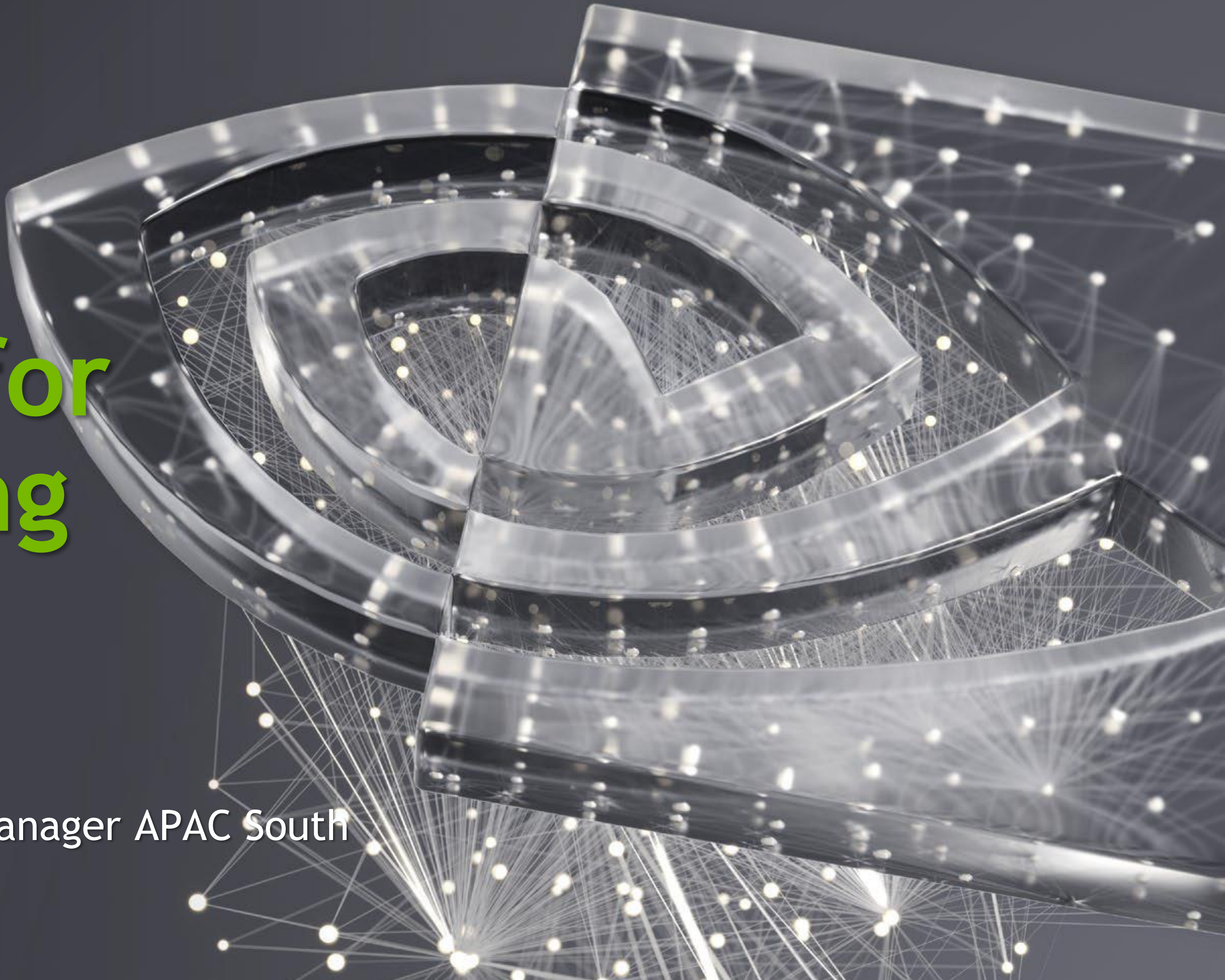
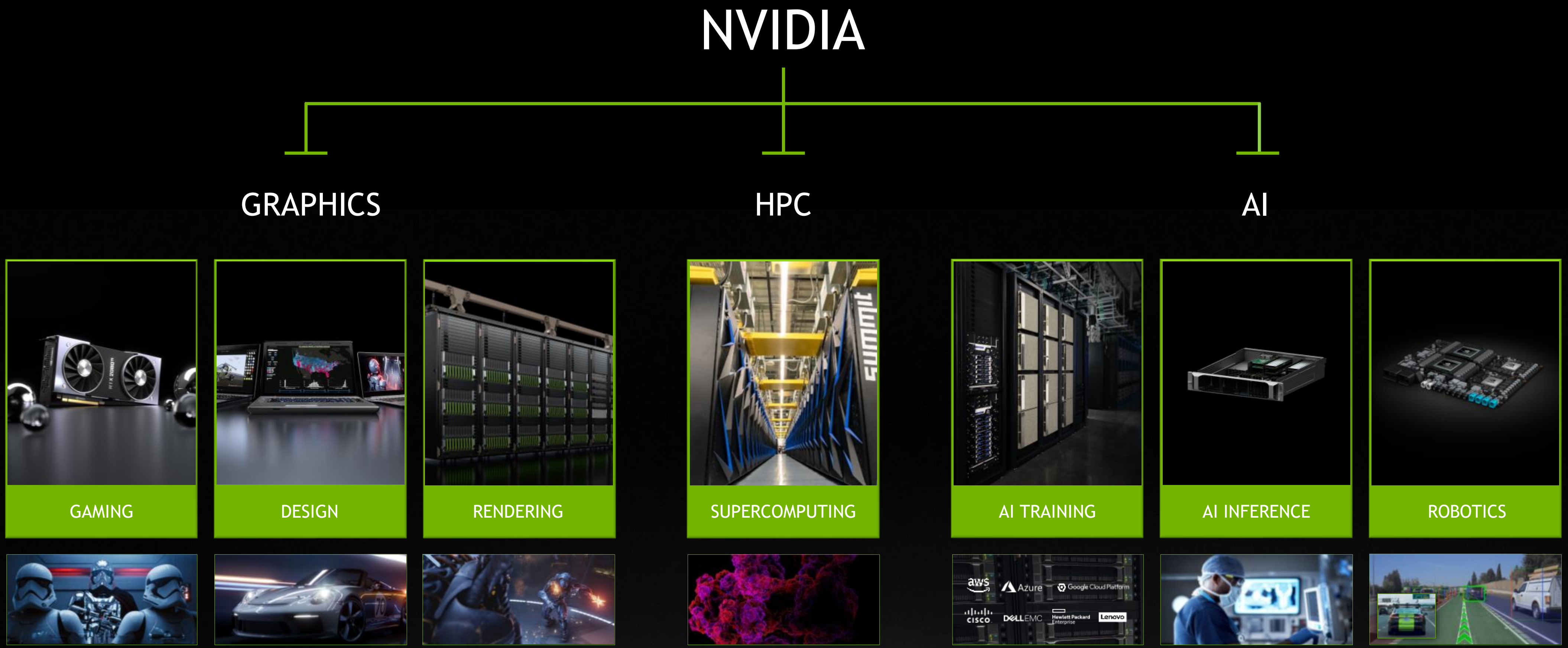
EGX Platform for edge computing



