




GPU ACCELERATED ANALYTICS WITH AZURE IOT

Dr. Ulrich Knechtel, February 2020, uknechtel@nvidia.com



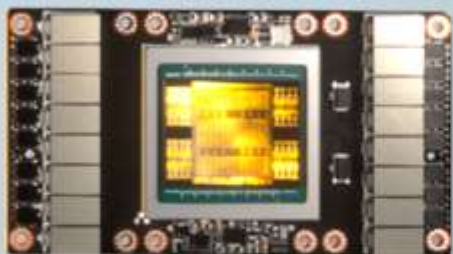
NVIDIA — A COMPUTING PLATFORM COMPANY



NVIDIA pioneered accelerated computing to solve problems that normal computers cannot solve. The approach is broadly recognized as the way to advance computing as Moore's law ends and AI lifts off. NVIDIA's platform is installed in several hundred million computers, is available in every cloud and from every server maker, powers 136 of the TOP500 supercomputers, and boasts 1.6 million developers.

AT THE INTERSECTION OF GRAPHICS, HPC, AI

NVIDIA innovates at the intersection of graphics, HPC, and AI. We simulate worlds, physics, and intelligence in real time. We make computers for the da Vincis and Einsteins of our time so that they can see and create the future.



THREE REVOLUTIONS HAPPENING

IOT



IOT devices projected to grow to >150B by 2025, >1T by 2035

5G



5G will deliver 1000X better bandwidth and 10X lower latency than 4G

AI



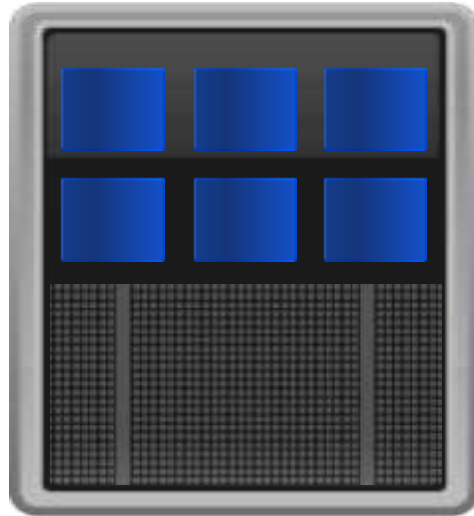
By 2025, AI at the edge has a potential total economic impact of up to \$11T/year

