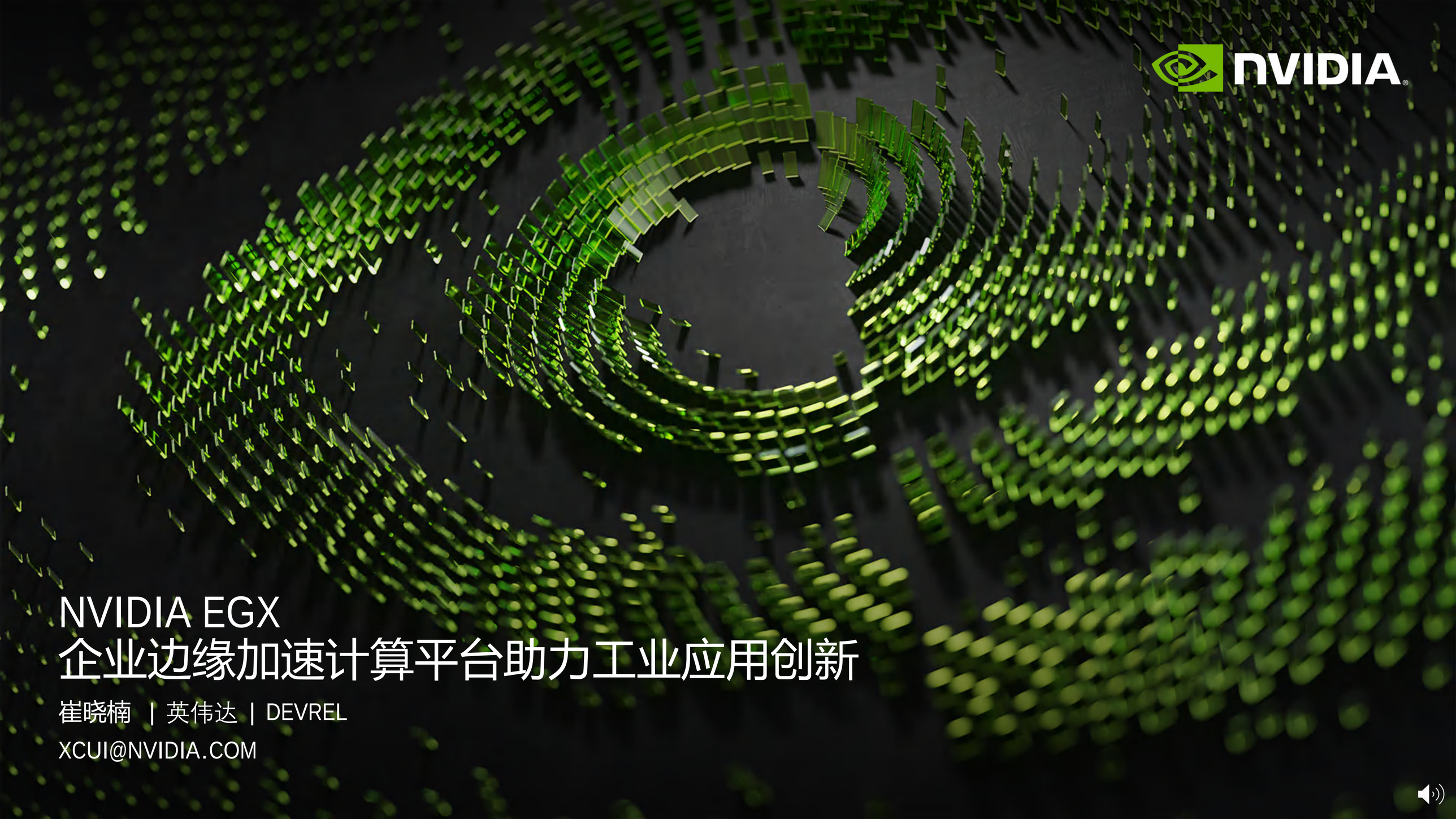


NVIDIA EGX
企业边缘加速计算平台助力工业应用创新

崔晓楠 | 英伟达 | DEVREL
XCUI@NVIDIA.COM

AGENDA

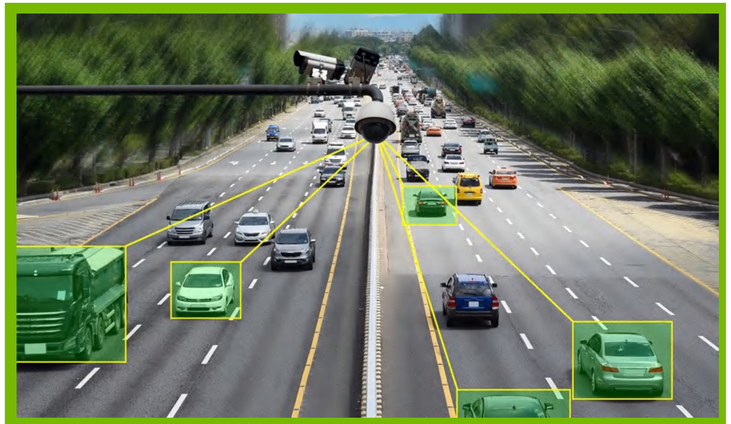
1. NVIDIA EGX Platform
2. EGX Reference Design - Inspection
3. FAQ

"Industrial edge AI will likely be the largest AI opportunity. A few percent of productivity and cost savings to trillion-dollar industries is a lot. The industrial edge will have hundreds of billions of things continuously operating at computer speeds."

NVIDIA CEO & Founder,
Jensen Huang

加速计算正在从云端向边缘侧扩展

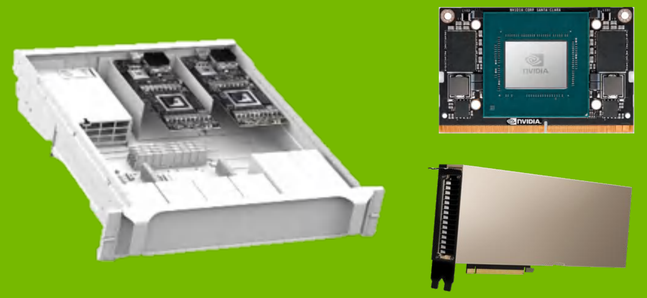
云 - DATA CENTER



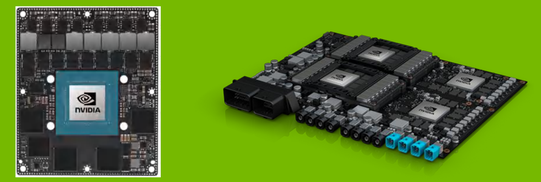
边 - EDGE COMPUTING

端 - AUTONOMOUS

DGX / HGX



EGX



AGX

One Architecture - CUDA

企业边缘加速计算平台

NVIDIA EGX PLATFORM

IVA　Retail　Logistics　ITS　Manufacturing　Telco　...