Michael Lang – Solutions Architecture Manager APAC South

September 2019



NVIDIA

GRAPHICS

HPC

AI

GAMING

DESIGN

RENDERING

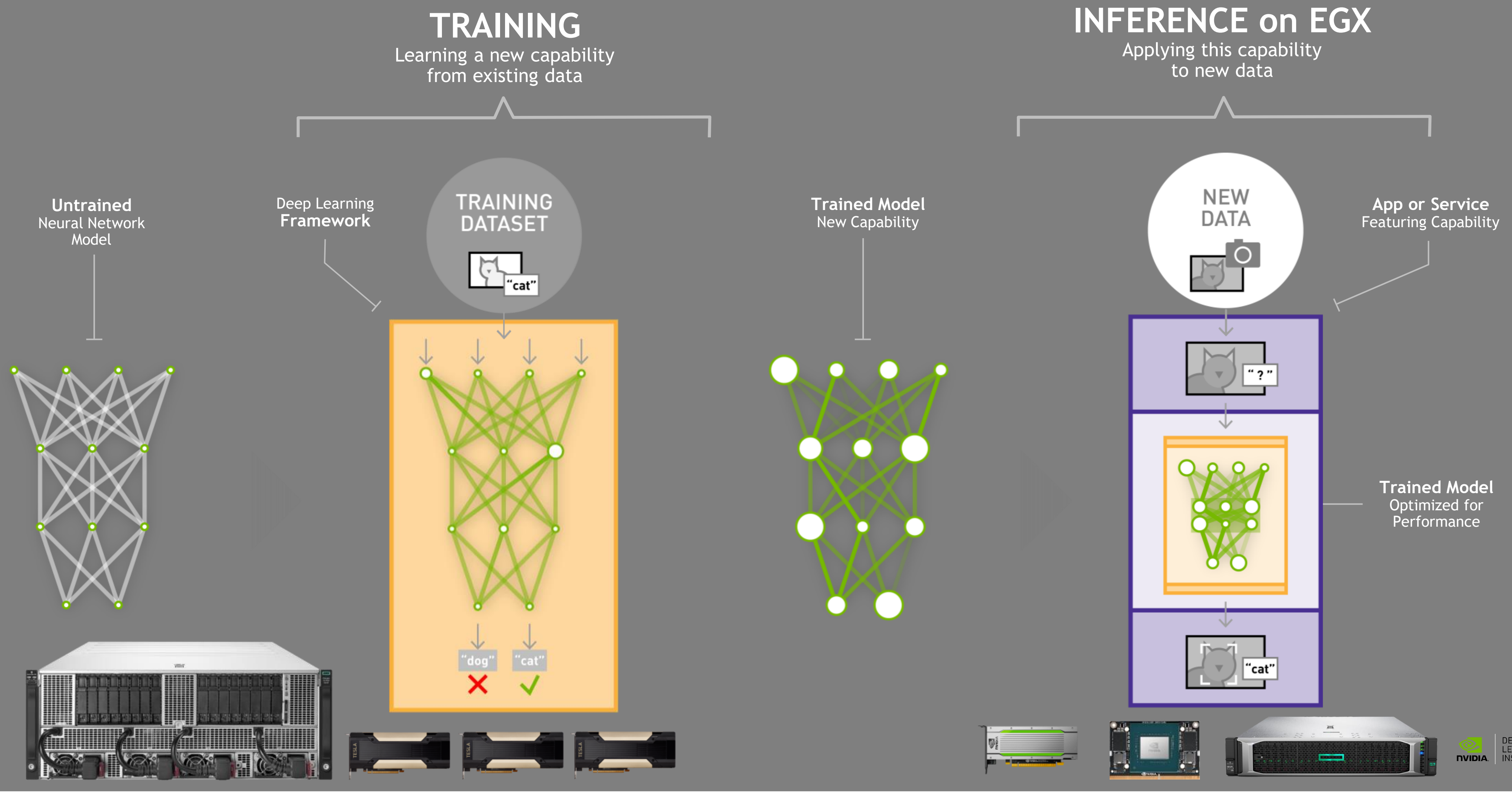
SUPERCOMPUTING

AI TRAINING

AI INFERENCE

ROBOTICS

DEEP LEARNING APPLICATION DEVELOPMENT



TRAINING
Learning a new capability
from existing data

INFERENCE on EGX
Applying this capability
to new data

Untrained
Neural Network
Model

Deep Learning
Framework

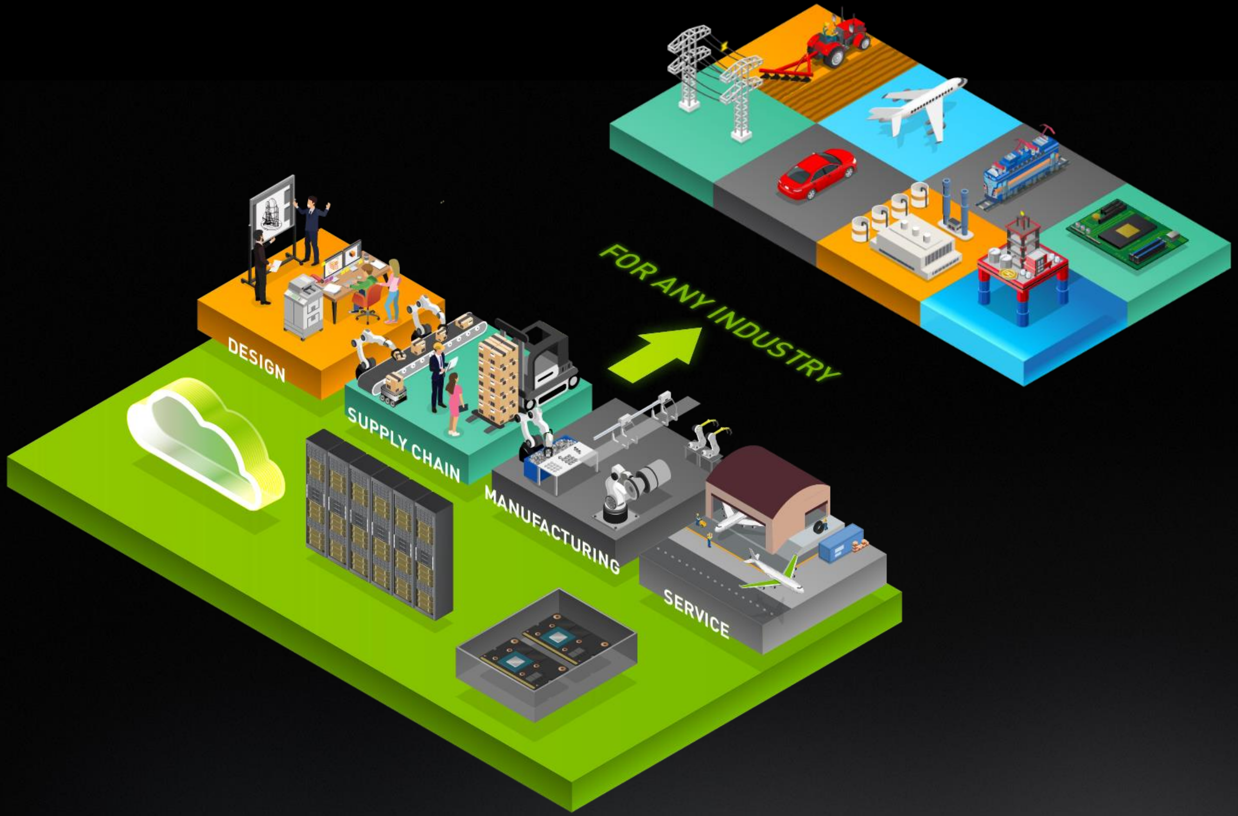
Trained Model
New Capability

App or Service
Featuring Capability

Trained Model
Optimized for
Performance

DEEP LEARNING INSTITUTE

ONE PLATFORM ACROSS ALL USE CASES



DESIGN

HPC
Modeling & Simulation
Design for Manufacturability

SUPPLY CHAIN

Forecasting & Inventory Management
Supply Chain Optimization
Robotics & Automation

