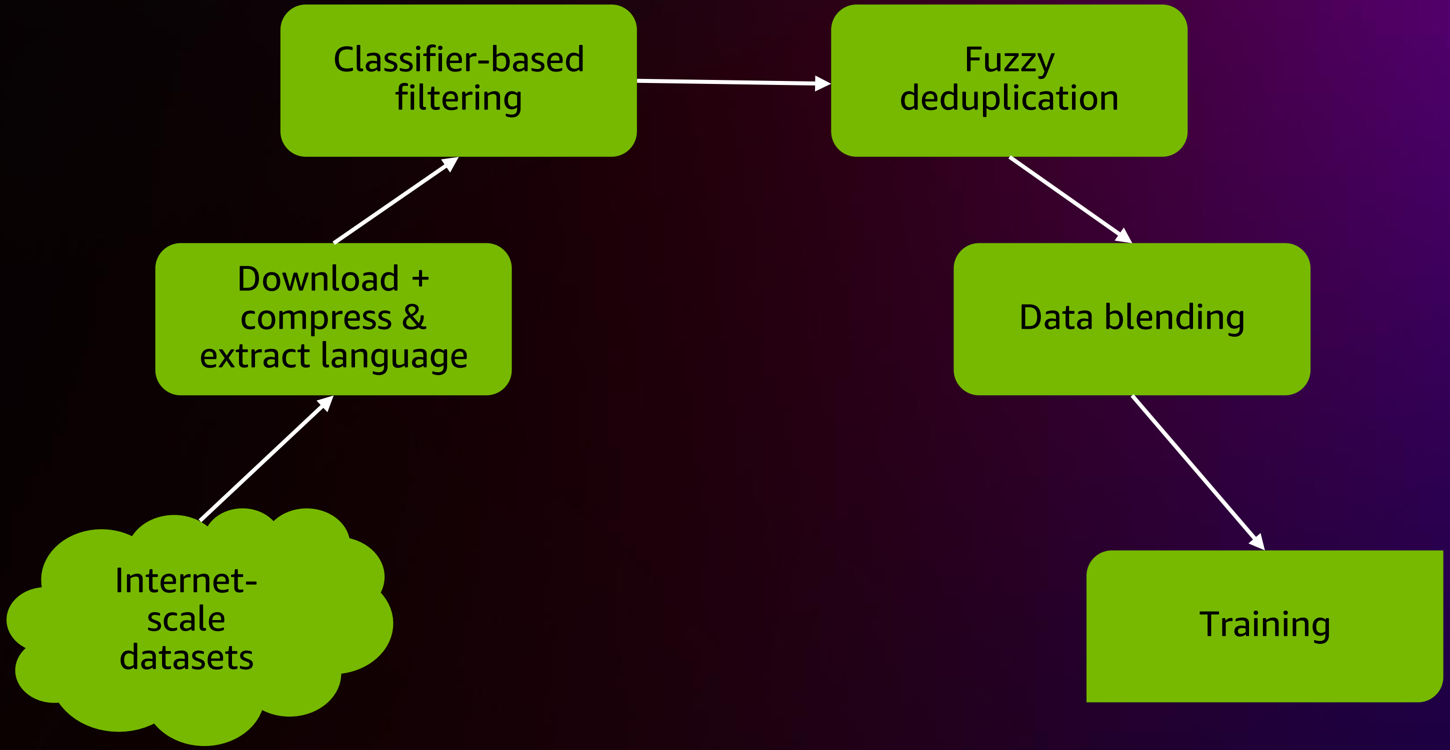


Training



Inference

Data curation & preprocessing

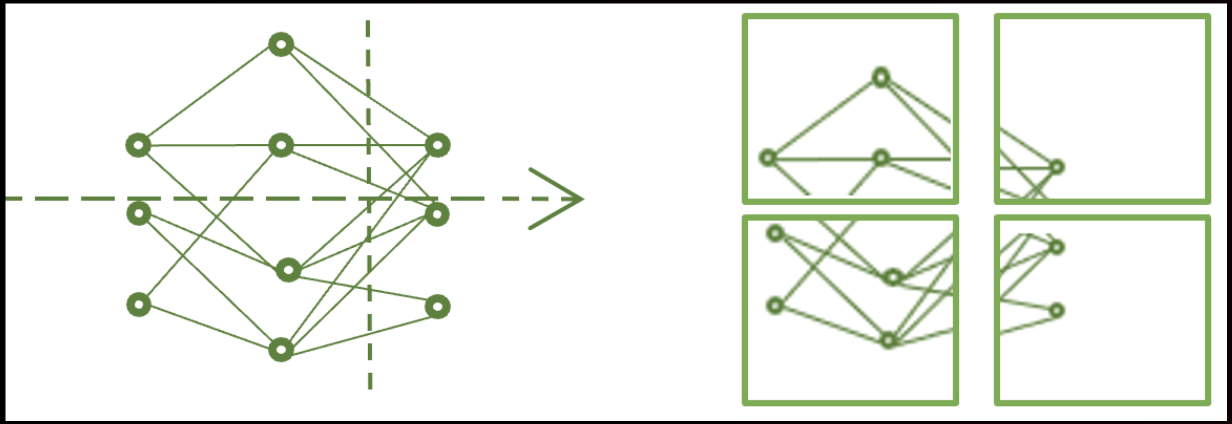


Bring your own dataset to train LLMs

Framework-agnostic distributed data curation tools for filtering, deduplication, and blending

Pipeline and Tensor parallelism for training

TRAINING MODELS AT SCALE



Maximize GPU utilization over InfiniBand and minimum latency within a single node

Pipeline (inter-layer) parallelism

- Split contiguous sets of layers across multiple GPUs
- Layers 0, 1, 2 and layers 3, 4, 5 are on different GPUs
- Exceptions and limitations: No interleaved scheduling

Tensor (intra-layer) parallelism

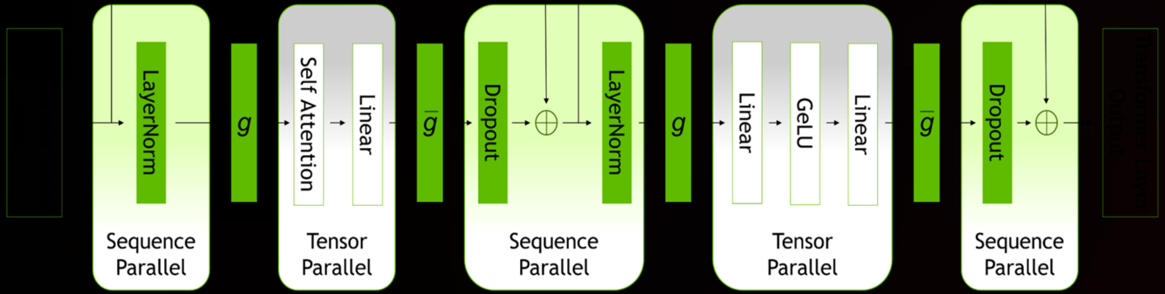
- Split individual layers across multiple GPUs
- Devices compute different parts of layers 0, 1, 2, 3, 4, 5
- Exceptions and limitations: Limited number of model architectures, GPT-3, and T5

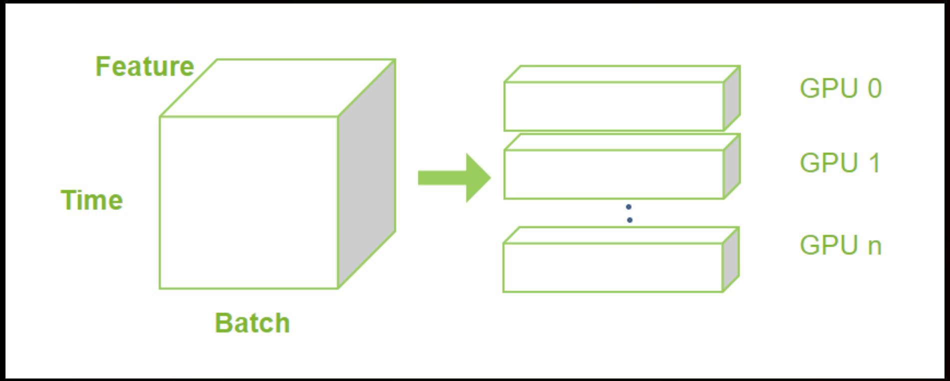
NeMo Megatron – Optimization techniques

SEQUENCE PARALLELISM AND SELECTIVE ACTIVATION RECOMPUTATION

Sequence parallelism

Increase throughput during back-propogation

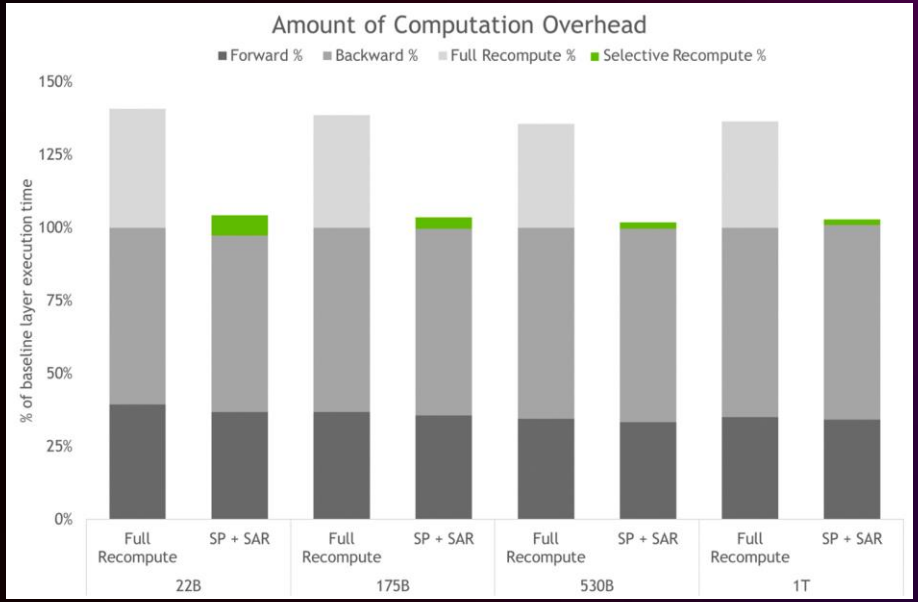


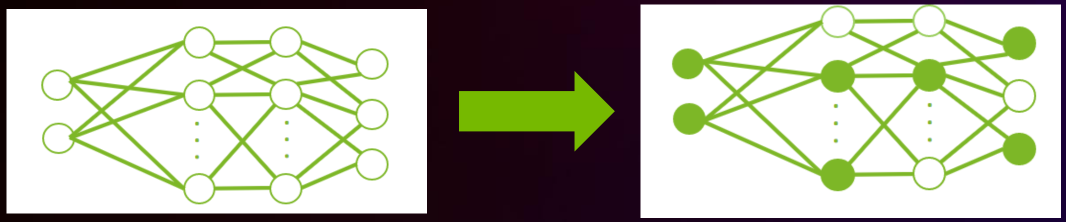


Reduce memory consumption of activation tensors to reduce recomputation of activations during back-prop