加速应用和框架生态系统

Fleet Command
云边协同

容器编排和管理集成

Bare Metal　Virtualization

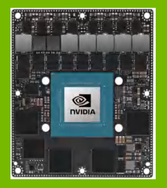
NVIDIA GPU

NVIDIA SmartNIC/DPU

NVIDIA CERTIFIED

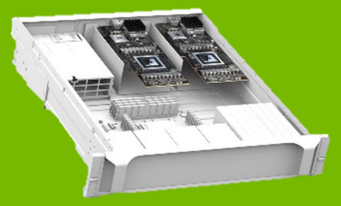
Embedded SoC

- Jetson NX/Xavier
- Jetson ORIN



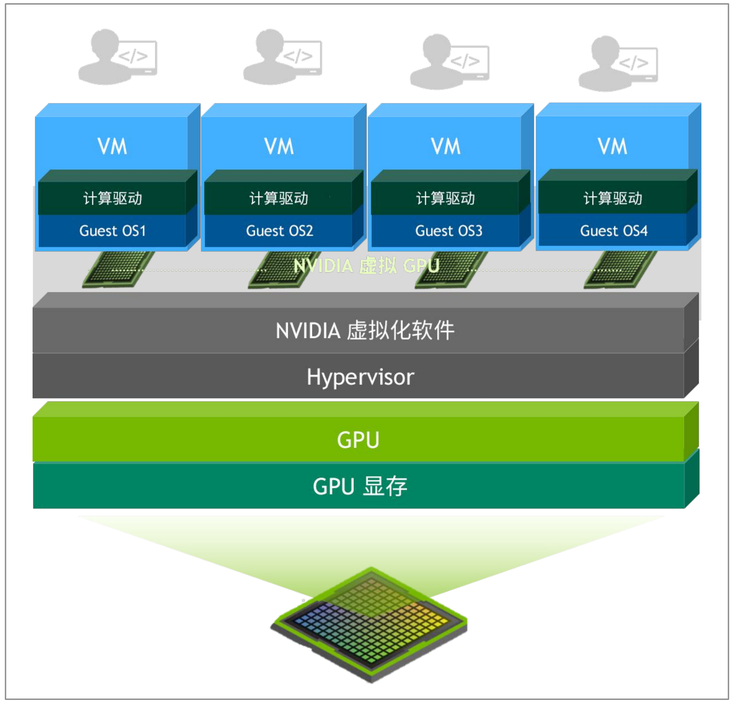
EGX Server

- A100, A30, A40
- ConnectX-6, ConnectX-6 Lx, BlueField-2



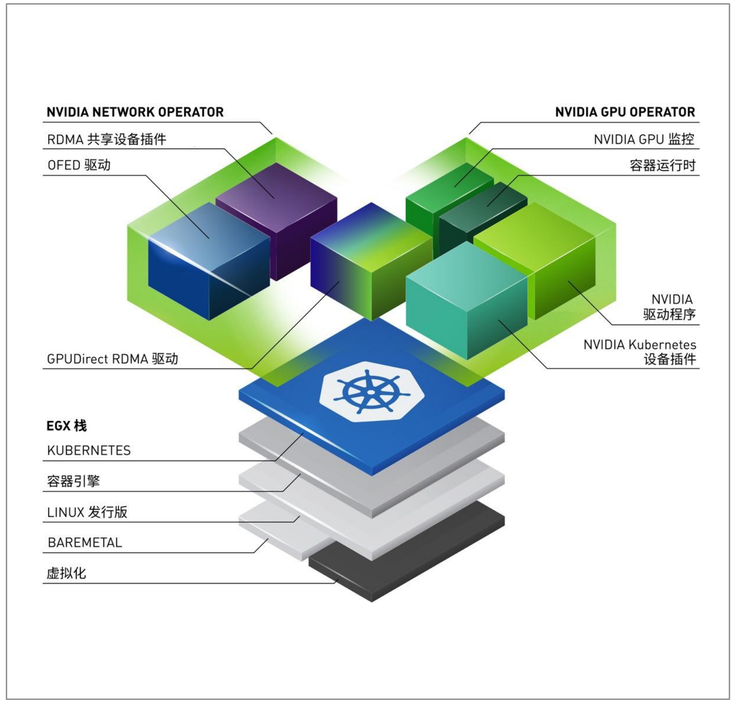
虚拟化容器编排和管理集成

实现统一的加速基础架构



虚拟化软件栈

集成了 VMware、Red Hat、Citrix、Nutanix 的虚拟基础架构



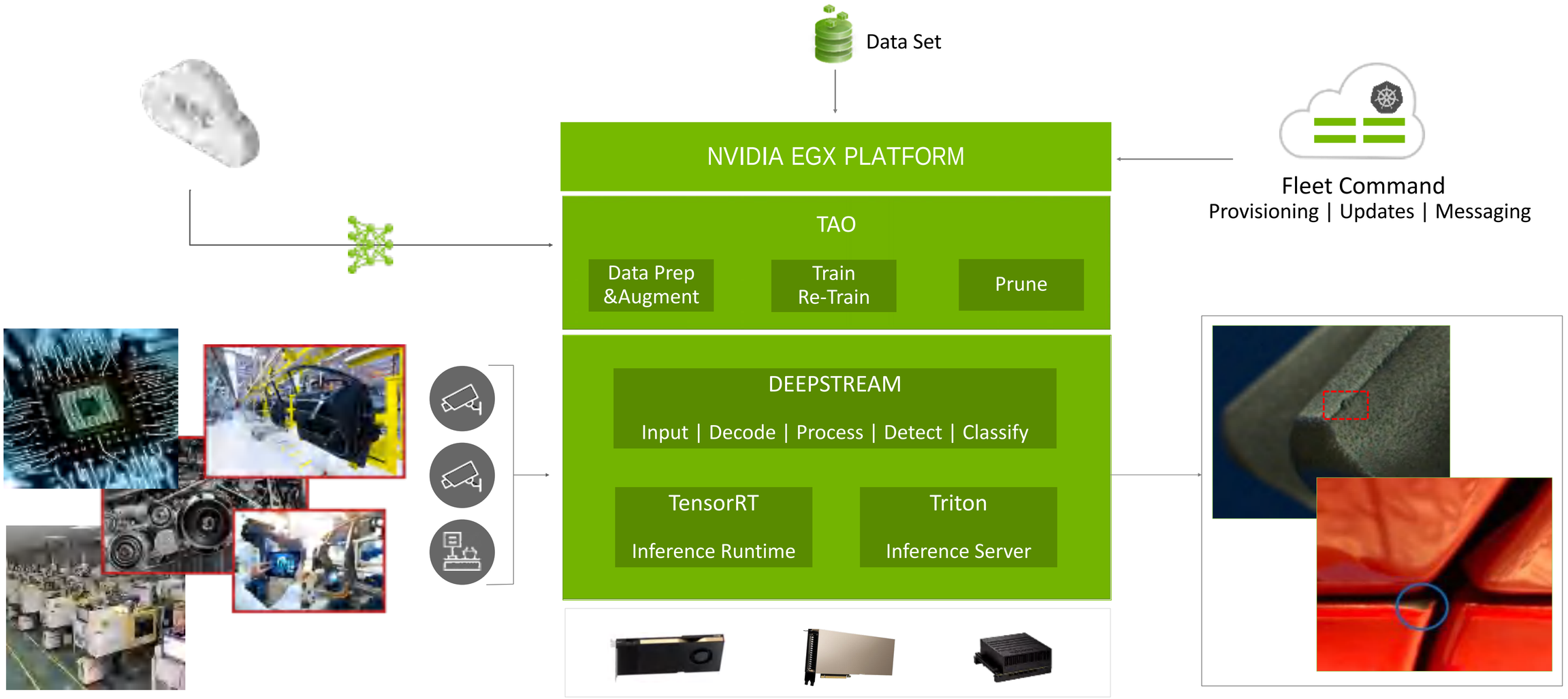
容器软件栈

适用于云原生工作负载的软件

NVIDIA Fleet Command

远程应用部署和管理

MANUFACTURE INSPECTION REFERENCE DESIGN



NVIDIA EGX PLATFORM EMPOWER AI WORKFLOWS

Fast-Track & End to End AI Application Development

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

2 Quickly train, adapt, and optimize models to your unique application

3 Integrate your customized models into your application and deploy

4 Scale out and deployment your mode in EGX or SoC platform

Start with NVIDIA-optimized Model Architecture

Image Classification
Object Detection
Segmentation
EfficientNet ResNet
RetinaNet YOLOV3/V4
UNET MaskRCNN

OR

Start with NVIDIA pre-trained Models

People Detection
Gaze
2D/3D Pose Estimation
Vehicle Classification
ALPR
Facial Landmarks
Gestures
ASR
Text Recognition