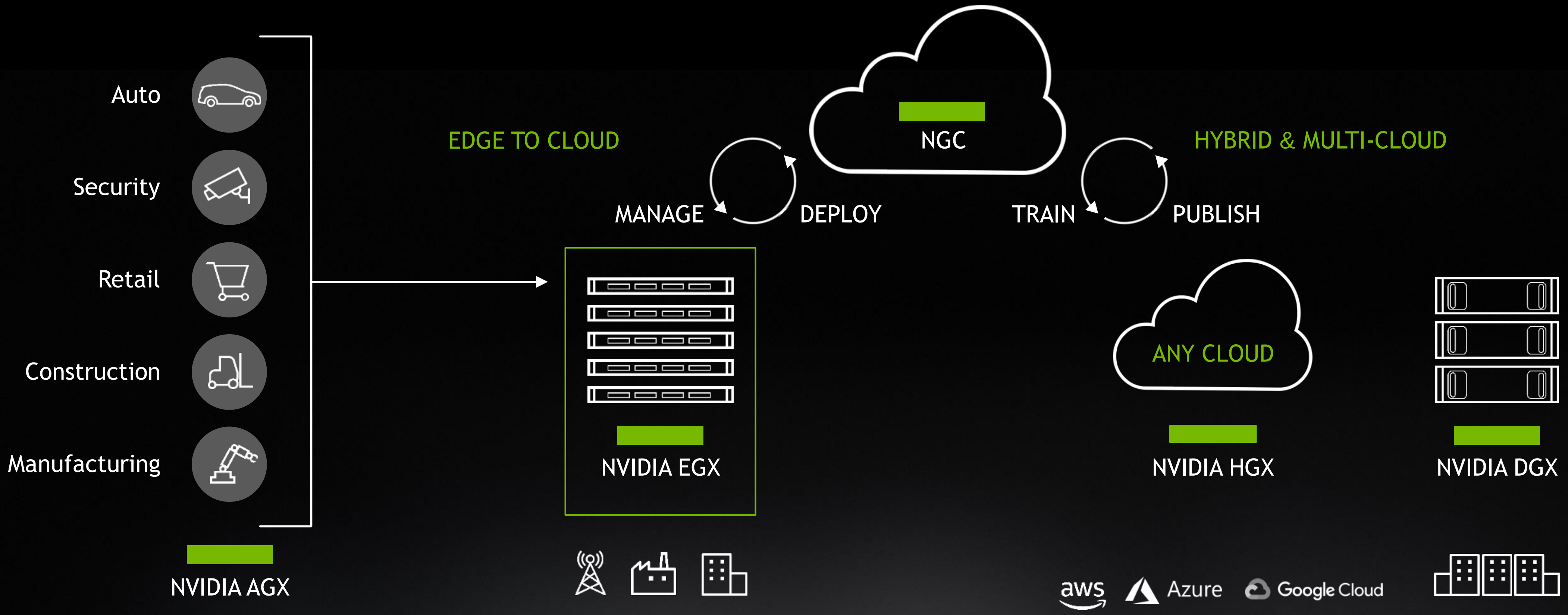
MANUFACTURE

Robotics & Automation
Inspection
Predictive Maintenance
Process Control

SERVICE

Predictive Maintenance
Field Inspection
Logistics Optimization
Parts Inventory Management

NVIDIA EGX EDGE COMPUTING



NVIDIA EGX EDGE COMPUTING

A new class of distributed AI computing systems designed to gather and analyze continuous streams of data at the edge of the network.

AI computation is performed largely or completely on the EGX systems close to the data or user.

NVIDIA EGX is for applications that require:

- Low-latency interactions
- Reduced bandwidth to the cloud
- Data privacy or sovereignty