Selective activation recomputation

Optimize memory-throughput tradeoff during back-propagation





Lower memory footprint of activations and increase throughput of network

Benchmarking

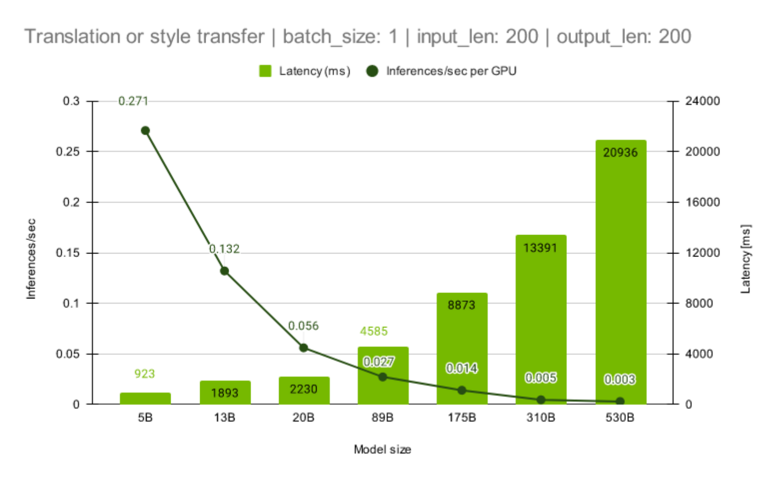
TRAINING AND INFERENCE

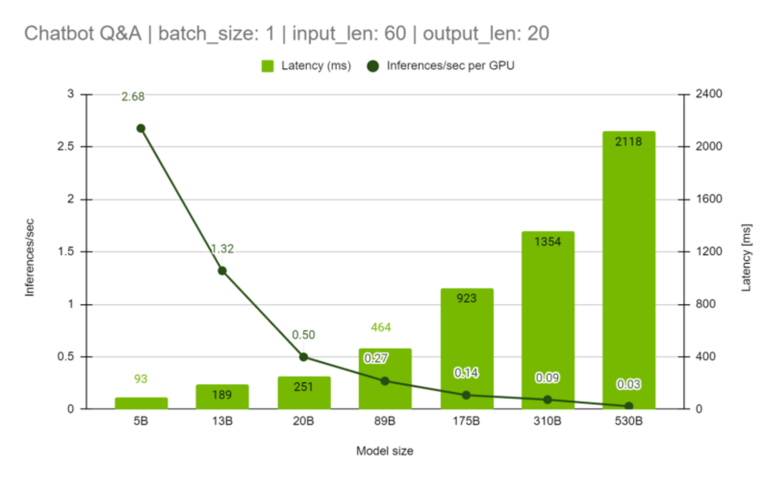
Training

GPT-3 Model Parameter Count	Estimated Time to Train			
	5 x SuperPod 100 nodes (days)	3 x SuperPod 60 nodes (days)	1 SuperPod 20 nodes (days)(weeks)	4 x DGX A100 (weeks)(years)
0.5B	0.21	0.21	0.62	0.44
1.7B	1	1	2	2
3.6B	1	1	4	3
7.5B	2	3	9	7
18B	5	8	23	16
39B	10	16	7	35
76B	19	31	13	1.3
145B	36	60	26	2.5
175B	43	72	31	3.0
310B	77	128	55	5.3
530B	131	219	94	9.0
1T	250	417	179	17.2

175B GPT-3 Training: ~1.5 months to train on 800 NVIDIA A100 80G GPUs

Inference





Solving pain points across the stack

NEMO MEGATRON SIMPLIFIES THE PATH TO AN LLM

Unmet needs		How we are helping
Large-scale data processing	→	Data curation and preprocessing tools
Multilingual data processing and training	→	Relative positional embedding (RPE) – multilingual support
Finding optimal hyperparameters	→	Hyperparameter tool
Convergence of models	→	Verified recipes for large GPT and T5-style models
Scaling on clouds	→	Scripts/configs to run on AWS
Deploying for inference	→	Model navigator + export to FT functionalities
Deployment at scale	→	Quantization to accelerate inferencing
Evaluating models in industry-standard benchmarks	→	Productization evaluation harness
Differing infrastructure setups	→	Full-stack support with FP8 and Hopper support
Lack of knowledge	→	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