Your Data

TAO TOOLKIT

Train
Adapt
Optimize

Workstations
Cloud
DGX

Your Production Model

MANY INDUSTRIES

Deployment Frameworks

DeepStream
RIVA
Triton

Edge to Cloud

Deployment Model

SCALE OUT WITH FLEET COMMAND

EGX or SoC

* Choose from over 100+ model combinations on NGC

NVIDIA EGX PLATFORM EMPOWER AI WORKFLOWS

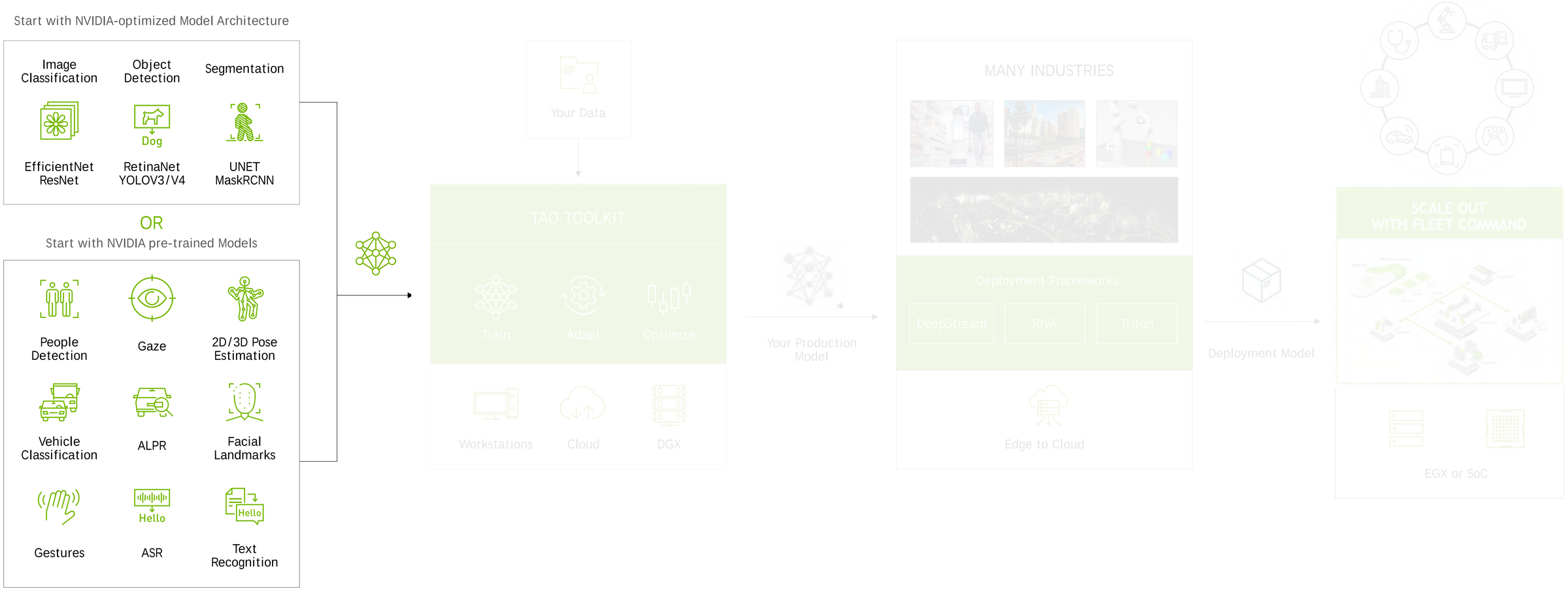
Fast-Track & End to End AI Application Development

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

2 Quickly train, adapt, and optimize models to your unique application

3 Integrate your customized models into your application and deploy

4 Scale out and deployment your mode in EGX or SoC platform



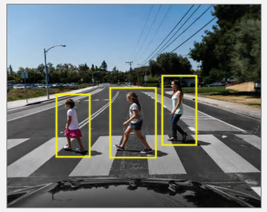
* Choose from over 100+ model combinations on NGC

HIGH PERFORMANCE PRE-TRAINED AI MODELS

Free download from – http://ngc.nvidia.com

Computing Vision (CV)

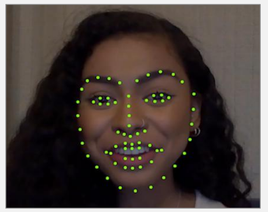
Optimized for high throughput
Trained for >80% accuracy


People Detection


People Segmentation

Face Detect IR

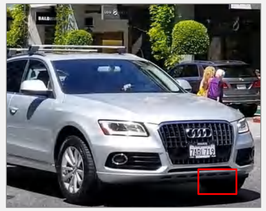

Gaze Estimation


Facial Landmark

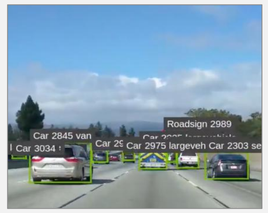

Pose Estimation


Dash Camera Vehicle Detection

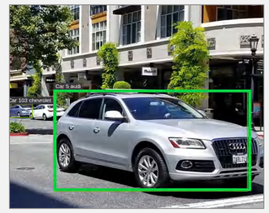

Vehicle & Pedestrian Detection


License Plate Detection

License Plate Recognition



Vehicle Type Net


Vehicle Make Net

Automatic Speech Recognition (ASR)

- Jasper
- QuartzNet
- CitriNet
- N-Gram

Natural Language Processing (NLP)

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

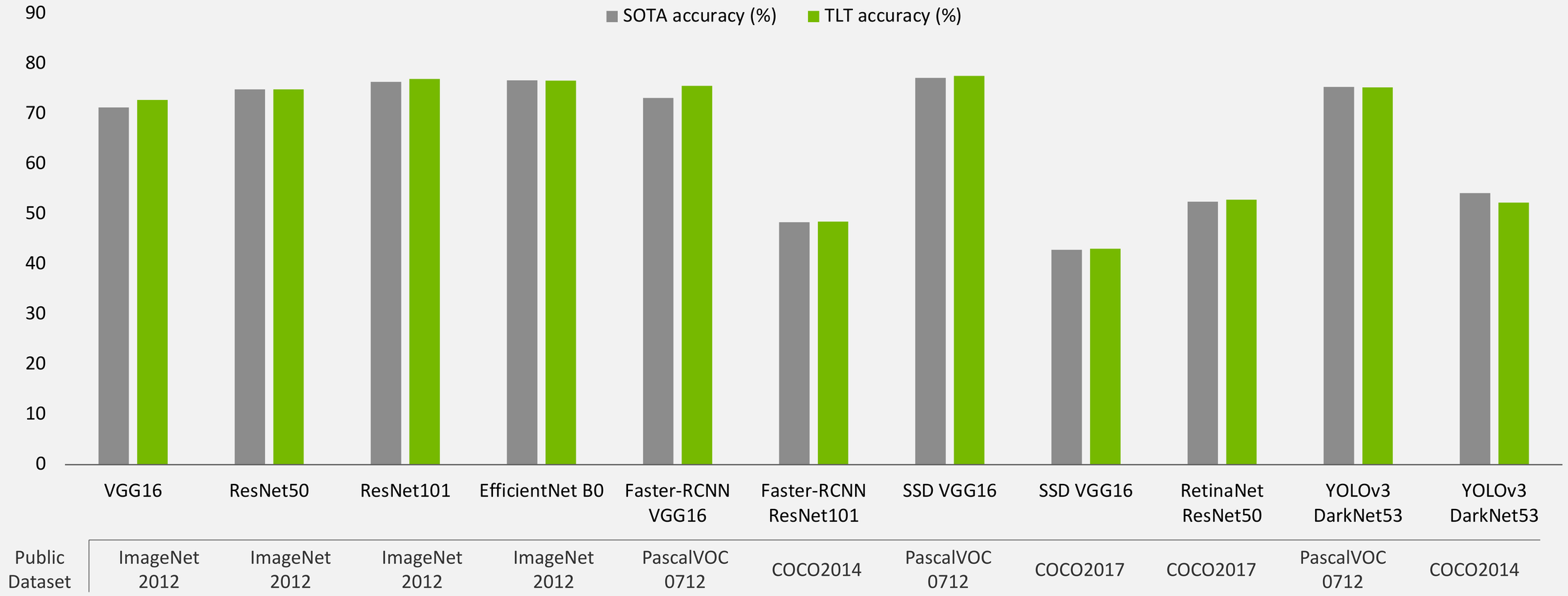
ENABLING BEYOND PRE-TRAINED AI MODELS

100+ Combination of Model Architectures and Backbones

	Image Classification	Object Detection							Segmentation	
		DetectNet_V2	FasterRCNN	SSD	YOLOV3	YOLOV4	RetinaNet	DSSD	MaskRCNN	UNET
ResNet10/18/ 34/50/101	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
VGG16/19	✓	✓	✓	✓	✓	✓	✓	✓		✓
GoogLeNet	✓	✓	✓	✓	✓	✓	✓	✓		
MobileNet V1/V2	✓	✓	✓	✓	✓	✓	✓	✓		
SqueezeNet	✓	✓		✓	✓	✓	✓	✓		
DarkNet 19/53	✓	✓	✓	✓	✓	✓	✓	✓		
CSPDarkNet 19/53	✓					✓				
EfficientNet B0/B1	✓		✓	✓			✓	✓		