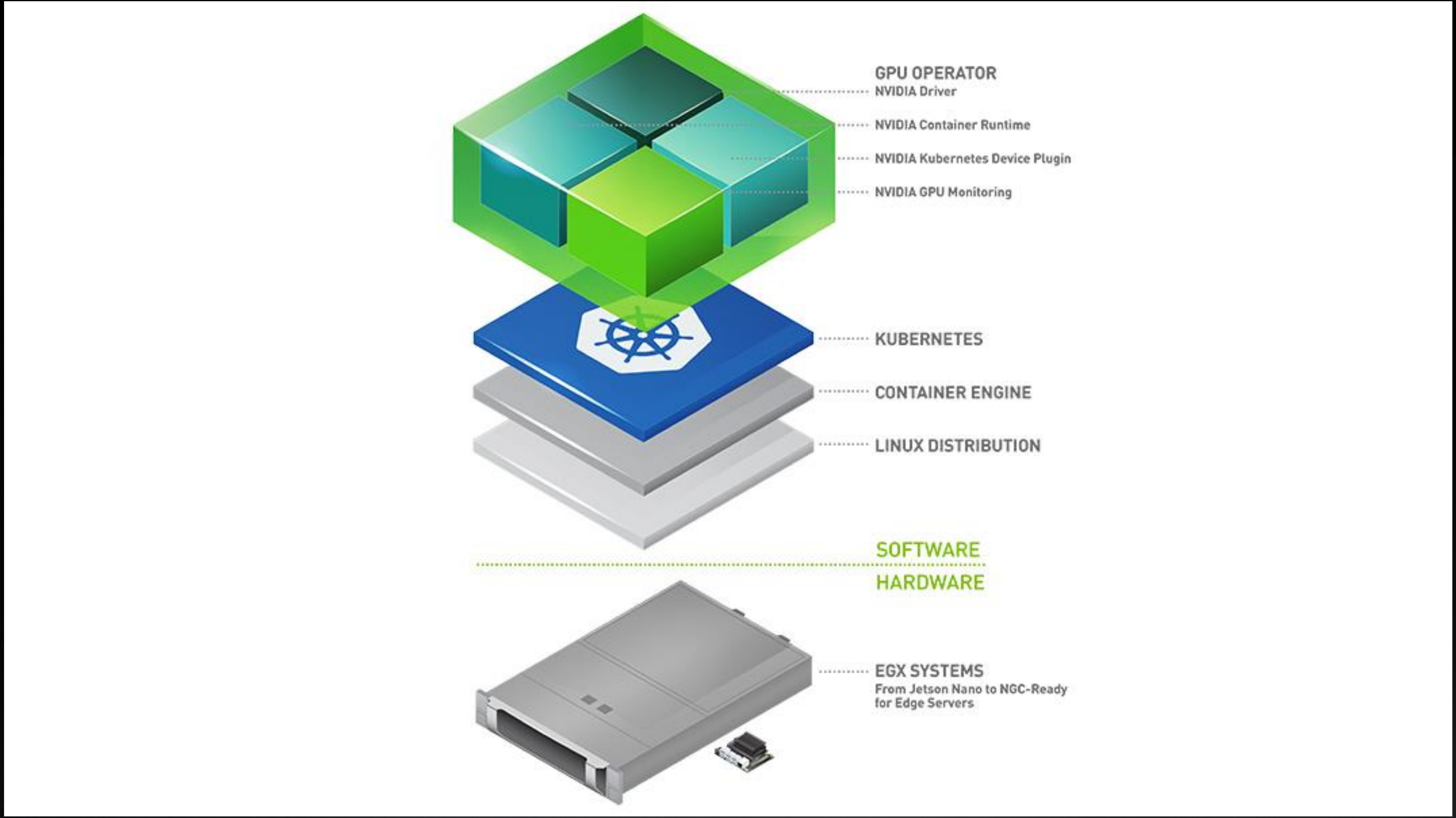
.5 TOPS

520 TOPS

10,000 TOPS

COMPUTE & AI BY NVIDIA
NETWORK, STORAGE, SECURITY BY MELLANOX

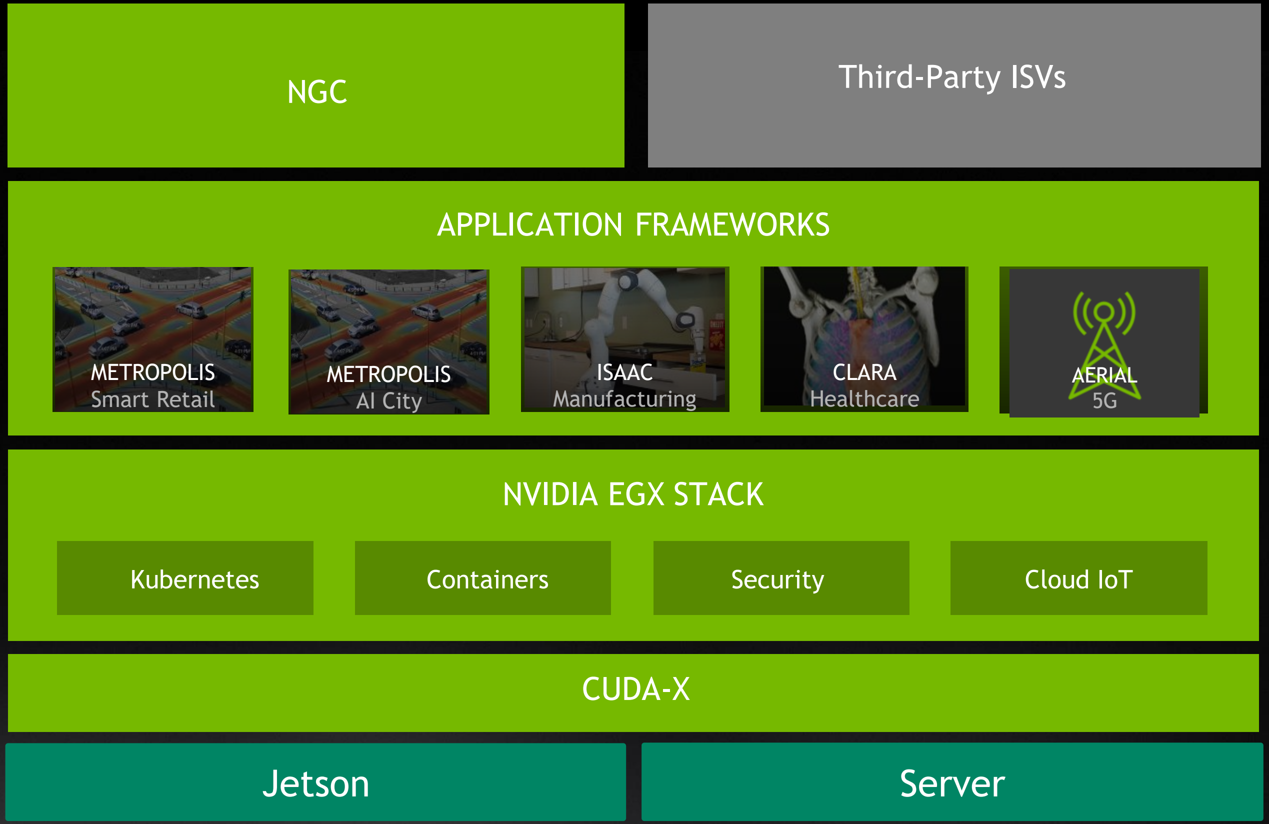
NVIDIA EGX EDGE COMPUTING - BREAKDOWN



NVIDIA EGX AI PLATFORM

Higher Performance Edge Computing Platform

- Powered by NVIDIA GPU
- Cloud-Native EGX Stack
- Vertical Industry SDKs
- Commercially off the shelf (COTS)
- Scale from 2W to 2 Petaflops