Find out more

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

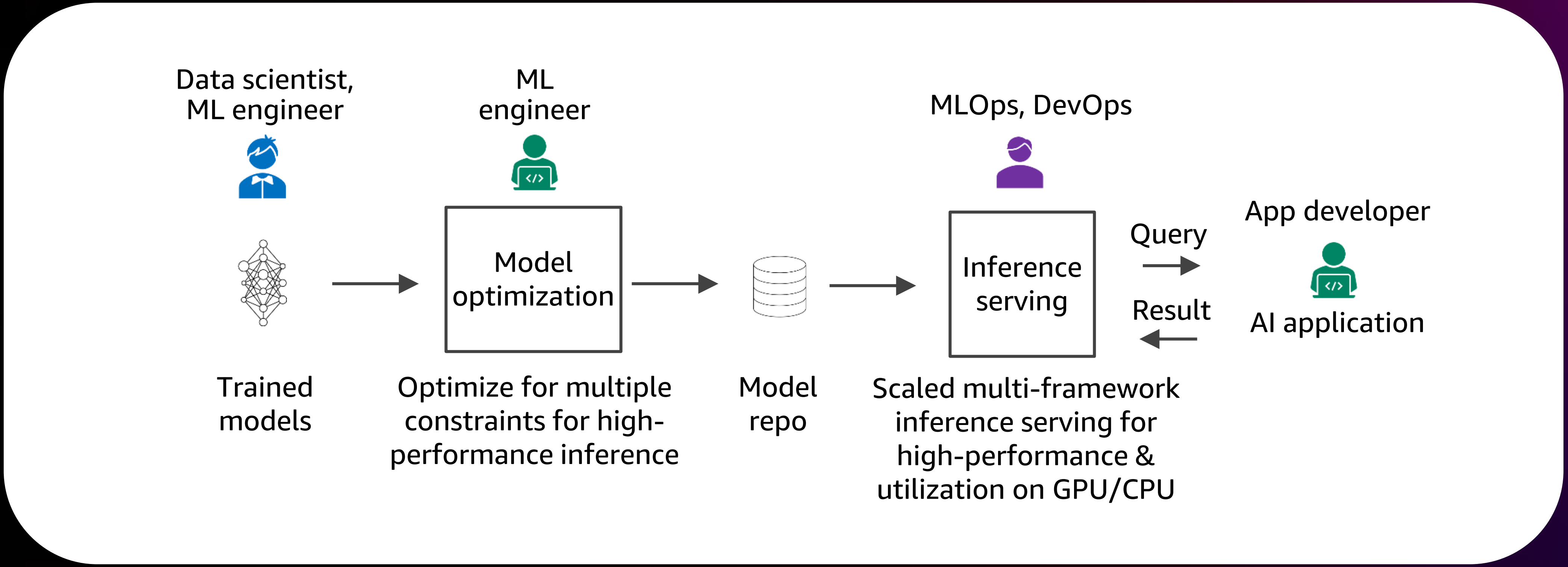
CUSTOMER STORIES

The King's Swedish: AI Rewrites the Book in Scandinavia

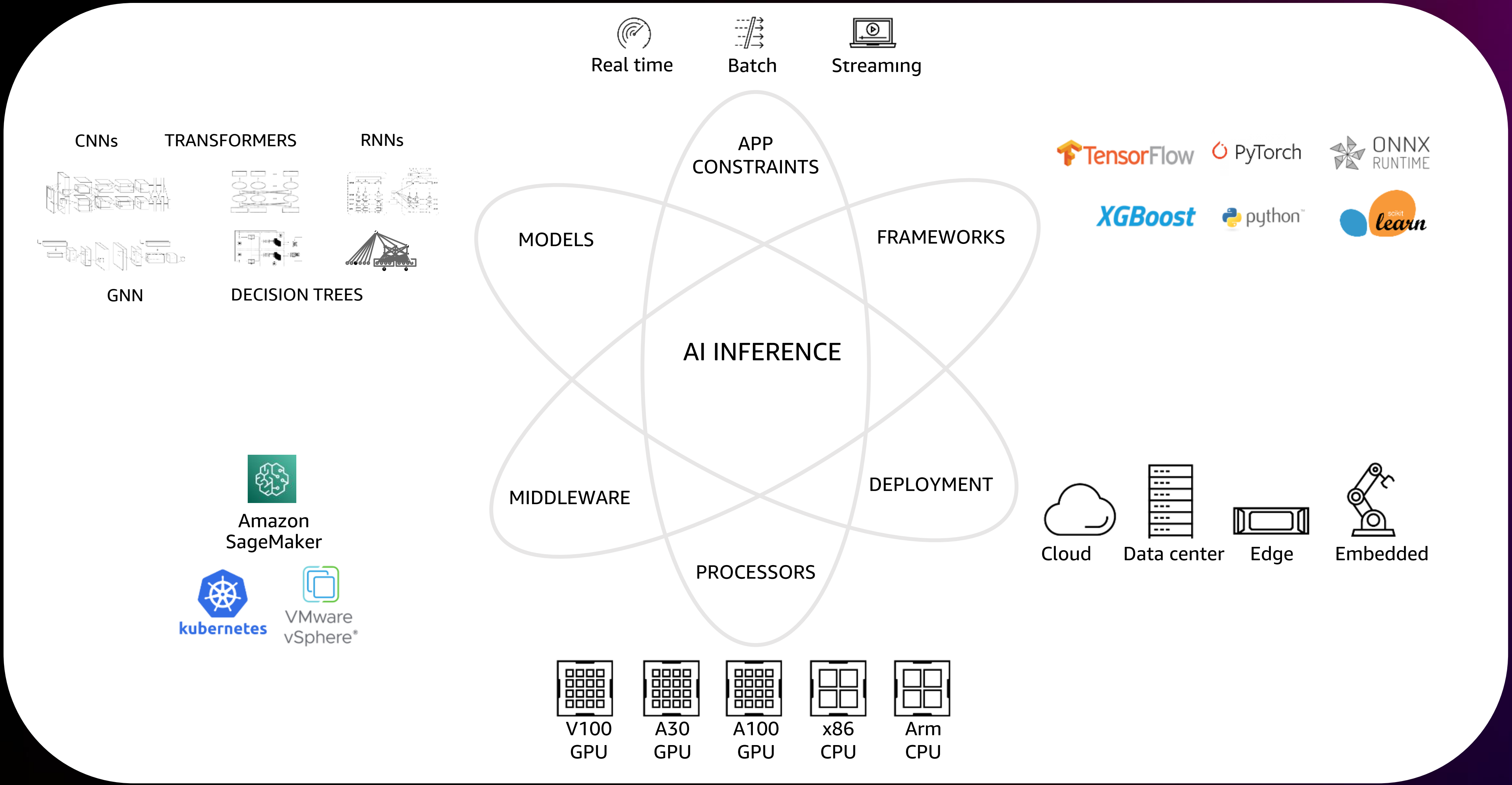
Deployment and inference

AI inference workflow

TWO-PART PROCESS IMPLEMENTED BY MULTIPLE PERSONAS

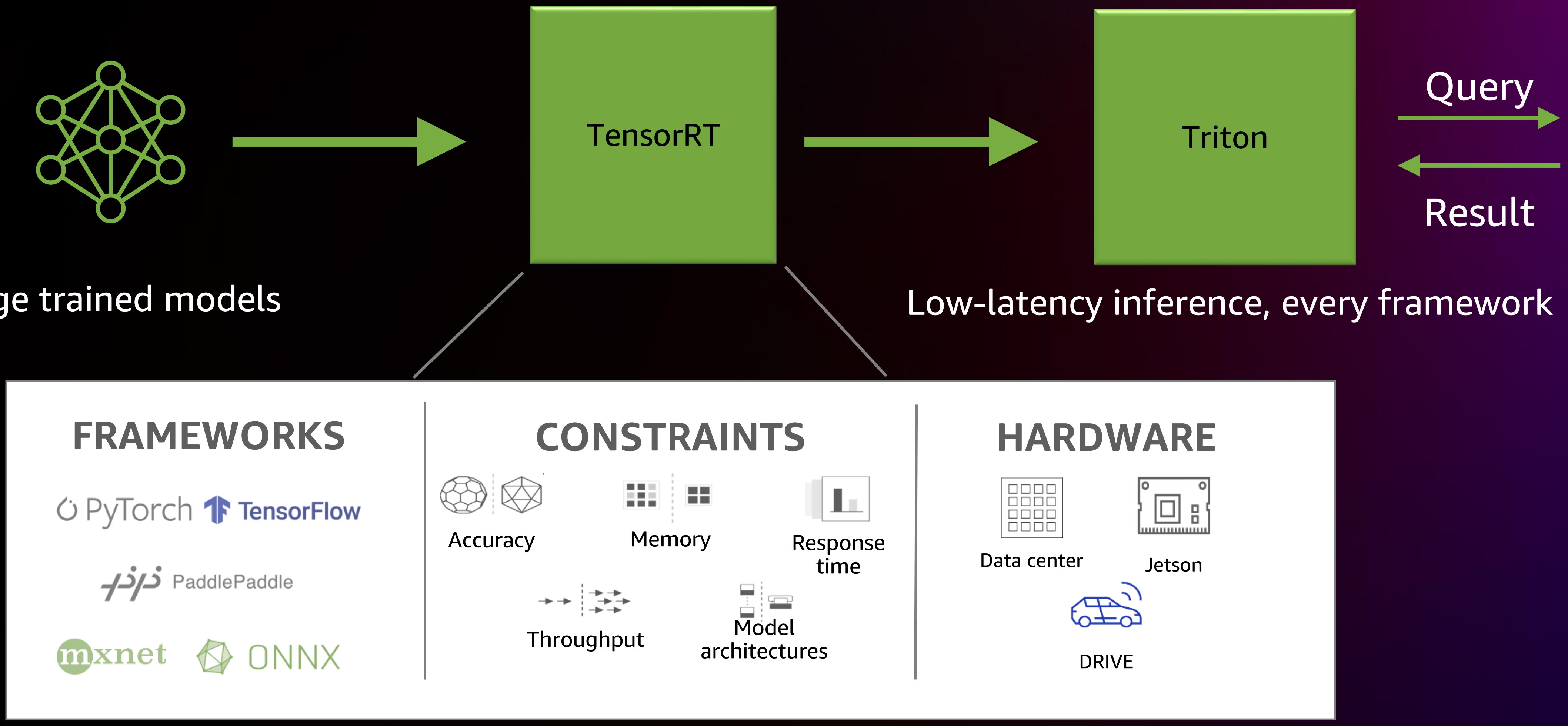


Challenges in AI inference



Inference is complex

REAL TIME | COMPETING CONSTRAINTS | RAPID UPDATES



A world-leading inference performance

TENSORRT ACCELERATES EVERY WORKLOAD

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

Hello

36x

Computer vision
< 7 ms

583x

Speech recognition
< 100 ms

21x

NLP
< 50 ms

10x

Reinforcement learning

Hello

178x

Text-to-speech
< 100 ms

12x

Recommenders
< 1 sec

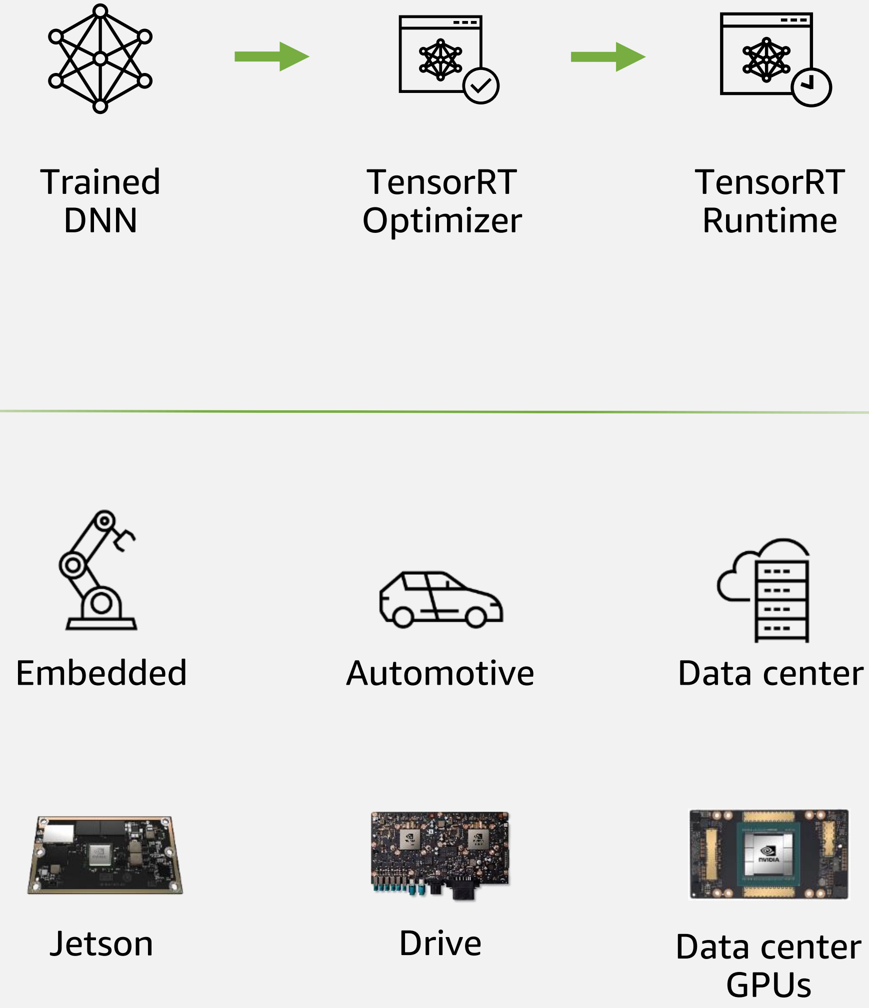
NVIDIA TensorRT

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



Layer and Tensor fusion

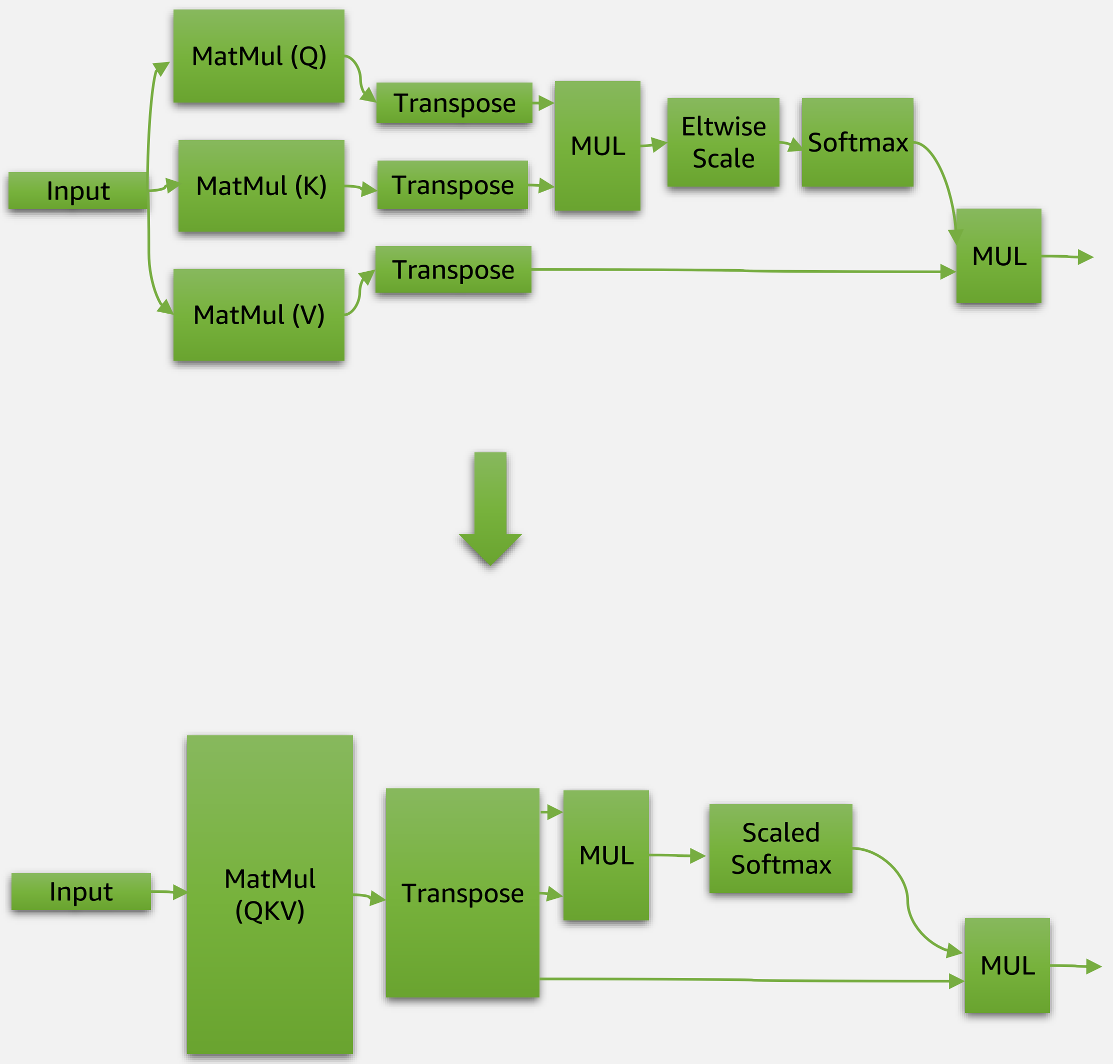
OPTIMIZES USE OF GPU MEMORY AND BANDWIDTH BY FUSING NODES IN A KERNEL

Combines successive nodes into a single node, making single-kernel execution

Significantly reduces number of layers to compute, resulting in faster performance