Pre-trained weights trained on OpenImage dataset

ACHIEVING STATE OF THE ART ACCURACY FOR PUBLIC DATASETS



Developer blog - Achieving SOTA accuracy

NVIDIA EGX PLATFORM EMPOWER AI WORKFLOWS

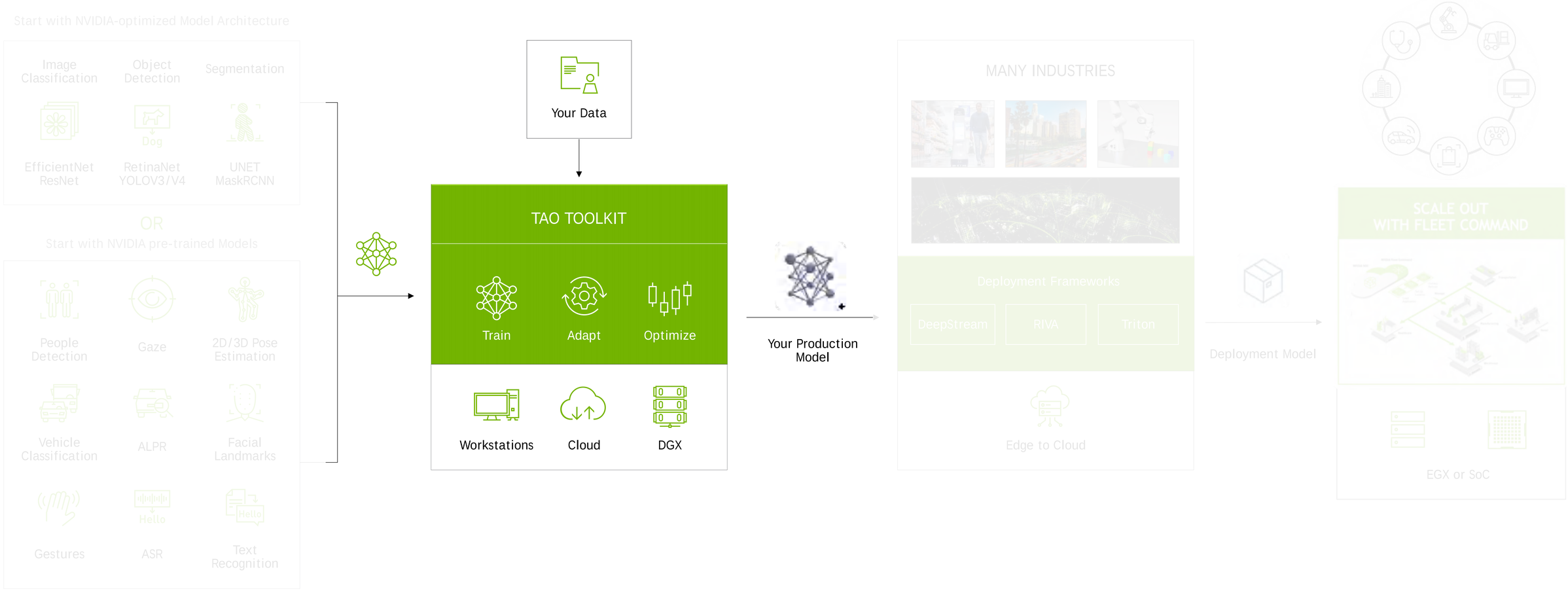
Fast-Track & End to End AI Application Development

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

2 Quickly train, adapt, and optimize models to your unique application

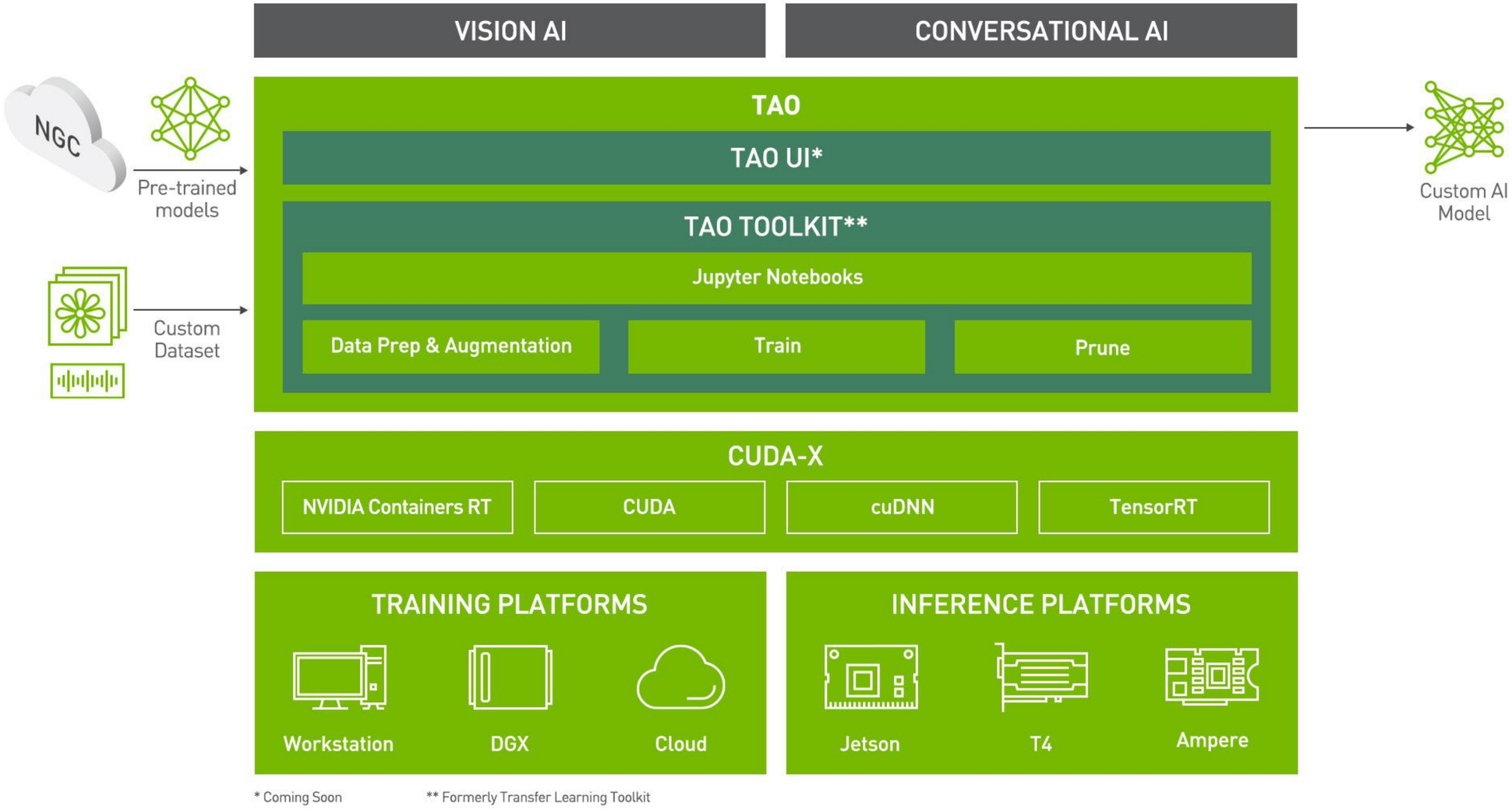
3 Integrate your customized models into your application and deploy