NGC

Third-Party ISVs

APPLICATION FRAMEWORKS

METROPOLIS Smart Retail

METROPOLIS AI City

ISAAC Manufacturing

CLARA Healthcare

AERIAL 5G

NVIDIA EGX STACK

Kubernetes

Containers

Security

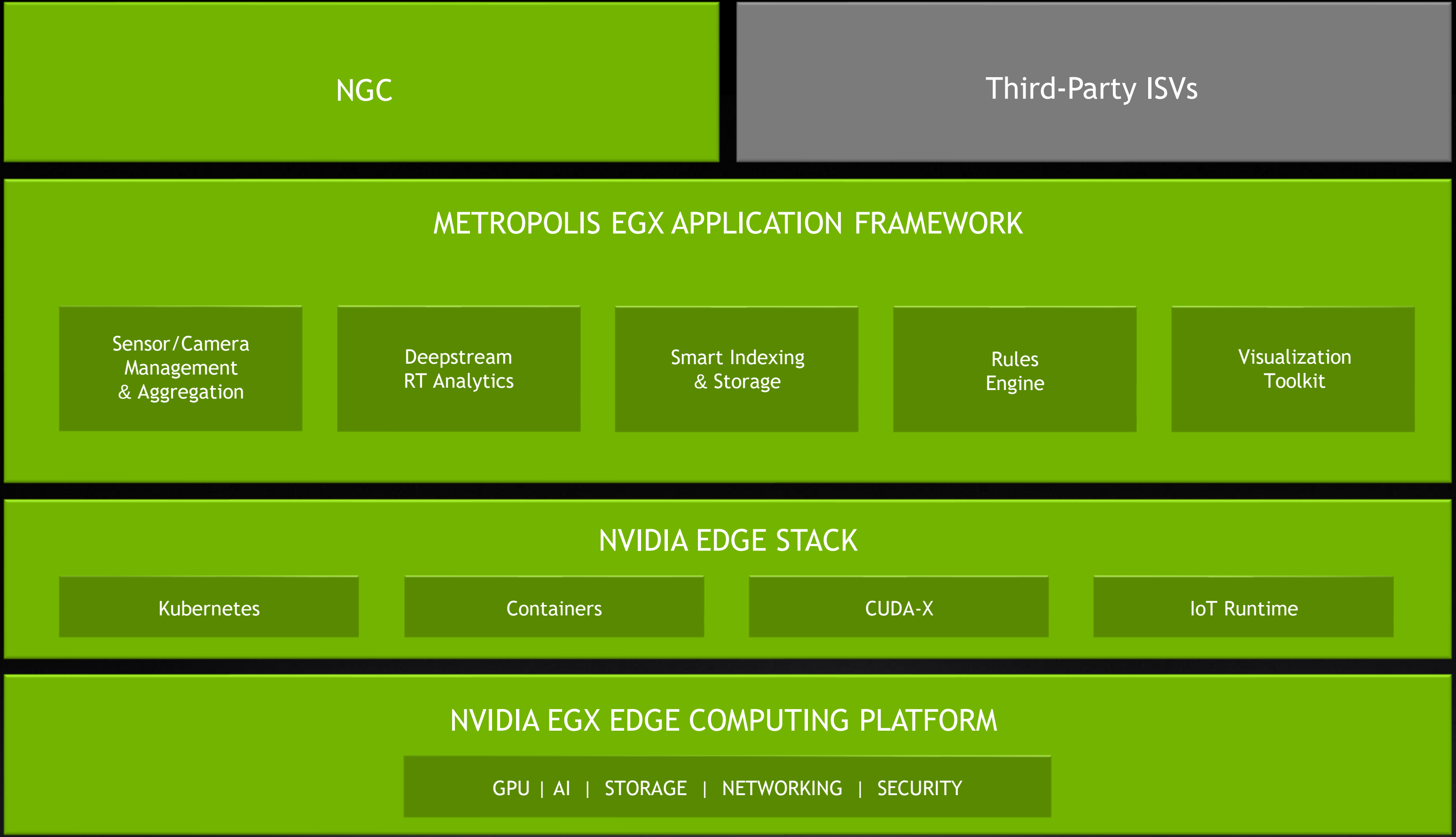
Cloud IoT

CUDA-X

Jetson

Server

METROPOLIS EGX OPEN AI RETAIL PLATFORM



NGC

Third-Party ISVs

METROPOLIS EGX APPLICATION FRAMEWORK

Sensor/Camera Management & Aggregation

Deepstream RT Analytics

Smart Indexing & Storage

Rules Engine

Visualization Toolkit

NVIDIA EDGE STACK

Kubernetes

Containers

CUDA-X

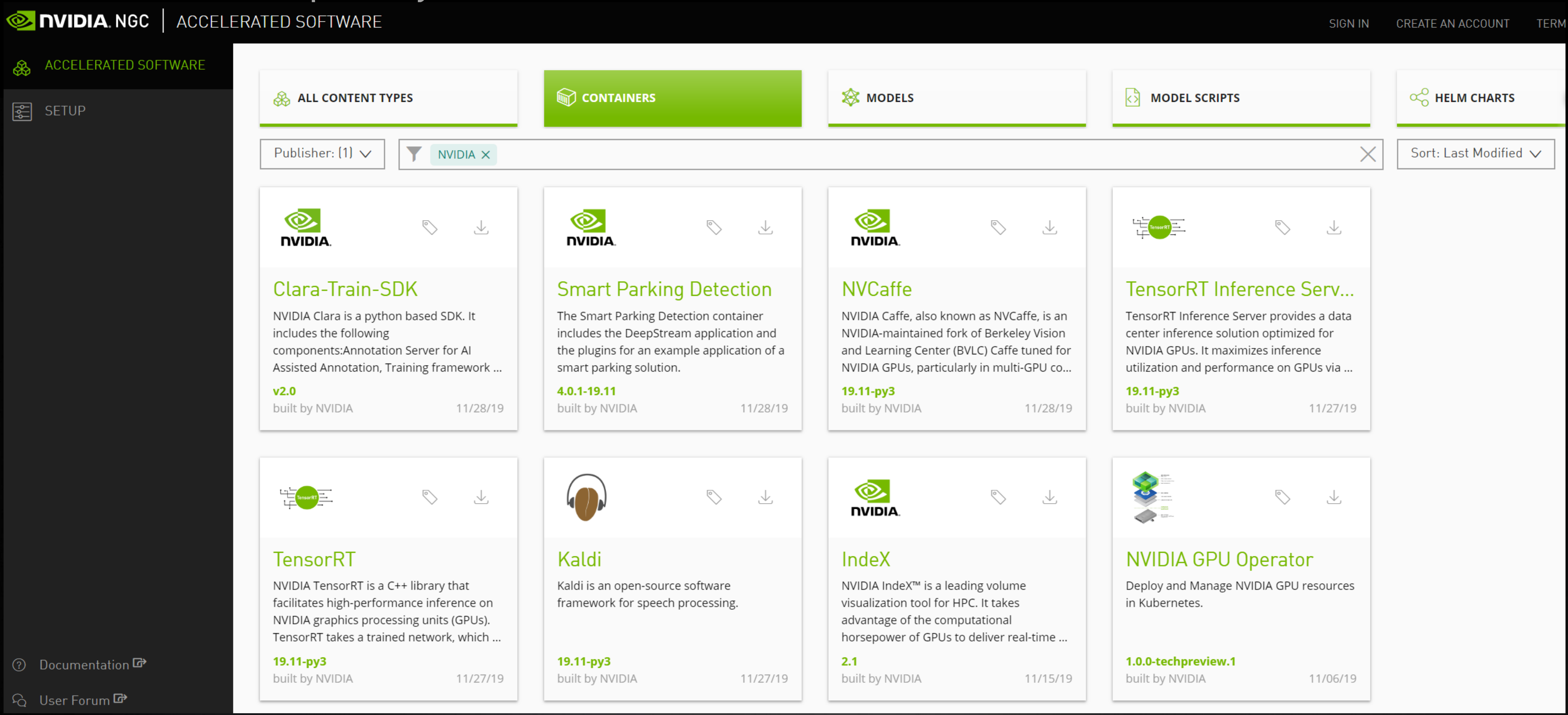
IoT Runtime

NVIDIA EGX EDGE COMPUTING PLATFORM

GPU | AI | STORAGE | NETWORKING | SECURITY

NVIDIA GPU CLOUD

Cloud repository of GPU enabled containers and frameworks NGC.NVIDIA.COM



NVIDIA NGC | ACCELERATED SOFTWARE

SIGN IN CREATE AN ACCOUNT TERM

ACCELERATED SOFTWARE

SETUP

ALL CONTENT TYPES | CONTAINERS | MODELS | MODEL SCRIPTS | HELM CHARTS

Publisher: (1) | NVIDIA ✕ | Sort: Last Modified

Clara-Train-SDK
NVIDIA Clara is a python based SDK. It includes the following components:Annotation Server for AI Assisted Annotation, Training framework ...
v2.0
built by NVIDIA 11/28/19

Smart Parking Detection
The Smart Parking Detection container includes the DeepStream application and the plugins for an example application of a smart parking solution.
4.0.1-19.11
built by NVIDIA 11/28/19

NVCaffe
NVIDIA Caffe, also known as NVCaffe, is an NVIDIA-maintained fork of Berkeley Vision and Learning Center (BVLC) Caffe tuned for NVIDIA GPUs, particularly in multi-GPU co...
19.11-py3
built by NVIDIA 11/28/19

TensorRT Inference Serv...
TensorRT Inference Server provides a data center inference solution optimized for NVIDIA GPUs. It maximizes inference utilization and performance on GPUs via ...
19.11-py3
built by NVIDIA 11/27/19

TensorRT
NVIDIA TensorRT is a C++ library that facilitates high-performance inference on NVIDIA graphics processing units (GPUs). TensorRT takes a trained network, which ...
19.11-py3
built by NVIDIA 11/27/19

Kaldi
Kaldi is an open-source software framework for speech processing.
19.11-py3
built by NVIDIA 11/27/19

IndeX
NVIDIA IndeX™ is a leading volume visualization tool for HPC. It takes advantage of the computational horsepower of GPUs to deliver real-time ...
2.1
built by NVIDIA 11/15/19

NVIDIA GPU Operator
Deploy and Manage NVIDIA GPU resources in Kubernetes.
1.0.0-techpreview.1
built by NVIDIA 11/06/19

Documentation

User Forum

SMALL, MEDIUM AND LARGE

JETSON AT THE EDGE

Jetson comes in a variety of sizes and carrier units for different use cases.