Eliminates unnecessary memory traffic by removing concat/slice layers

See the supported fusion list



developer.nvidia.com/tensorrt

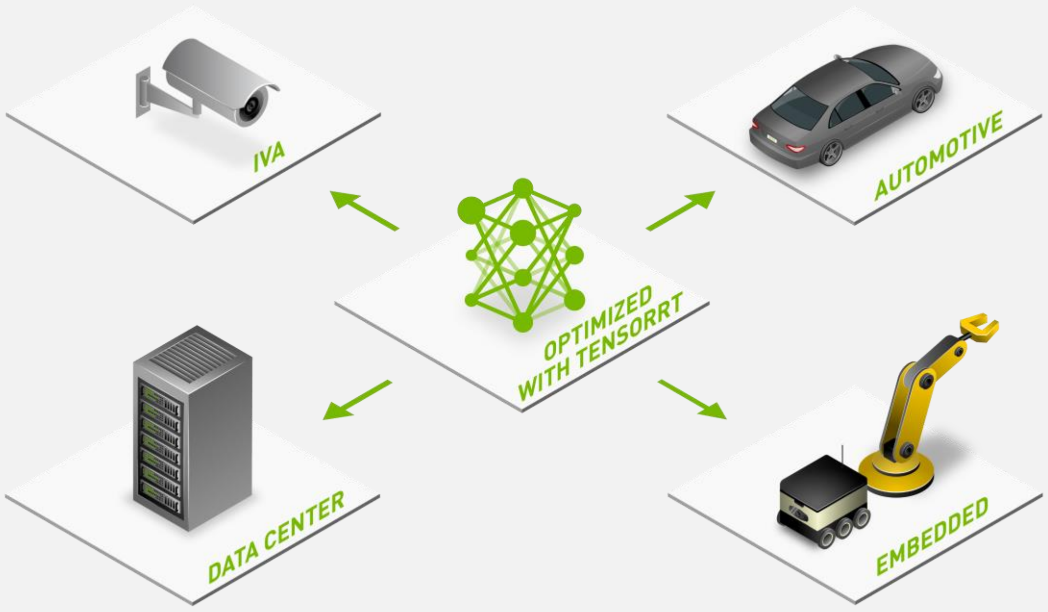
Kernel auto-tuning

SELECTS BEST DATA LAYERS AND ALGORITHMS BASED ON THE TARGET GPU PLATFORM

Hundreds of specialized kernels optimized for every GPU platform

TensorRT Optimizer uses runtime profile to select the best-performance kernels

Ensures best performance for specific deployment platform and specific neural network



developer.nvidia.com/tensorrt

Dynamic Tensor memory

MINIMIZES MEMORY FOOTPRINT AND REUSES MEMORY FOR TENSORS EFFICIENTLY

Reduces memory footprint and improves memory reuse

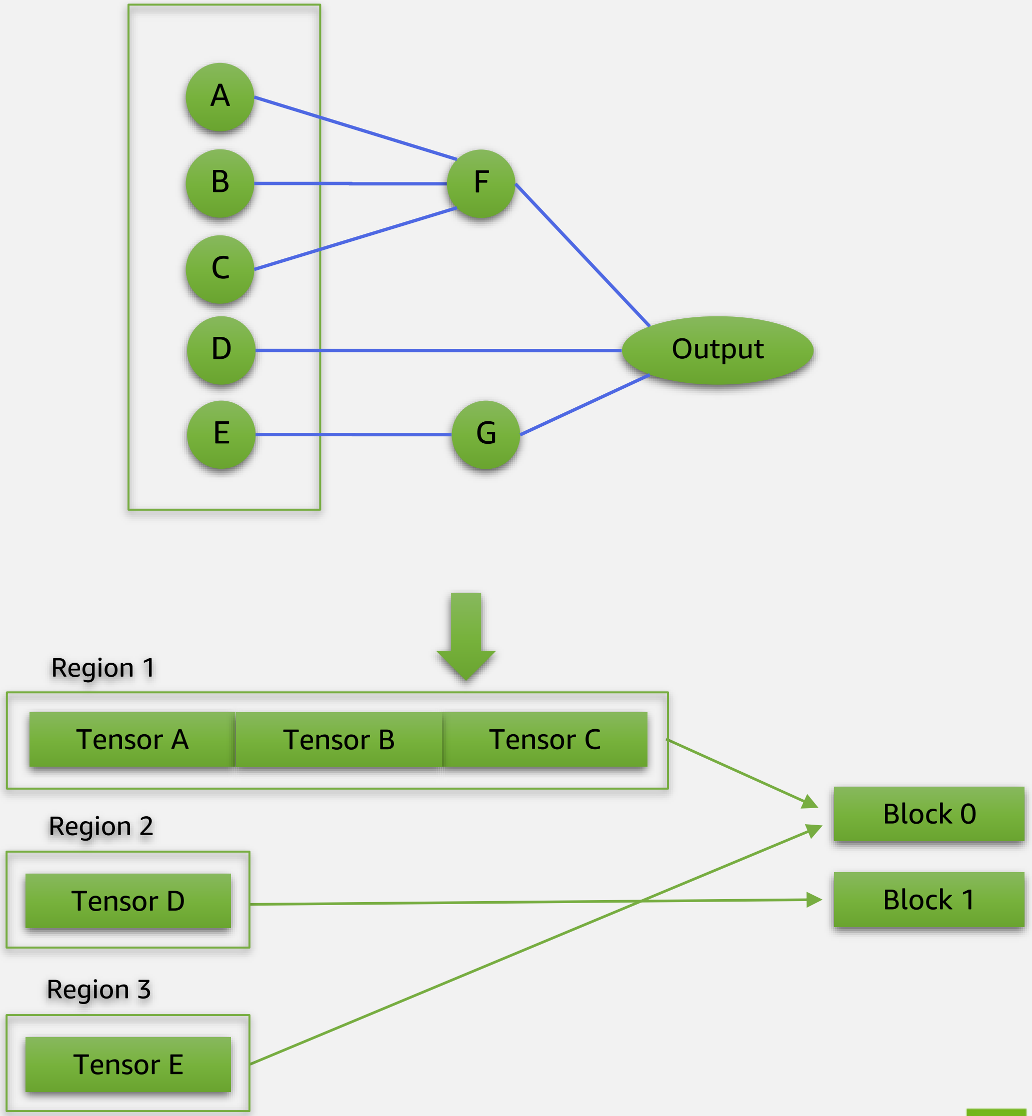
Graph optimizer combines tensors into regions

Region lifetime is a section of network execution time

Memory optimizer assigns regions to blocks; regions assigned to a block have disjoint lifetimes

Just like register allocation

developer.nvidia.com/tensorrt



Dynamic Tensor memory

MINIMIZES MEMORY FOOTPRINT AND REUSES MEMORY FOR TENSORS EFFICIENTLY

Reduces memory footprint and improves memory reuse

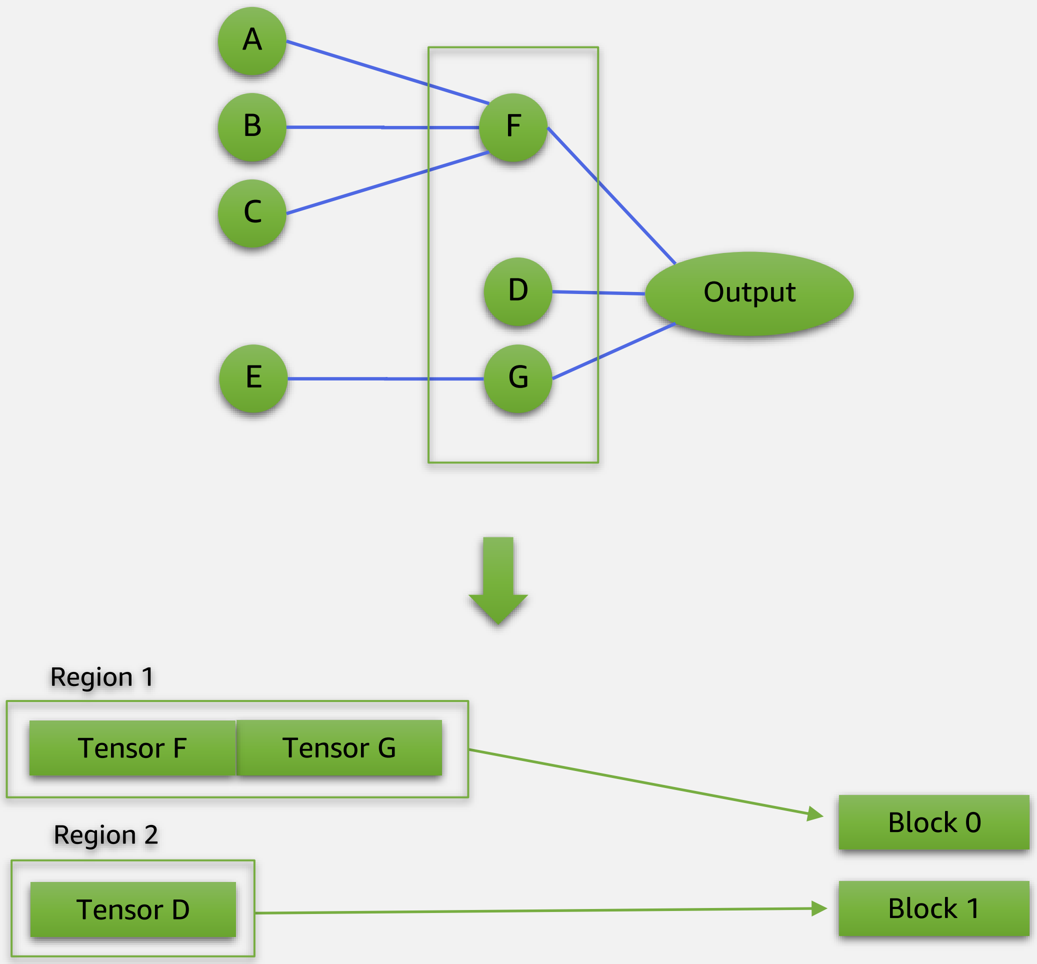
Graph optimizer combines tensors into regions

Region lifetime is a section of network execution time

Memory optimizer assigns regions to blocks; regions assigned to a block have disjoint lifetimes

Just like register allocation

developer.nvidia.com/tensorrt



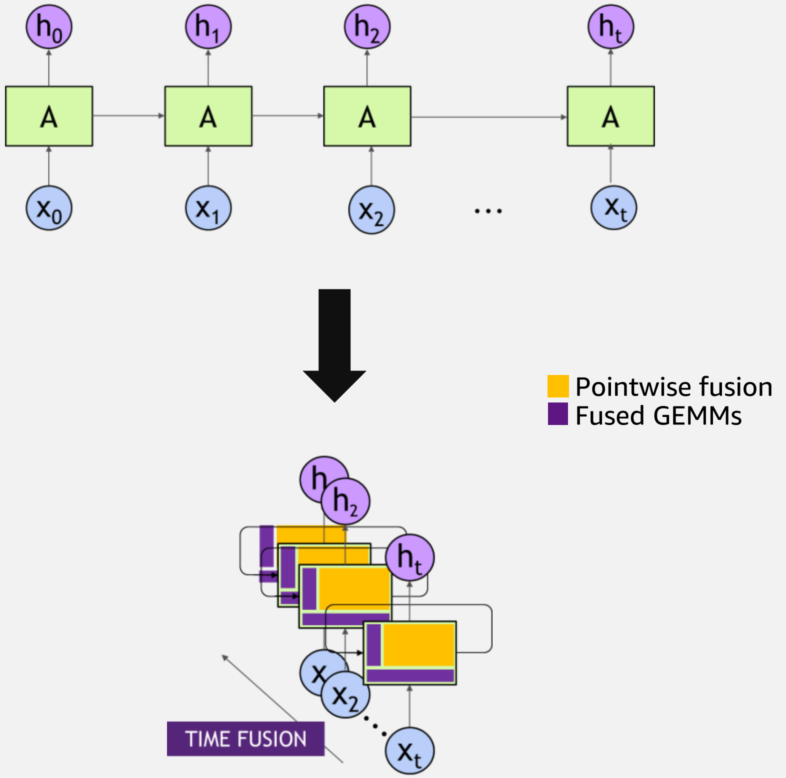
Time fusion

OPTIMIZES RECURRENT NEURAL NETWORKS OVER TIME STEPS WITH DYNAMICALLY GENERATED KERNELS

Recurrent neural network optimizations

Deploy highly optimized ASR and TTS

Compiler fuses pointwise ops, fuses GEMMs, and computes efficiently across time steps