4 Scale out and deployment your mode in EGX or SoC platform



* Choose from over 100+ model combinations on NGC

NVIDIA TAO SOLUTION STACK

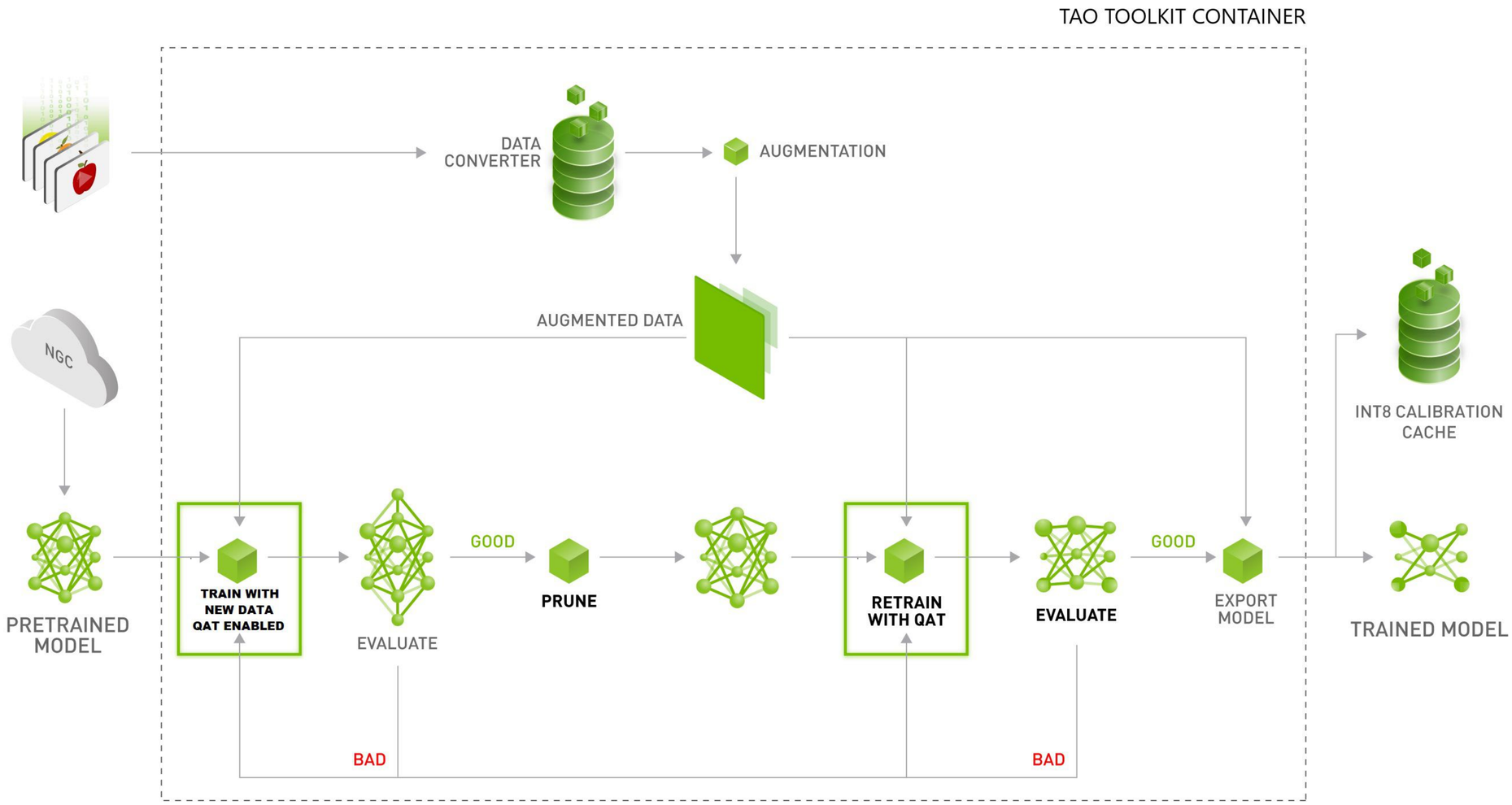


* Coming Soon

** Formerly Transfer Learning Toolkit

TAO TOOLKIT WORKFLOW

Automatic Mixed Precision | Quantization Aware Training | Pruning



https://developer.nvidia.com/tao-toolkit

NVIDIA EGX PLATFORM EMPOWER AI WORKFLOWS

Fast-Track & End to End AI Application Development

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

2 Quickly train, adapt, and optimize models to your unique application

3 Integrate your customized models into your application and deploy

4 Scale out and deployment your mode in EGX or SoC platform

Start with NVIDIA-optimized Model Architecture

Image Classification | Object Detection | Segmentation

EfficientNet ResNet | RetinaNet YOLOV3/V4 | UNET MaskRCNN

OR

Start with NVIDIA pre-trained Models

People Detection | Gaze | 2D/3D Pose Estimation

Vehicle Classification | ALPR | Facial Landmarks

Gestures | ASR | Text Recognition

* Choose from over 100+ model combinations on NGC

Your Data

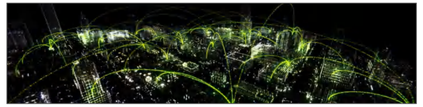
TAO TOOLKIT

Train | Adapt | Optimize

Workstations | Cloud | DGX

Your Production Model

MANY INDUSTRIES



Deployment Frameworks

DeepStream | RIVA | Triton

Edge to Cloud

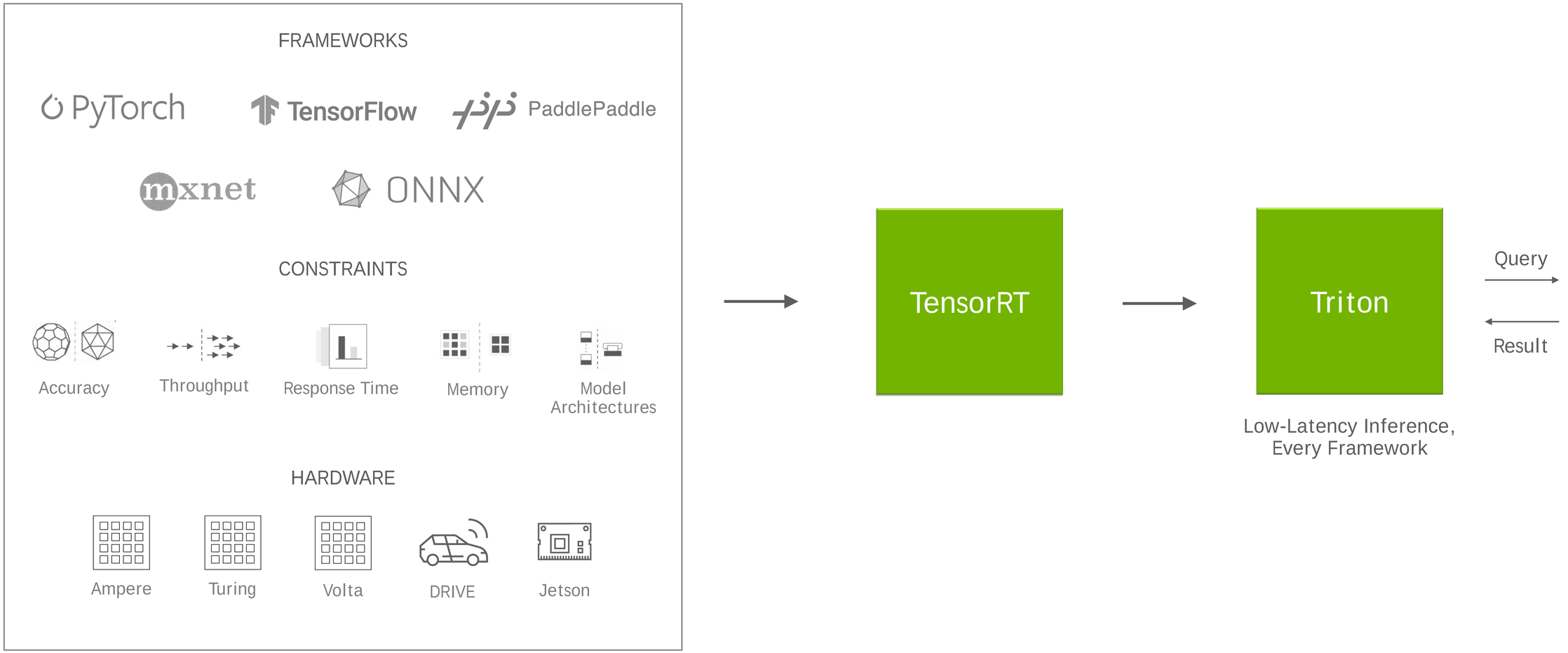
Deployment Mode

SCALE OUT WITH FLEET COMMAND