MICRO SERVERS

Small rack based servers that can support 1 – 4 Tesla T4 cards

HEAVY LIFTING SERVERS

Rack based 2RU servers that can support up to 7 x Telsa T4 cards for scale out sizing.

EGX HARDWARE DESIGN CONSIDERATIONS

EGX in the DC or at the Edge?

Latency considerations

High throughput & low latency

One architecture scalable from device to cloud

Sizing for Video streams

How many streams?

Resolution, 720P/2K/4K

FPS

What protocol(s)? H265/5?

Sizing for AI models

How many models?

How many FPS?

How complex is the model

Scale up vs Scale out

More smaller devices or fewer larger ones

Power considerations

Storage

MetaData Considerations

Data vs Metadata

Which one goes where

Different platforms?

Decentralised Data vs Centralized MD?

Maintenance and support

Who will support it and how

Remote updates & capability

Support contracts through OEM or bespoke?

Consumer vs Commercial HW

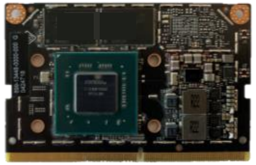
Ruggedized/NEBS or standard kit?

THE JETSON FAMILY

for AI at the Edge and Autonomous System designs

JETSON NANO

0.5 TFLOPS (FP16)



5 - 10W
45mm x 70mm

JETSON TX2 series

1.3 TFLOPS (FP16)

7.5 – 15W*
50mm x 87mm

JETSON Xavier NX

6 TFLOPS (FP16)
21 TOPS (INT8)

10 - 15W
45mm x 70mm

JETSON AGX XAVIER series

11 TFLOPS (FP16)
32 TOPS (INT8)

10 – 30W
100mm x 87mm

AI at the edge

Fully autonomous machines

Same software

Listed prices are for 1000u+ | Full specs at developer.nvidia.com/jetson

* TX2i: 10-20W

	JETSON NANO	JETSON TX2	JETSON XAVIER NX	JETSON AGX XAVIER
GPU	128 Core Maxwell 0.5 TFLOPs (FP16)	256 Core Pascal 1.3 TFLOPS (FP16)	384 Core Volta 21 TOPs (INT8)	512 Core Volta + NVDLA 10 TFLOPS (FP16) 32 TOPS (INT8)
CPU	4 core ARM A57	6 core Denver and A57 (2x) 2MB L2	6 core Carmel ARM CPU (3x) 2MB L2 + 4MB L3	8 core Carmel ARM CPU (4x) 2MB L2 + 4MB L3
Memory	4 GB 64-bit LPDDR4 25.6 GB/s	Up to 8 GB 128b LPDDR4 58 GB/s	8 GB 128-bit LPDDR4x 51.2 GB/s	Up to 16GB 256-bit LPDDR4x 137 GB/s
Storage	16 GB eMMC	Up to 32 GB eMMC	16 GB eMMC	32 GB eMMC
Encode	4K @ 30 (H.265)	4K @ 60 (H.265)	2x 4K @ 30 (H.265)	4x 4K @ 60 (H.265)
Decode	4K @ 60 (H.265)	2x 4K @ 60 (H.265)	2x 4K @ 60 (H.265)	6x 4K @ 60 (H.265)
Camera	12 (3x4 or 4x2) MIPI CSI-2 D-PHY 1.1 lanes (18 Gbps)	12 lanes MIPI CSI-2 D-PHY 1.2 (30 Gbps) C-PHY (41 Gbps)	12 lanes (3x4 or 6x2) MIPI CSI-2 D-PHY 1.2 (30 Gbps)	16 lanes MIPI CSI-2 | 8 lanes SLVS-EC D-PHY (40 Gbps) C-PHY (59 Gbps)
Mechanical	69.6mm x 45mm 260 pin edge connector	87mm x 50mm 400 pin connector	69.6mm x 45mm 260 pin edge connector	100mm x 87mm 699 pin connector
Software	JetPack SDK – Unified software release across all Jetson products			

CONTROLLING AIR TRAFFIC WITH AI

From autopilot systems to customer service to predicting weather, AI is transforming aviation. With Aimee—a GPU-powered framework for AI solutions from Searidge Technologies—Air Traffic Control no longer needs a direct sightline. Aimee analyzes video feeds from hundreds of cameras, enabling ATC to look past occlusions and "see" every runway, taxiway, tarmac, and gate without looking away from their workstations.

NVIDIA

SEARIDGE TECHNOLOGIES

SUPERHUMAN INSPECTION ACCURACY

Delivering impeccable quality is a great opportunity for high precision manufacturers to differentiate but raises the bar for accurate detection of the smallest micron-scale product defects.

Foxconn Interconnect Technology Group (FIT) is deploying AI-powered inspection systems with NVIDIA HGX-1, Tesla V100/P4, and Jetson TX2, and has improved its CPU socket defect detection escape rate from 4.3% to 0.015% - 287x

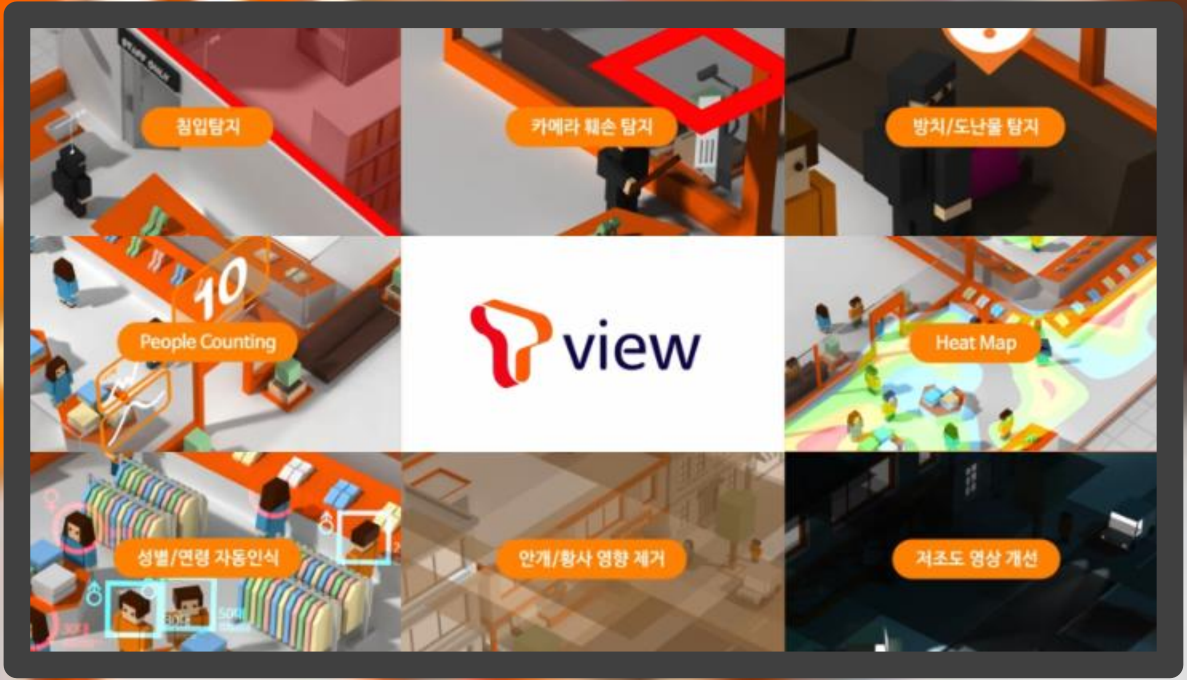
FIT

NVIDIA

ACCELERATING IVA FOR SMART CITIES

Intelligent video analysis (IVA) can safeguard citizens and property and is a key element of smart cities but analyzing data from millions of cameras in real-time requires deep learning and intensive computing power. SK Telecom uses NVIDIA GPUs to power T View, it's AI VSaaS (Video Surveillance as a Service) solution. With Tesla GPUs, SKT speeds training 5x, and with TensorRT to scale its inference engine, SKT achieves cost-efficiencies without sacrificing accuracy.