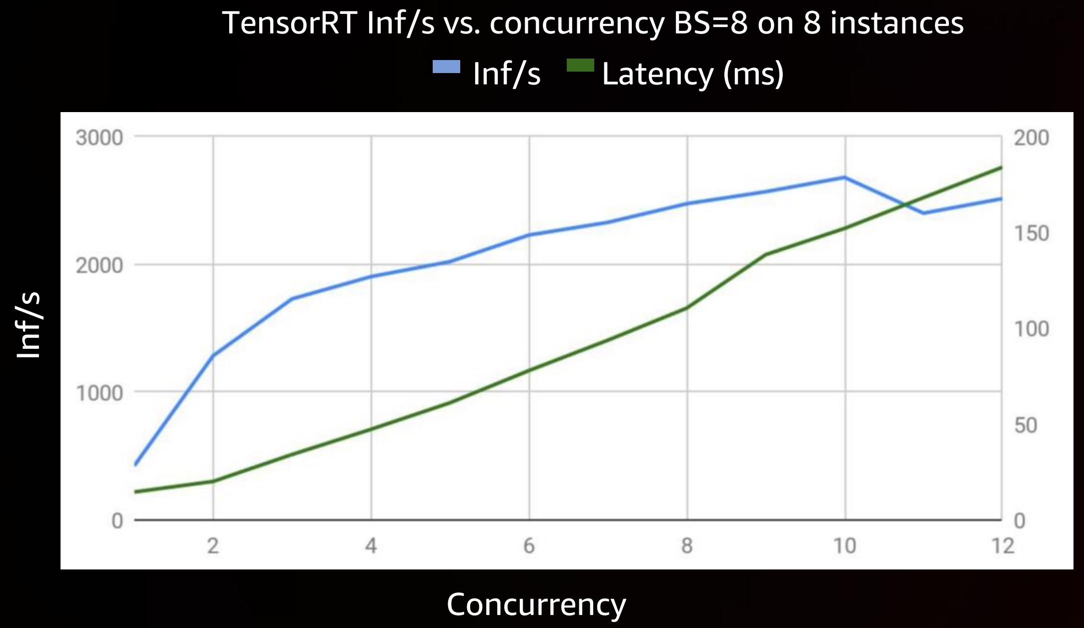


developer.nvidia.com/tensorrt

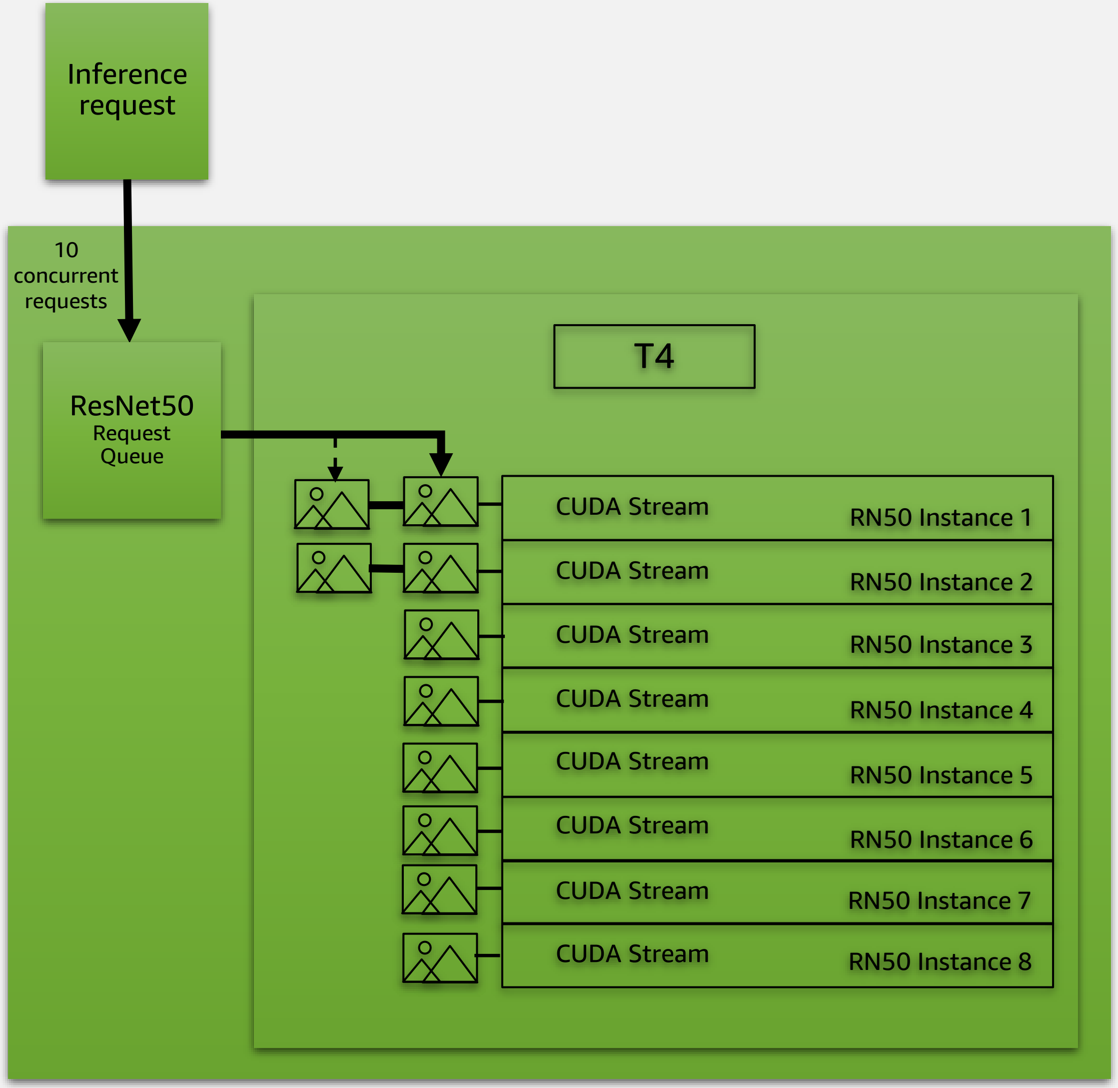
Multi-stream concurrent execution

USES A SCALABLE DESIGN TO PROCESS MULTIPLE INPUT STREAMS IN PARALLEL

6x better performance and improved GPU utilization through multi-stream concurrent execution