EGX or SoC

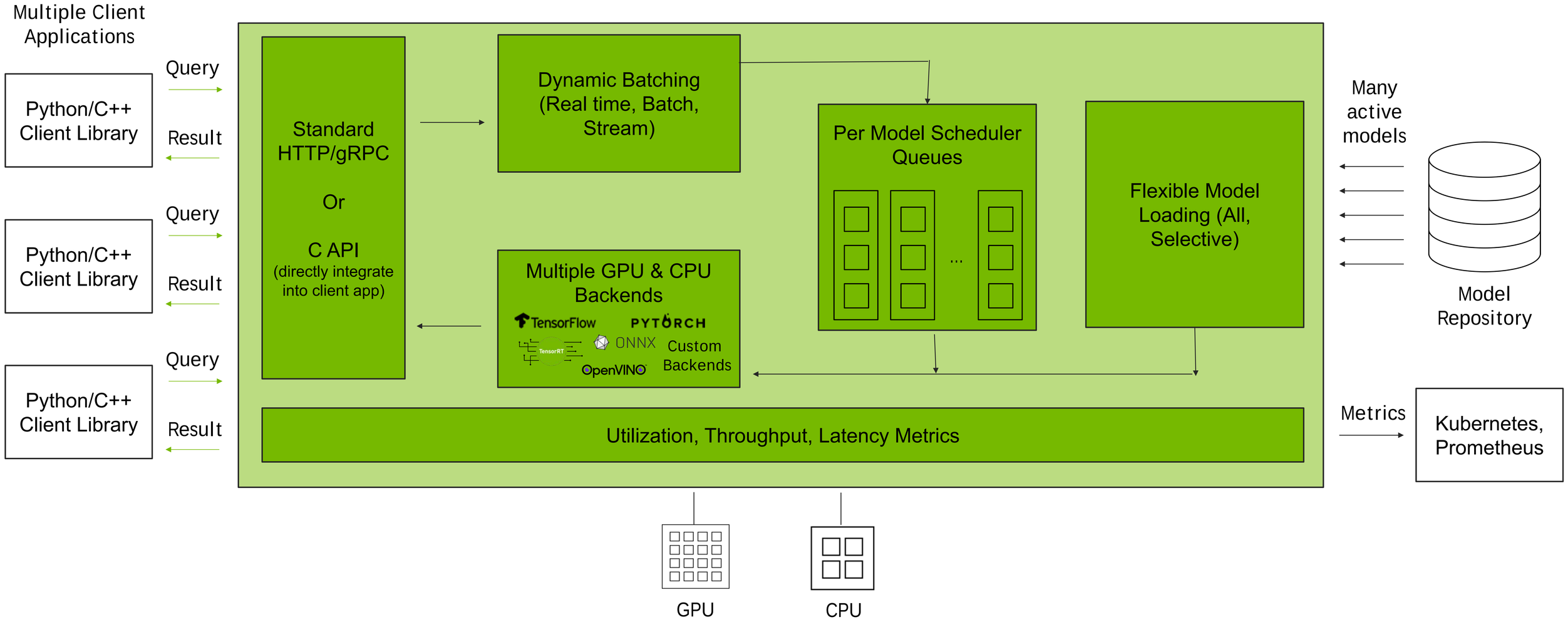
INFERENCE IS COMPLEX

Real Time | Competing Constraints | Rapid Updates



TRITON INFERENCE SERVER

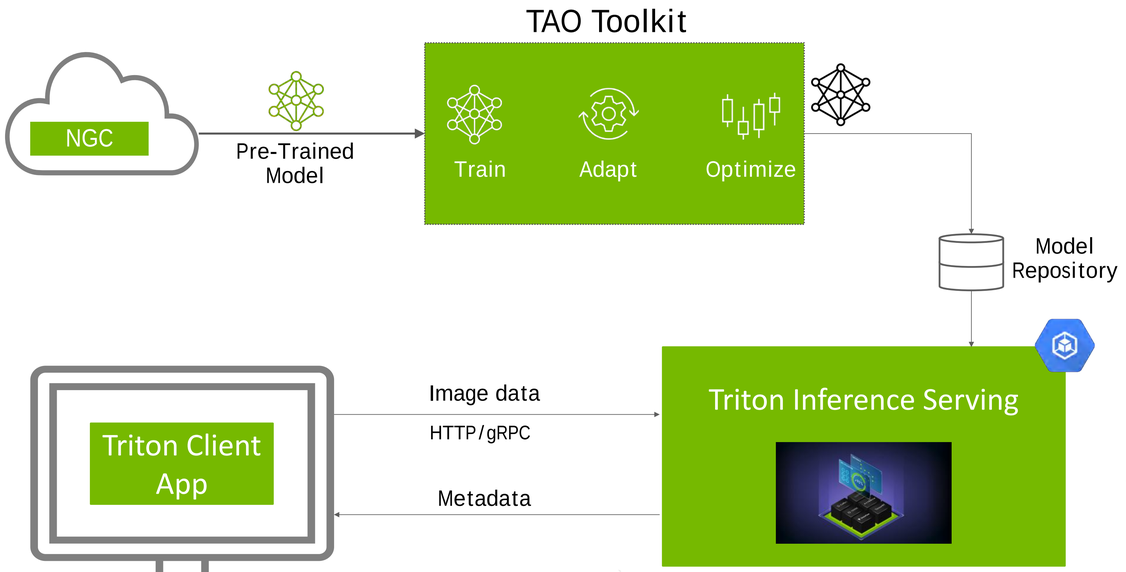
Open-Source Software For Scalable, Simplified Inference Serving

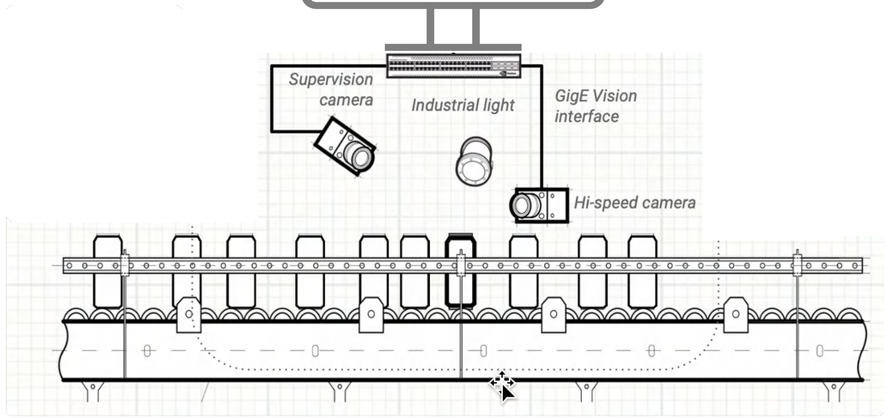


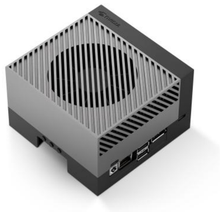
Monthly releases available on GitHub and on NGC as docker container

DEPLOY AI MODELS WITH TRITON

GitHub Project: https://github.com/NVIDIA-AI-IOT/tao-triton-apps

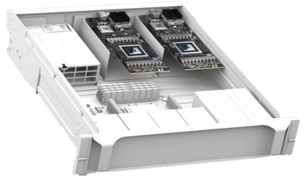






EGX - Jetson SoC

- Realtime inference (ms)
- 1:N deployment (n<4)
- Light weight model
- Advanced IPC



EGX - Server

- Batch inference
- 1:N deployment
- Complex model & pipeline
- NVIDIA Full Stack Certified

NVIDIA EGX PLATFORM EMPOWER AI WORKFLOWS

Fast-Track & End to End AI Application Development

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

2 Quickly train, adapt, and optimize models to your unique application

3 Integrate your customized models into your application and deploy

4 Scale out and deployment your mode in EGX or SoC platform

Start with NVIDIA-optimized Model Architecture

Image Classification
Object Detection
Segmentation
EfficientNet ResNet
RetinaNet YOLOV3/V4
UNET MaskRCNN

OR

Start with NVIDIA pre-trained Models

People Detection
Gaze
2D/3D Pose Estimation
Vehicle Classification
ALPR
Facial Landmarks
Gestures
ASR
Text Recognition