NVIDIA. SK telecom

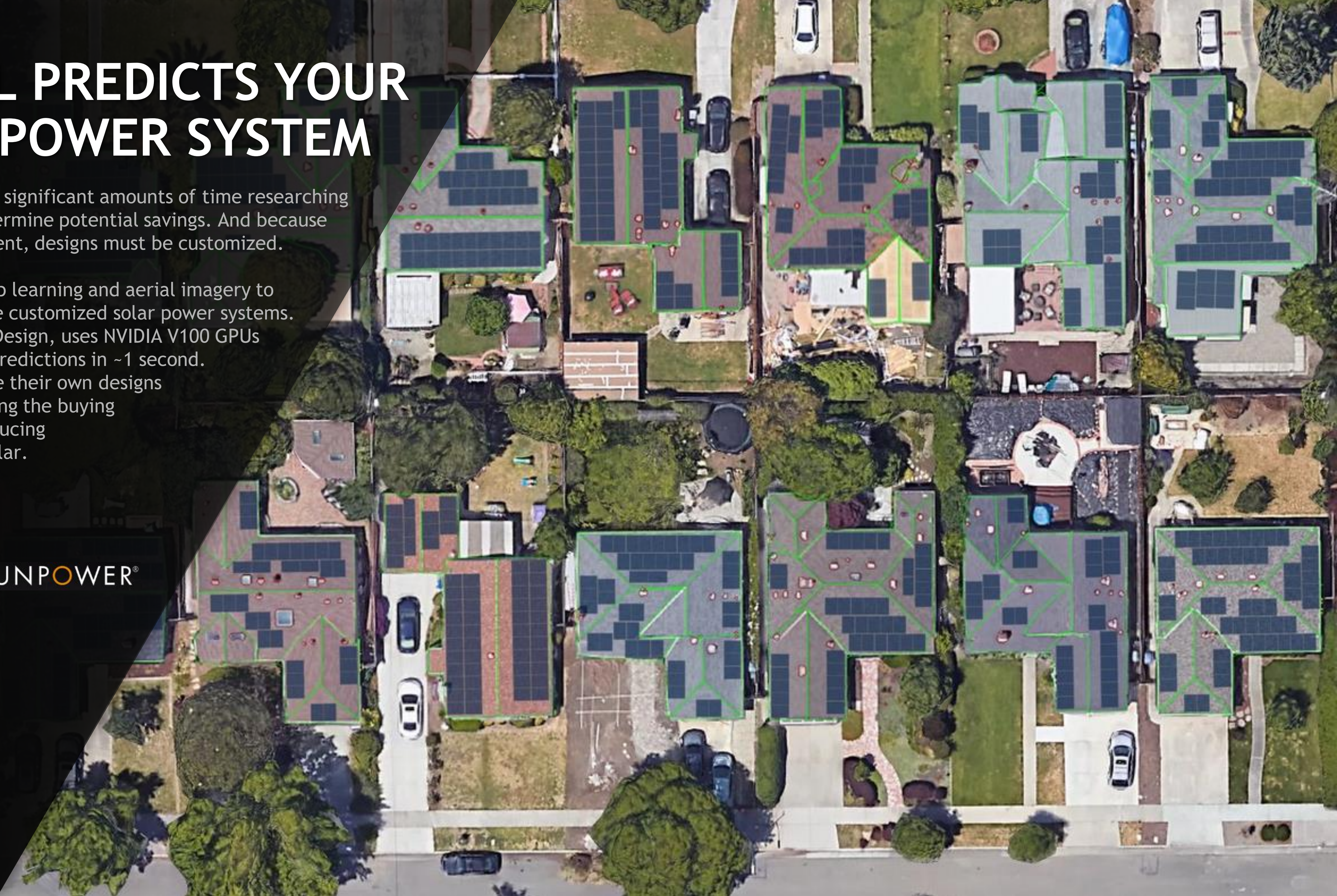


AI TOOL PREDICTS YOUR SOLAR POWER SYSTEM

Homeowners spend significant amounts of time researching solar panels to determine potential savings. And because every roof is different, designs must be customized.

SunPower uses deep learning and aerial imagery to design and visualize customized solar power systems. Its AI tool, Instant Design, uses NVIDIA V100 GPUs on GCP to deliver predictions in ~1 second. Homeowners create their own designs instantly — improving the buying experience and reducing barriers to going solar.

NVIDIA. SUNPOWER®



TRANSMISSION LINE FIELD INSPECTION

Industrial Inspection Automation

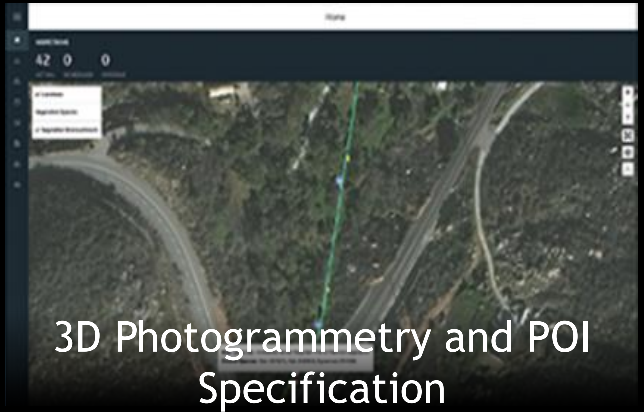
Costly, hazard prone and slow manual inspection of industrial assets results in downtimes and safety hazards

AI Workbench:

- Multi-Sensor: RGB Optical, Laser, Infrared and Exogenous data, Fugitive emission, Ultra sound
- RTX: 3D Visualization
- EGX: Data Capture, Path Planning, Continuous learning and Inferencing
- DGX: Training and Continuous Learning

Detection: Corrosion levels, Damaged/missing parts, Encroaching Vegetation volumes

Outcomes: 25-50% reduction in inspection cost and 50% avoidance of asset downtime.*



3D Photogrammetry and POI Specification

Autonomous Path Planning

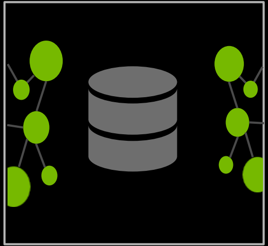
Intelligent Inspection

Avitas Systems
a GE venture

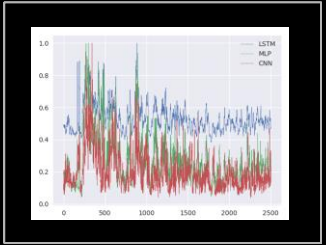
* ssurce: Avitas systems, a ge-venture-partners-nvidia-enhance-ai-robotic-inspection