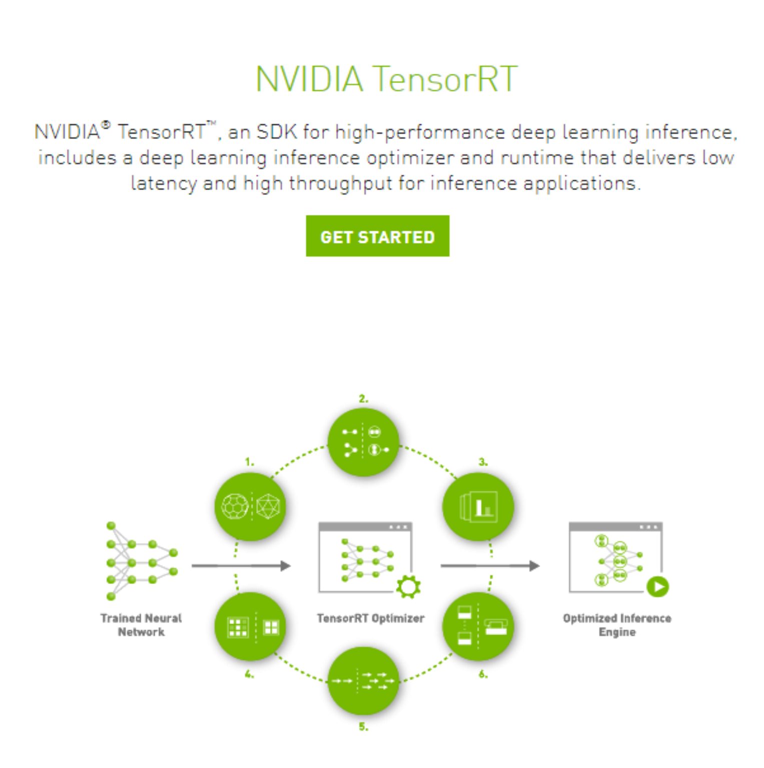
developer.nvidia.com/tensorrt



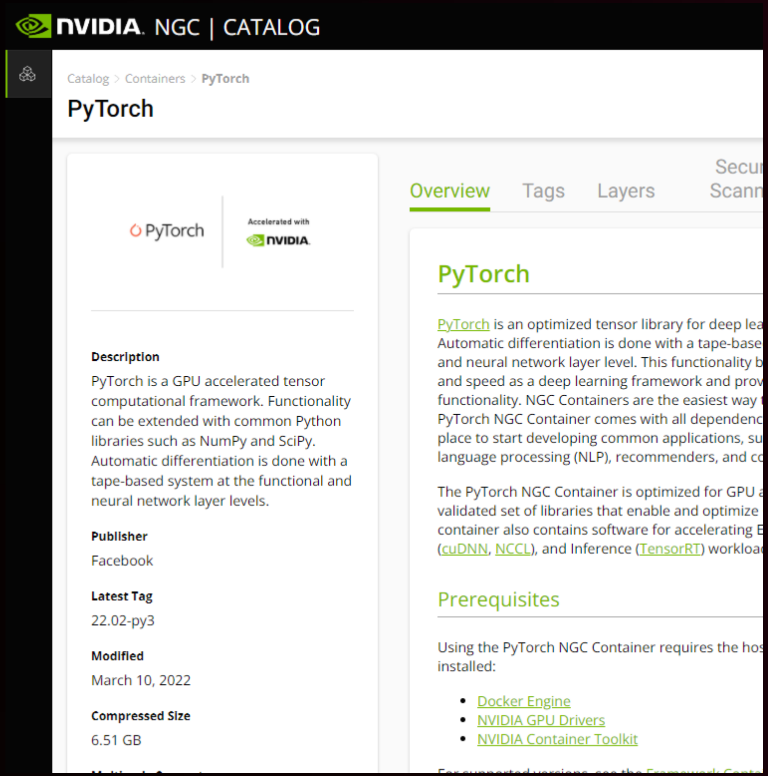
Download TensorRT today

TENSORFLOW WITH TENSORRT

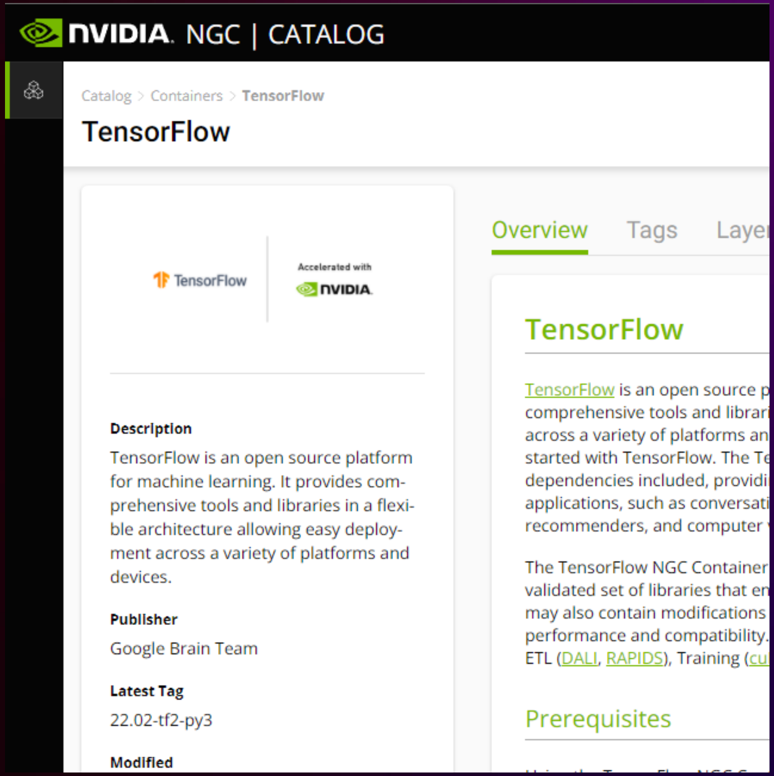
TensorRT



Torch-TensorRT



TensorFlow-TensorRT



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

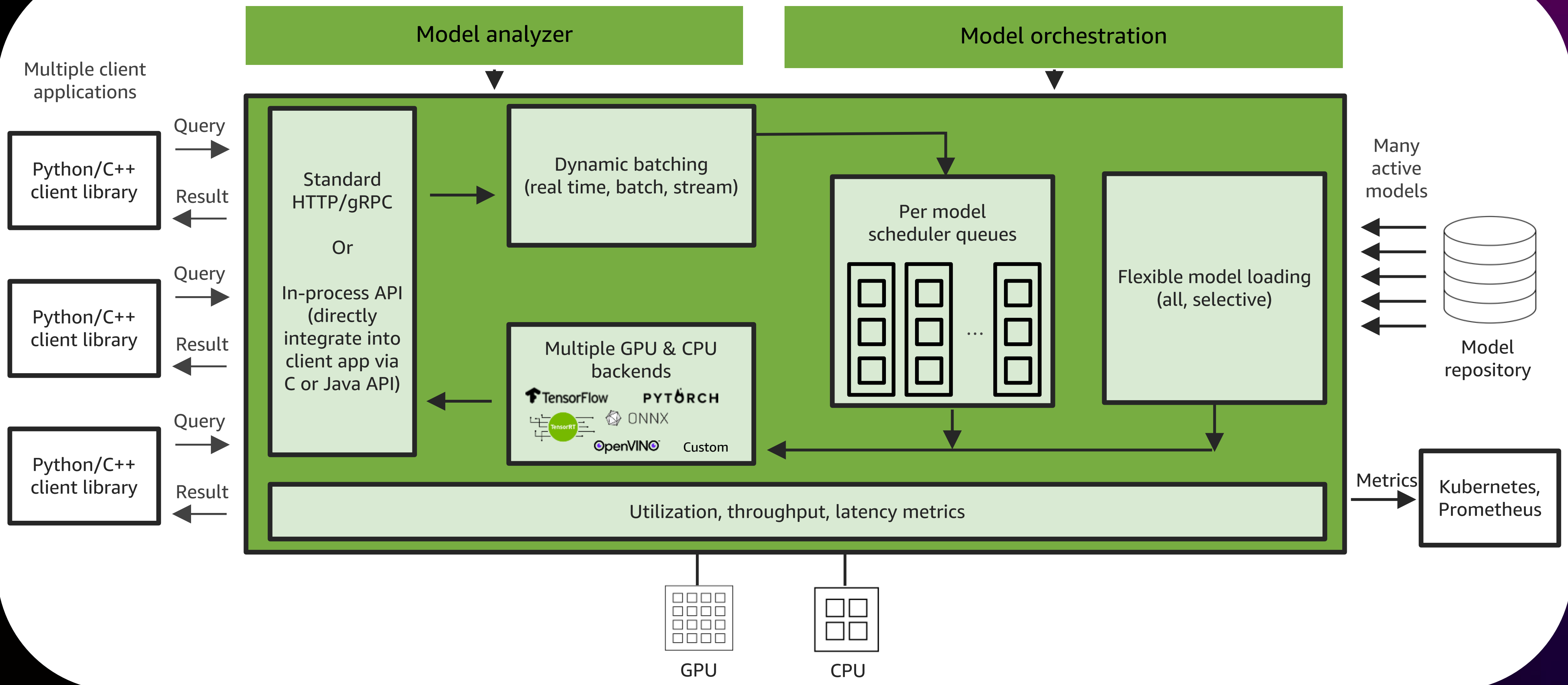
Performance & utilization

Model Analyzer for optimal configuration

Optimized for high GPU/CPU utilization, high throughput & low latency

Triton's architecture

DELIVERING HIGH PERFORMANCE ACROSS FRAMEWORKS



Inference with many natively supported backends

TensorFlow 1.x/2.x Any model SavedModel | GraphDef	**PyTorch** Any model JIT/TorchScript | Python	**TensorRT** All TensorRT-optimized models	**TF-TensorRT & TorchTRT** Any TensorFlow and PyTorch model
Forest Inference Library (FIL) Tree-based models (e.g., XGBoost, Scikit-learn RandomForest, LightGBM)	**ONNX RT** ONNX converted models	**Python** Custom code in Python; e.g., pre-/post-processing, any Python model	**Faster transformer backend (beta)** Multi-GPU, multi-node inferencing for large transformer models (GPT and T5)
OpenVINO OpenVINO-optimized models on Intel architecture	**Custom C++ backend** Custom framework in C++	**DALI** Pre-processing logic using DALI operators	**NVTabular** Feature engineering and preprocessing library for tabular data
HugeCTR Recommender model with large embeddings			