* Choose from over 100+ model combinations on NGC

Your Data

TAO TOOLKIT

Train
Adapt
Optimize

Workstations
Cloud
DGX

Your Production Model

MANY INDUSTRIES

Deployment Frameworks

DeepStream
RIVA
Triton

Edge to Cloud

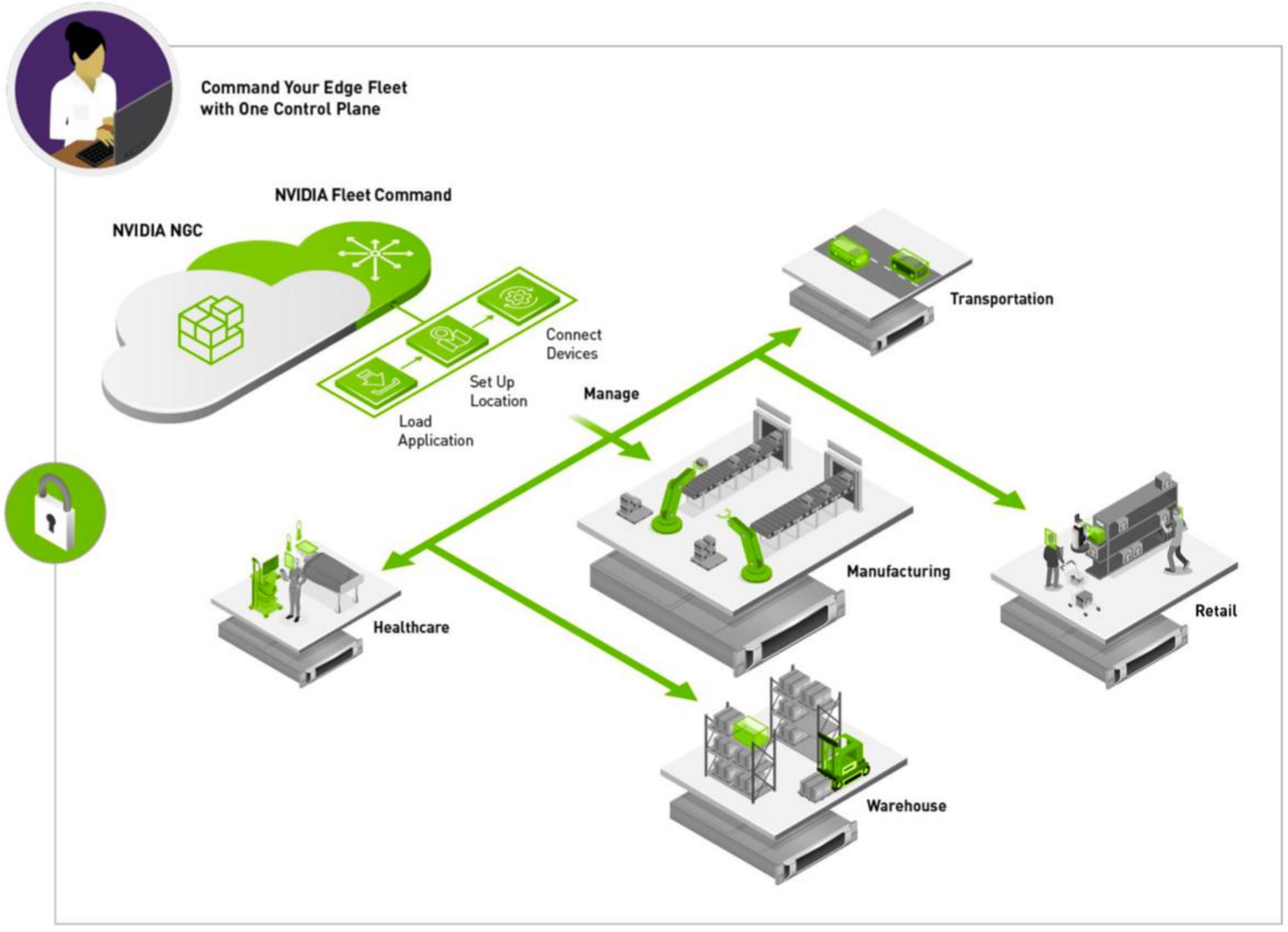
Deployment Model

SCALE OUT WITH FLEET COMMAND

EGX or SoC

NVIDIA FLEET COMMAND

Central, Secure Management for Every Node, Every Site



Cloud Service

One-Click Deployment

Open APIs

Layered Security

Monitoring & Alerting

Unified Dashboard

Resilient Architecture

NVIDIA EGX PLATFORM EMPOWER AI WORKFLOWS

Fast-Track & End to End AI Application Development

1 Choose from NVIDIA's Library of Pre-trained Models OR Model Architectures

2 Quickly train, adapt, and optimize models to your unique application

3 Integrate your customized models into your application and deploy

4 Scale out and deployment your mode in EGX or SoC platform

Start with NVIDIA-optimized Model Architecture

Image Classification
Object Detection
Segmentation
EfficientNet ResNet
RetinaNet YOLOV3/V4
UNET MaskRCNN

OR

Start with NVIDIA pre-trained Models

People Detection
Gaze
2D/3D Pose Estimation
Vehicle Classification
ALPR
Facial Landmarks
Gestures
ASR
Text Recognition

Your Data

TAO TOOLKIT

Train
Adapt
Optimize

Workstations
Cloud
DGX

Your Production Model

MANY INDUSTRIES

Deployment Frameworks

DeepStream
RIVA
Triton

Edge to Cloud

Deployment Model

SCALE OUT WITH FLEET COMMAND

EGX or SoC

* Choose from over 100+ model combinations on NGC

NVIDIA EGX PLATFORM SOFTWARE

End-to-End AI SDK and Software

Application Iteration - AIOPS

MODEL & DATA PREP	TRAIN AT SCALE	OPTIMIZED APPLICATIONS	DEPLOY AT SCALE
			
NGC	TAO	SDK	Fleet Command
Pre-Trained Models General-Purposed Models Data Set	Train Adapt Optimize	Application SDK • DeepStream • Riva • Triton Simulation • ISAAC	Cloud Native Cloud to Edge Scale-Out Deployment

NVIDIA EGX Platform Software Toolkit

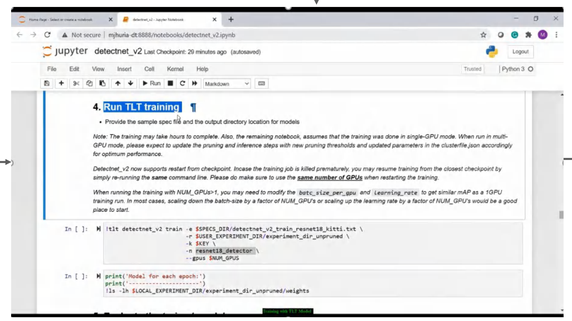
AUTOMATE OPTICAL INSPECTION

End-to-end workflow

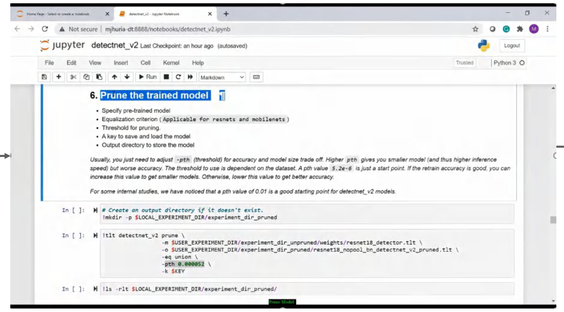
Dataset

Continuous Learning using data from factory floor

Select UNET pre-trained model from NGC

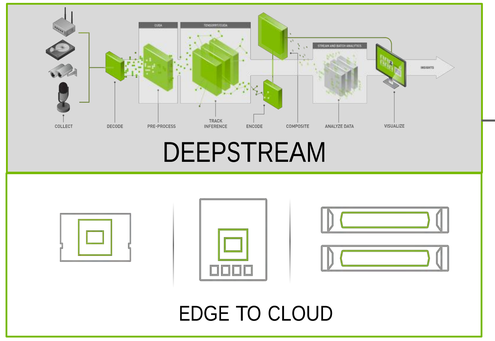


Re-train on the TAO Toolkit with site specific data

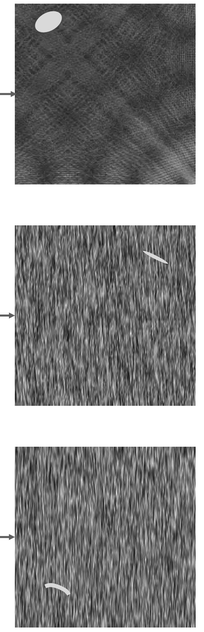


Prune and optimize

Deploy



Deploy inference at the edge



Defect detected in real-time

< 1 day

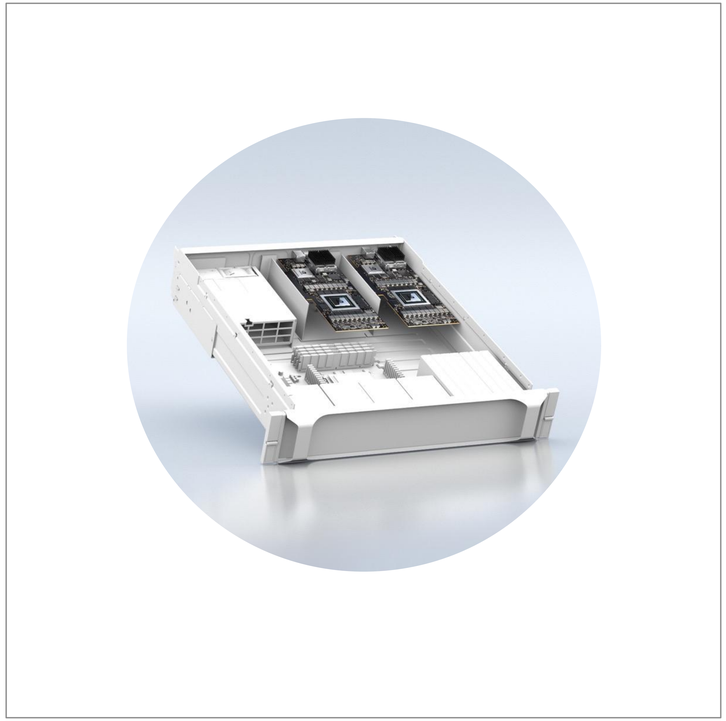
https://github.com/NVIDIA-AI-IOT/deepstream-segmentation-analytics

NVIDIA INDUSTRIAL SCALE-AI

GPU-accelerated computing enables AI for
industrial manufacturing applications across industries
with predictive maintenance, industrial inspection, and robotics.