INDUSTRIAL AI PLATFORM

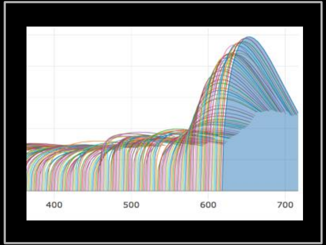
Data Management



Anomaly Detection



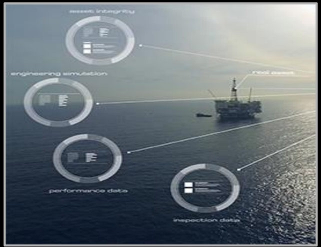
Predicting Failures



Inspection

Video Analytics

Digital Twin



AI FRAMEWORK FOR TRAINING

Visualization, Data Prep, Architecture Optimization, Model and Analytics Orchestration,
Serving, Lifecycle, Modeling Templates, Repository

METROPOLIS FOR EDGE

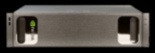
TensorRT, DeepStream, Smart Indexing and Storage, IOT Runtime

Jetson SDK
Anomaly Detection, Inspection

AGX

Tesla

DGX
Station



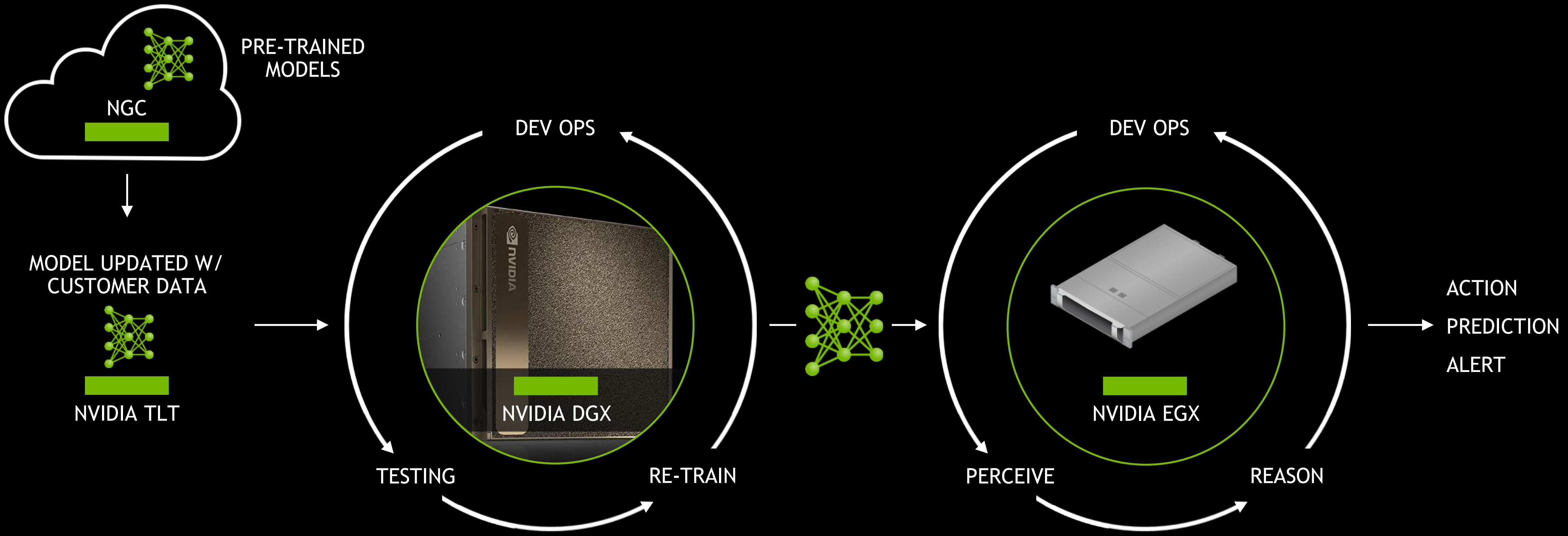
DGX,
HGX

Data
Center

Google Cloud
Azure
aws

Cloud

NVIDIA AI END-TO-END WORKFLOW



NGC

PRE-TRAINED MODELS

MODEL UPDATED W/ CUSTOMER DATA

NVIDIA TLT

DEV OPS

TESTING

RE-TRAIN

NVIDIA DGX

DEV OPS

PERCEIVE

REASON

NVIDIA EGX

ACTION

PREDICTION

ALERT

DEVELOP AI

DEPLOY AI

NVIDIA EGX METROPOLIS BENEFITS

NVIDIA is Industry’s Most Advanced AI Computing Platform

Largest domain of AI models

High throughput & low latency

One architecture scalable from device to cloud

NVIDIA EGX is AI-Optimized Hyper-converged Infrastructure

Hardware and software optimized for AI, storage, networking & security

Easy development & deployment of AI at edge

Optimized AI models in NGC

NVIDIA is an Open Platform

Support for every platform — VMW, RH, NTNX, Azure, AWS, GCP

Rich 3rd party ISV ecosystem

Rich OEM and integrator ecosystem

NVIDIA is Pervasive AI Platform

Every cloud

Hybrid cloud

Edge to cloud

NVIDIA has Deep AI Expertise

End-to-end, from development to deployment, from tools to experts

NVIDIA Research

DLI to reskill talent

SA & DevTech to co-engineer

NVIDIA has Global Reach and Support

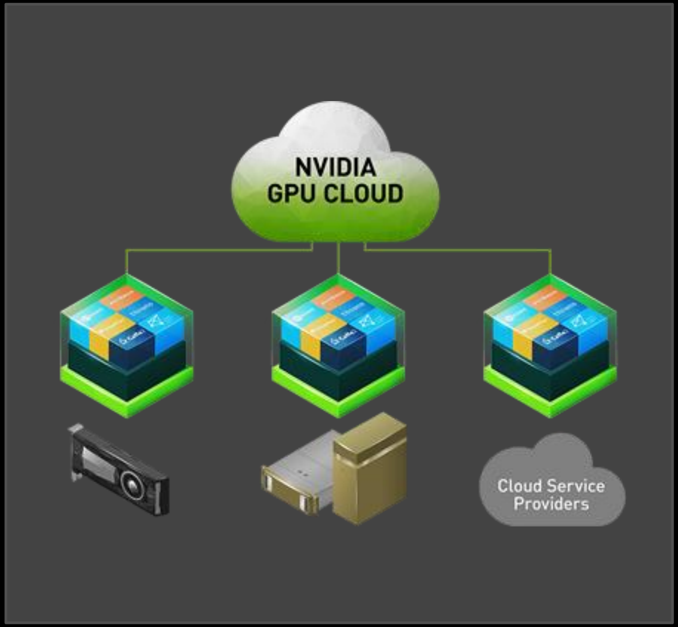
Expertise in large verticals — M&E, healthcare, retail, manufacturing, transportation, and more

BD, SA, DevRel, DevTech, Research in every region

MANUFACTURING RESOURCES

Industrial AI Content

NVIDIA Deep Learning Institute
www.nvidia.com/en-us/deep-learning-ai/education



NGC
www.nvidia.com/en-us/gpu-cloud

NVIDIA®