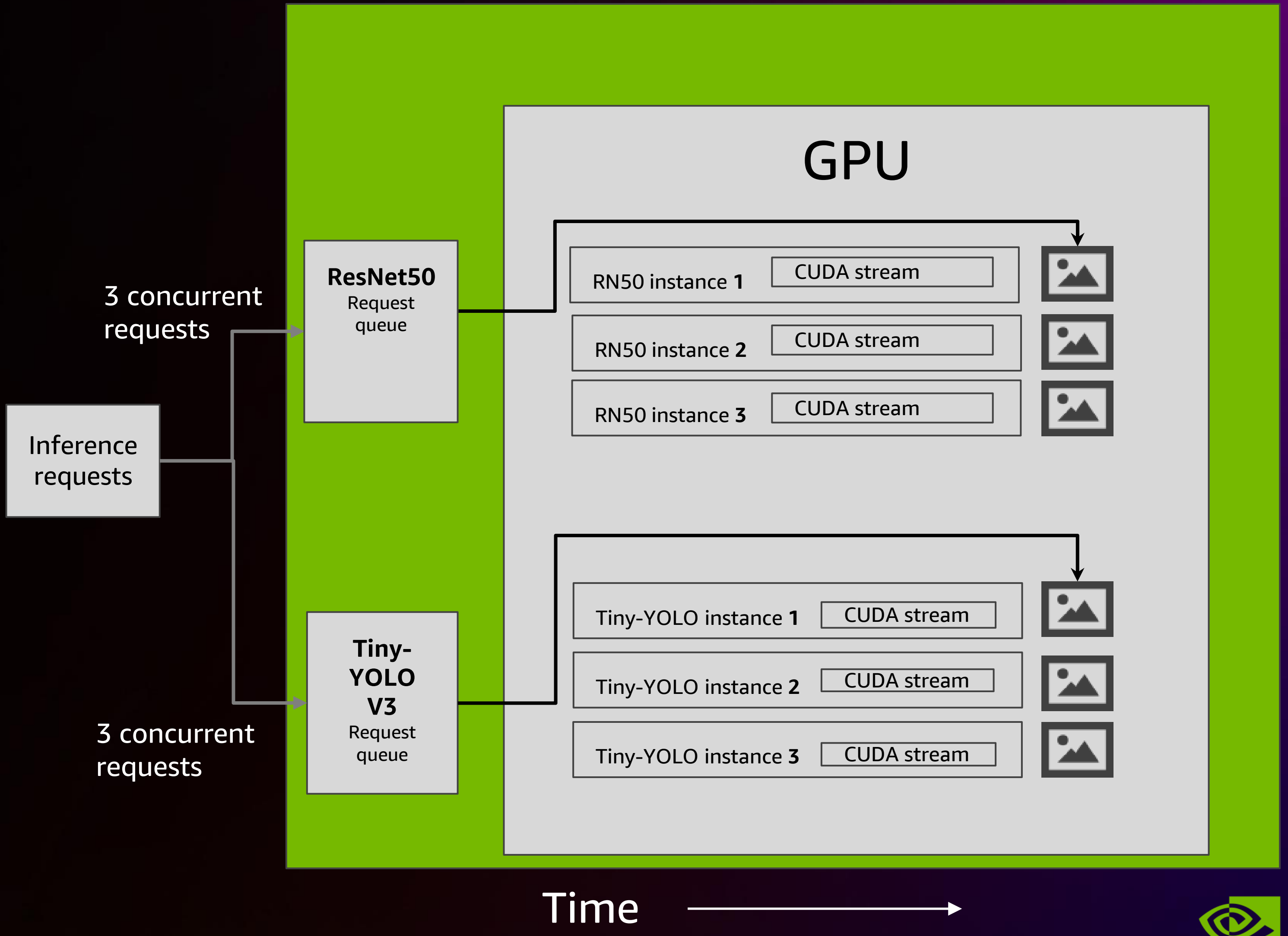
Concurrent model execution

INCREASE THROUGHPUT AND UTILIZATION

Triton can run concurrent inference on

1. Multiple different models
2. And/or multiple copies of the same model in parallel on the same system

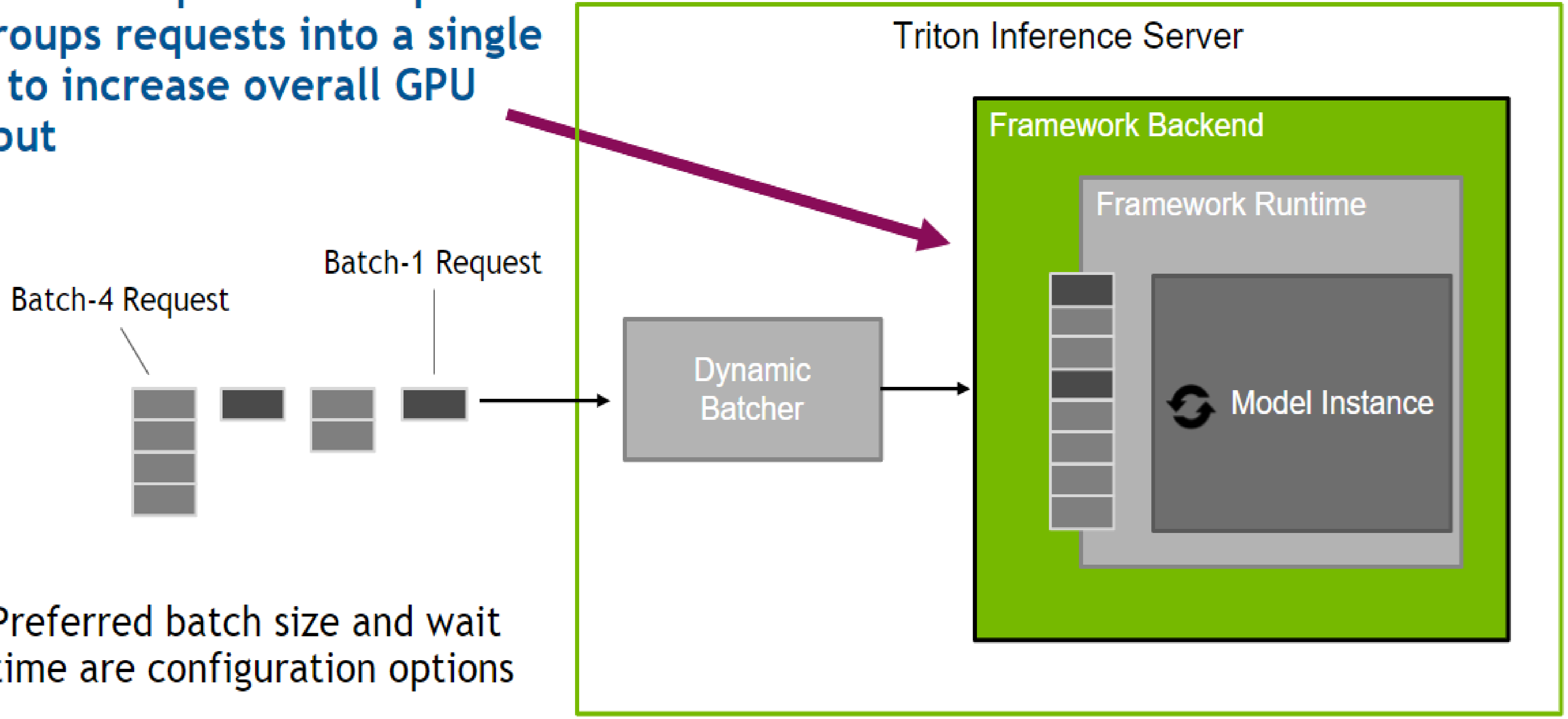
Maximizes GPU utilization, enabling better performance and lowering the cost of inference



Dynamic batching scheduler

GROUP REQUESTS TO FORM LARGER BATCHES, INCREASE GPU UTILIZATION

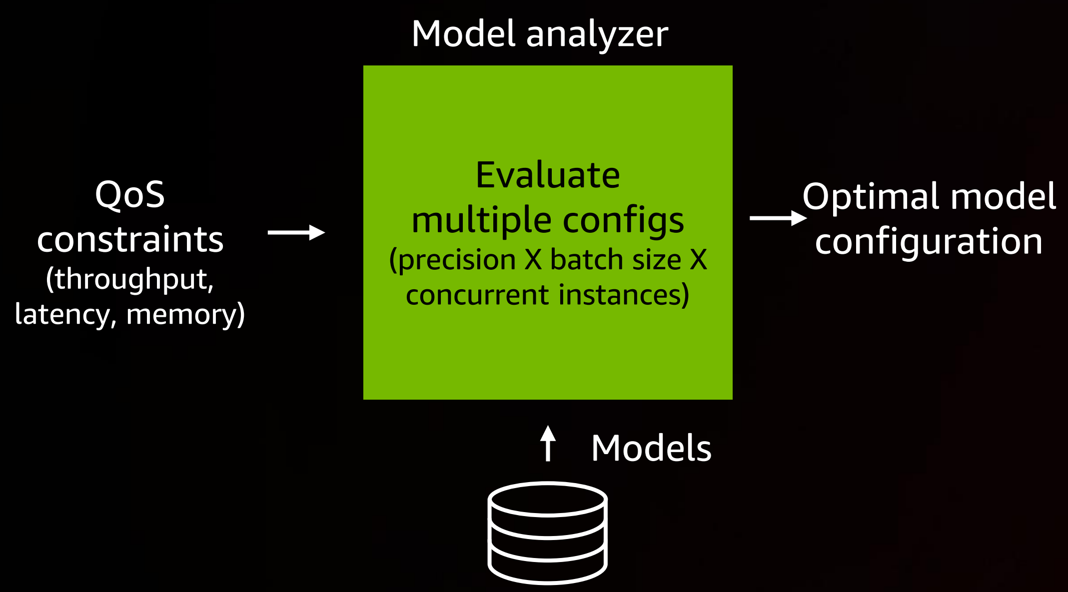
Client sends independent requests. Triton groups requests into a single “batch” to increase overall GPU throughput

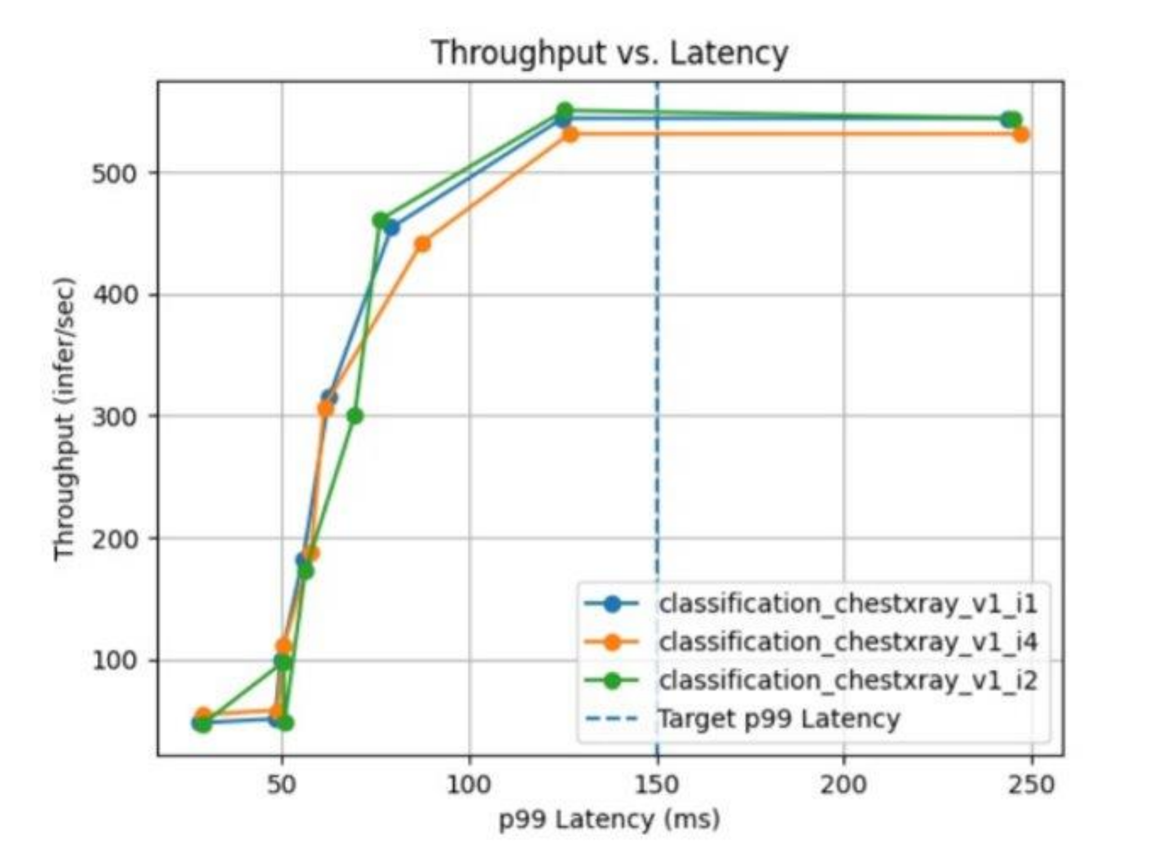


Preferred batch size and wait time are configuration options

Optimal model configuration

USING THE MODEL ANALYZER CAPABILITY





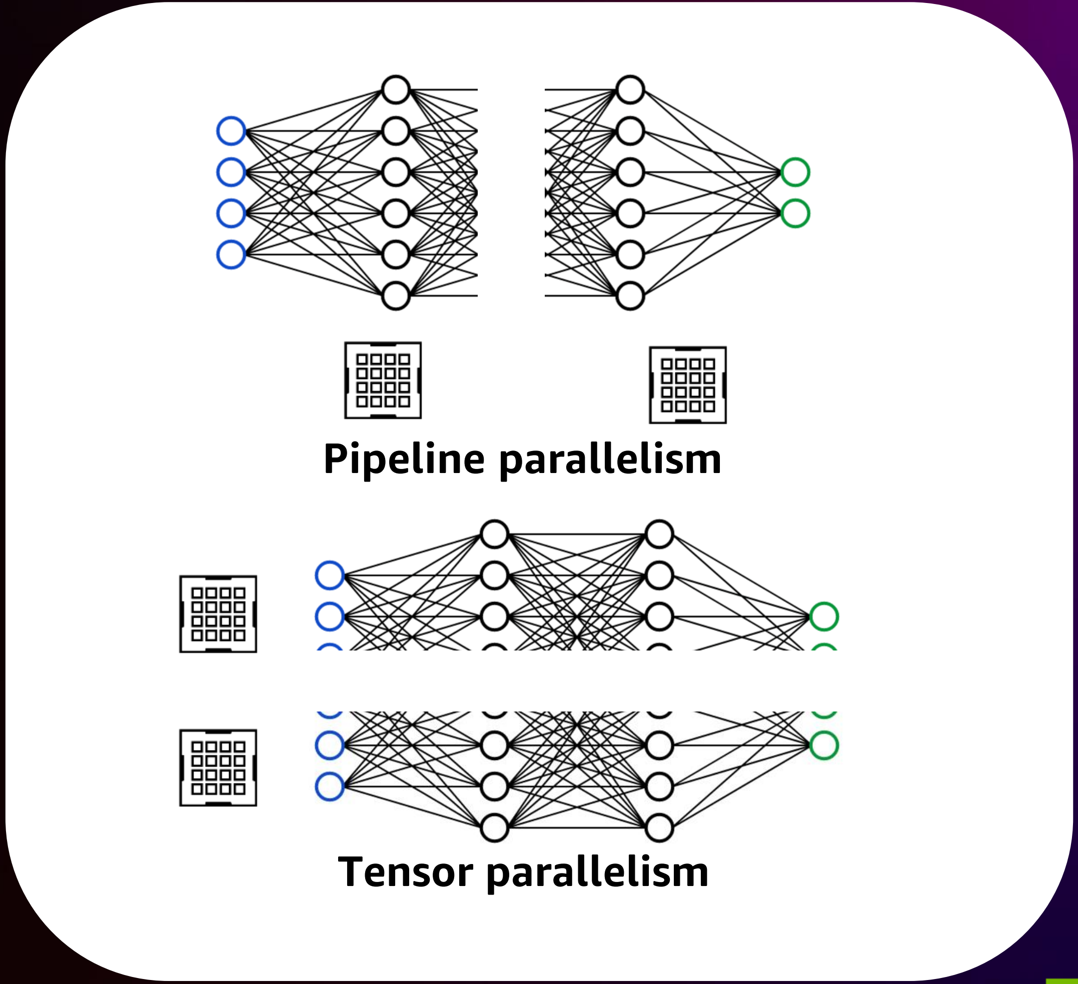
Throughput vs. Latency curves for 3 configurations of classification_chestxray_v1