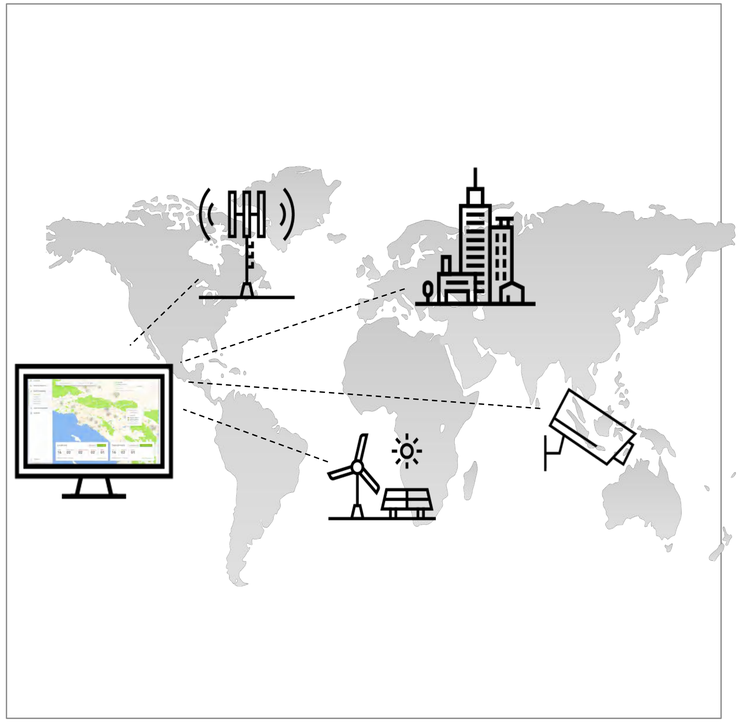
NVIDIA EGX - 企业边缘加速计算平台

平台的优势



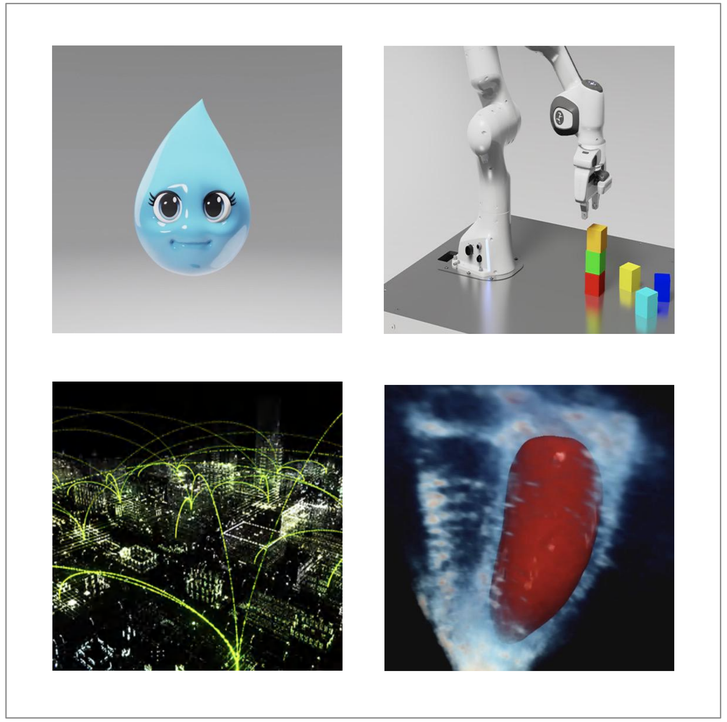
NVIDIA 认证系统

高带宽、低延迟
节能高效



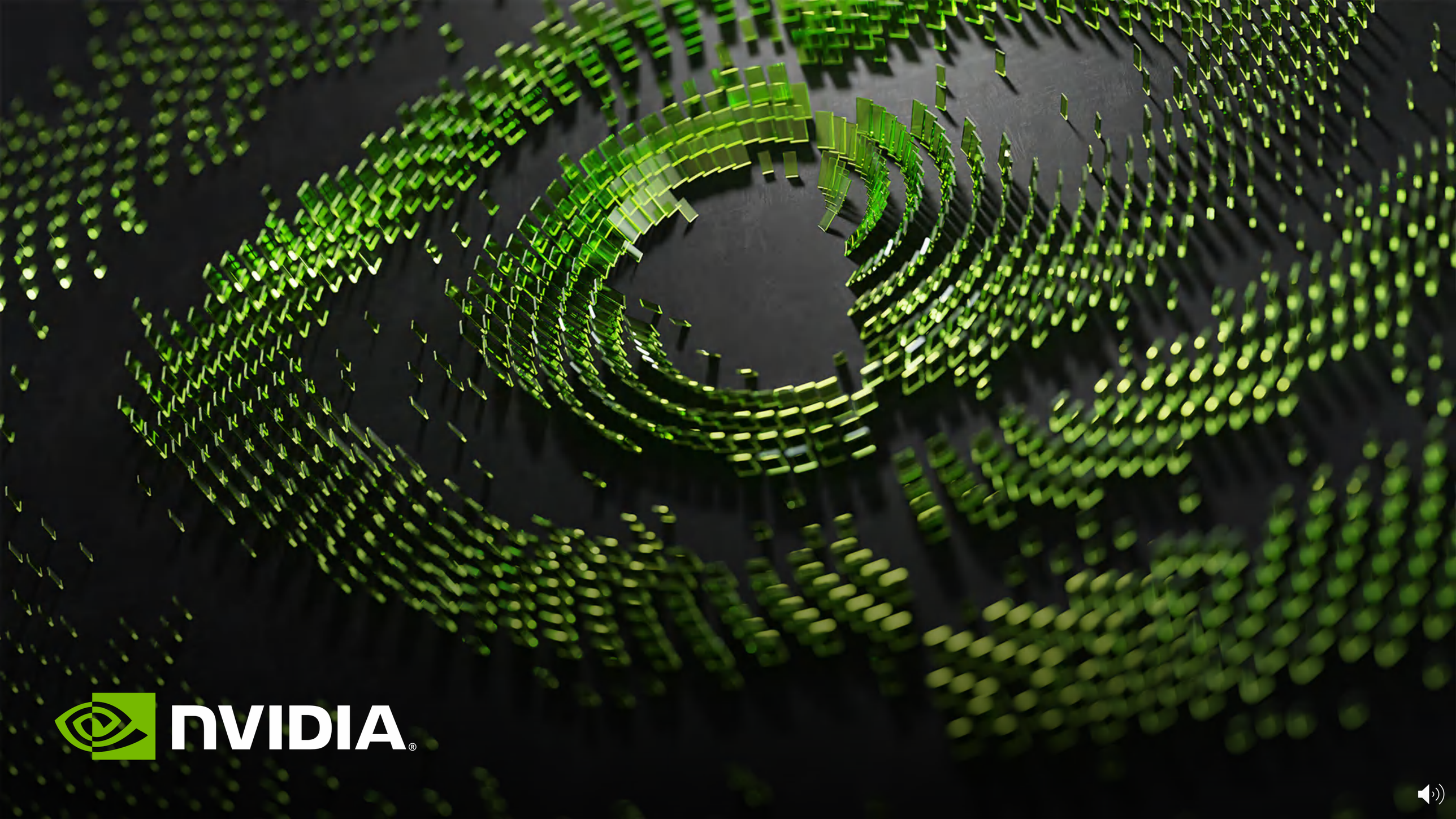
适用于远程应用部署的 Fleet Command

集中管理边缘设备
全栈端到端安全性
监控和远程修复系统



NVIDIA 软件开发包

加速应用开发
AI 应用构建块
适用于各行各业的工具


NVIDIA