GitHub repo and docs: https://github.com/triton-inference-server/model_analyzer

Large language model inference

USING TRITON'S FASTERTRANSFORMER BACKEND

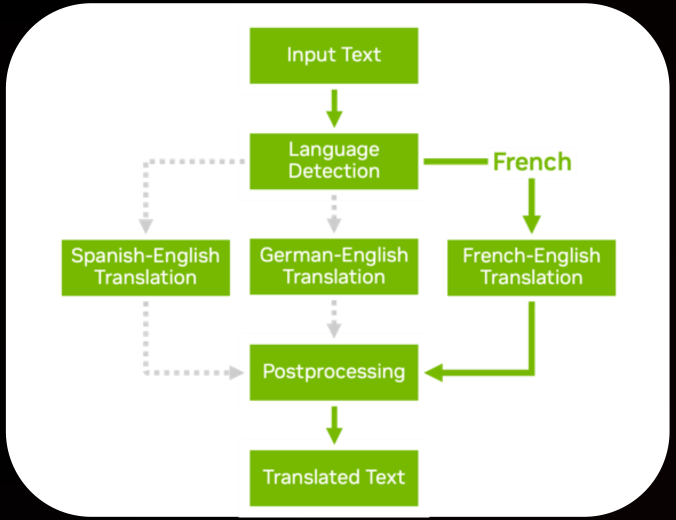
- Multi-GPU multi-node inference of large transformer models: GPT, GPT-J, GPT-NeoX, T5, UL2, OPT
- Tensor and pipeline parallelism
- Uses MPI and NCCL for efficient inter/intra node communication
- Supports FP32, FP16, BF16 in beta release, available via:
  - NeMo Megatron EA program: https://developer.nvidia.com/nemo-megatron-early-access
  - GitHub repo: https://github.com/triton-inference-server/fastertransformer_backend



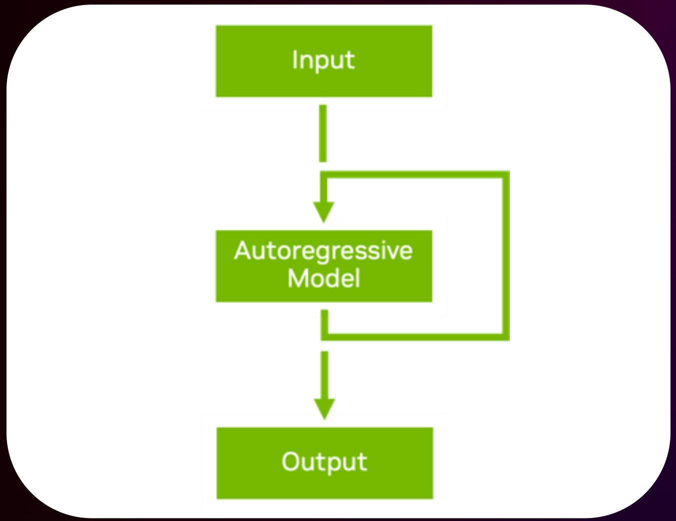
Model pipelines with business logic scripting

CONTROL FLOW AND LOOPS IN MODEL ENSEMBLES

- **Model ensembles beyond simple pipelines:** Enables arbitrary ordering of models, with conditionals, loops, and other custom control flow
- **Call any other backend from the Python or C++ backend:** Triton will efficiently pass data to and from the other backend



Conditional model execution

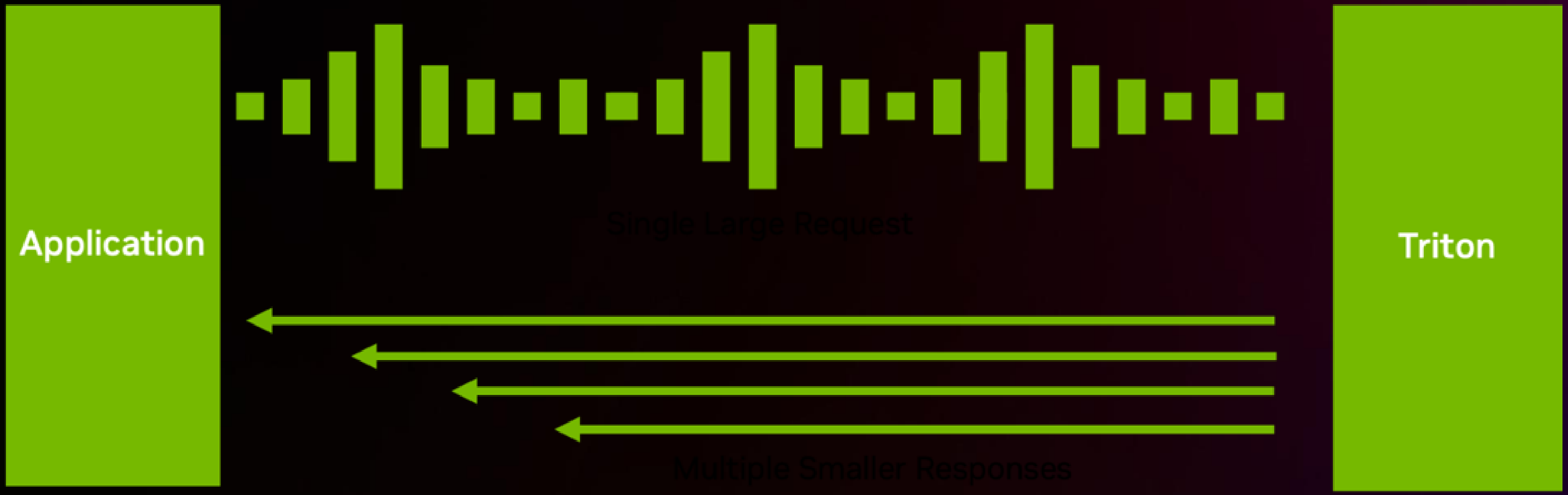


Looping execution for autoregressive models

Decoupled models

ALLOWS 0, 1, OR 1+ RESPONSES PER REQUEST

- For situations where requests and responses are not strictly 1 to 1 (e.g., automated speech recognition use case)
- Supported in C++ and Python backend
- Requires use of gRPC bidirectional streaming API or in-process C/Java API



Delivering value across industries

Search & ads

Payment fraud detection

Digital copyright management

Medical imaging

Alibaba text to speech for smart speaker

Image processing for autonomous driving

Grammar check

Meeting transcription

Document translation

Image classification & recommendation

Preventive maintenance

Defect detection

Multiple use cases

Contact center speech analytics

Package analytics

ByteDance Volcengine ML Platform

Fraud, fintech

Similar image search

Retail product identification

Financial fraud detection

Clinical notes analytics

NLP services

Coding assistant

AI character generation

Real-time spell check for product search

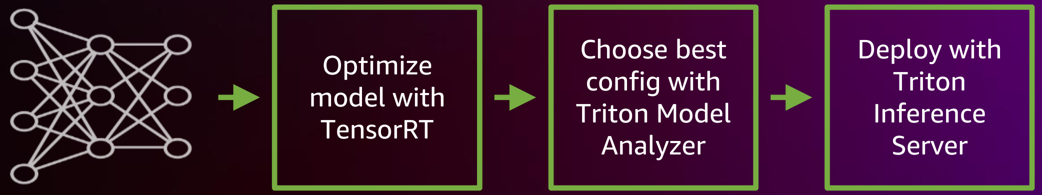
AMAZON SEARCH

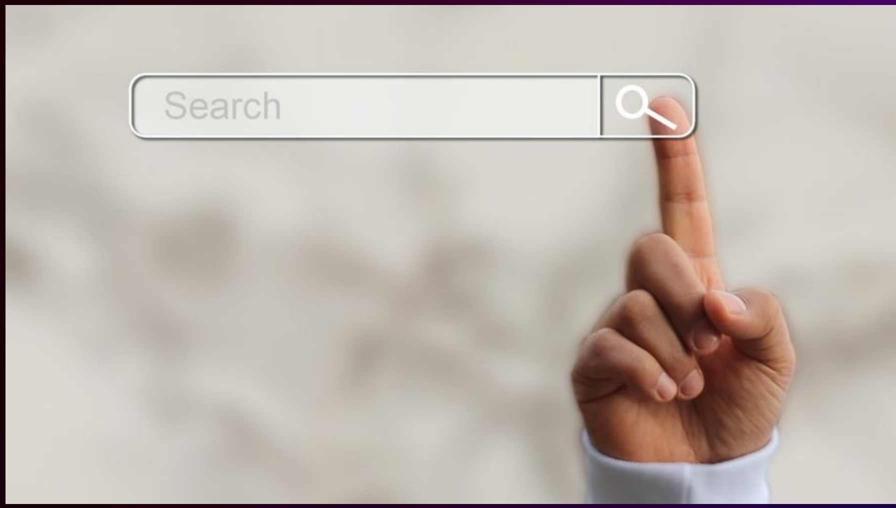
One of the most visited ecommerce websites

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

Benefits of Triton Inference Server on Amazon SageMaker

Cost-effective: Amazon SageMaker optimizes scale and performance to reduce costs

MLOps-ready: Includes metadata persistence, model management, logging metrics to Amazon CloudWatch, and monitoring data drift and performance

Flexible: Ability to run real-time inference for low latency, offline inference on batch data, and asynchronous inference for longer inference times

Secure: High bar on security, with available mechanisms including encryption at rest and in transit, VPC connectivity and fine-grained IAM permissions

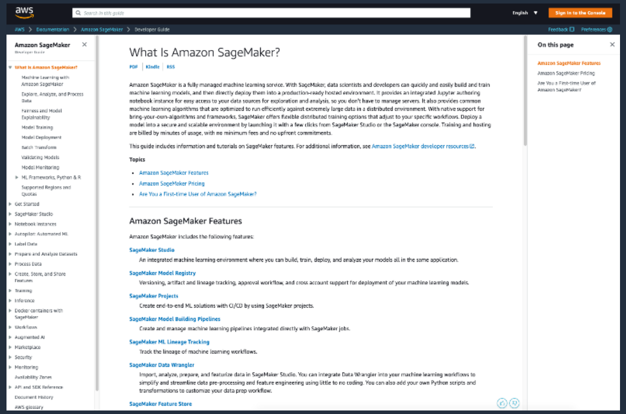
Less heavy-lifting: No container registry management, no custom installation of libraries, no extra SDK configuration

Get started with Triton Inference Server on Amazon SageMaker

Sample notebooks

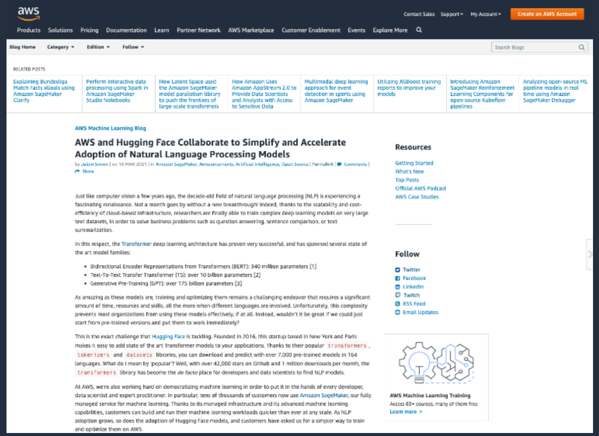
[Access GitHub](#)

Documentation



[AWS documentation](#)

Launch blog



[Read the blog](#)

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

Call to action and conclusions

Conclusions

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://developer.nvidia.com/developer-program

Thank you!

Jiahong Liu
jiahongl@nvidia.com

Edwin Weill
eweill@nvidia.com

Please complete the session survey in the **mobile app**