ACCELERATED COMPUTING

Focus on Performance, Energy Efficiency and Throughput

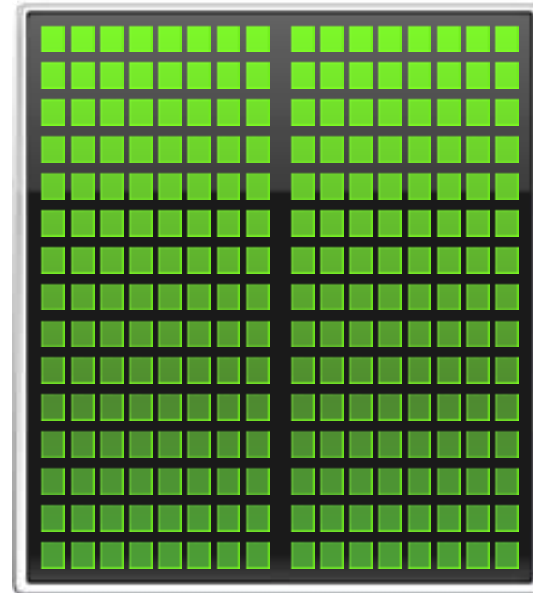
CPU

Optimized for
Serial Tasks

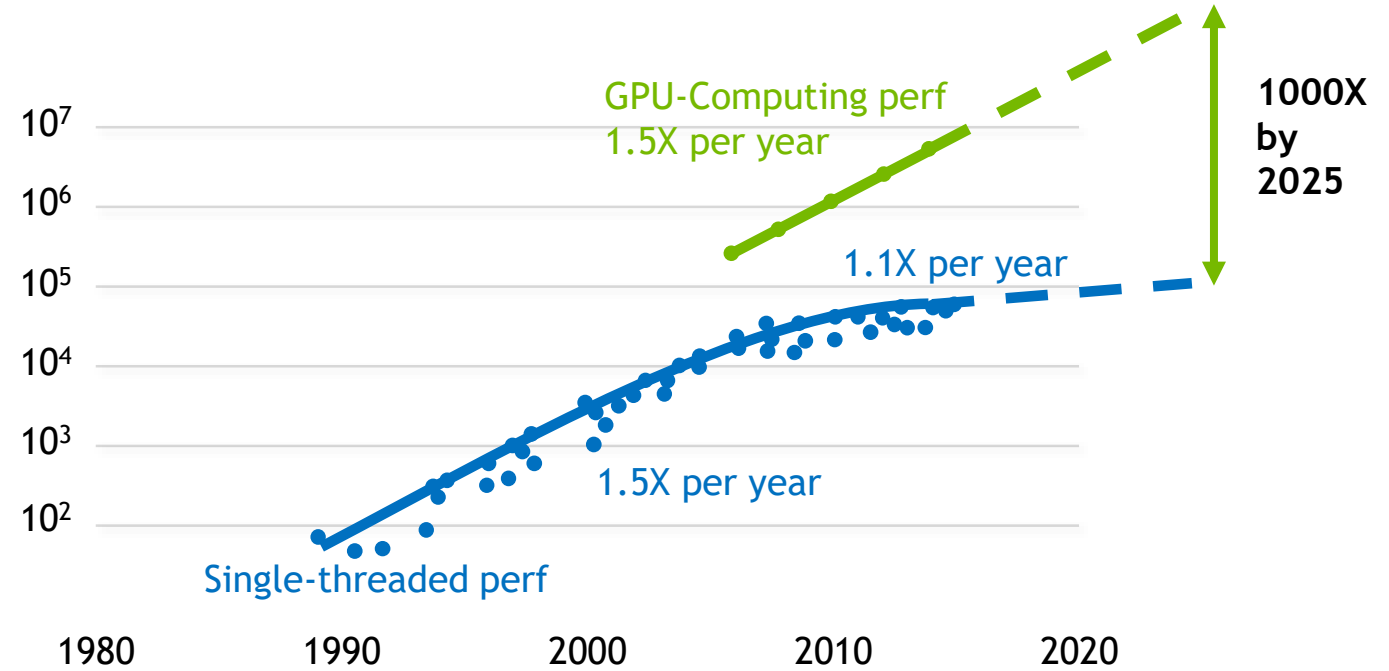
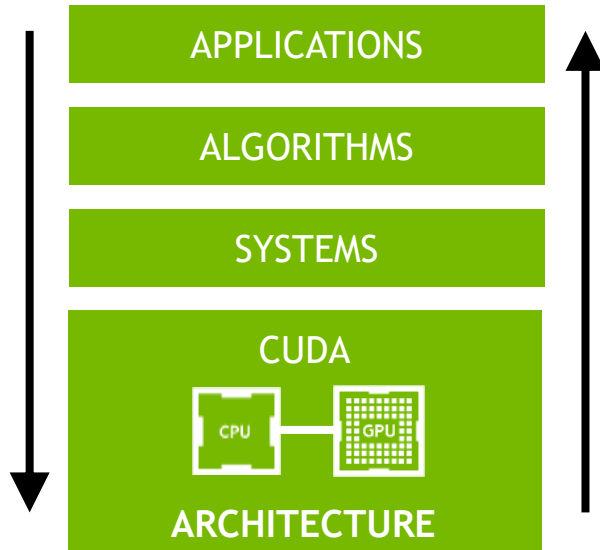


GPU Accelerator

Optimized for
Parallel Tasks



RISE OF GPU COMPUTING



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

NVIDIA DATA CENTER PLATFORM

Single Platform Drives Utilization and Productivity

CUSTOMER USE CASES



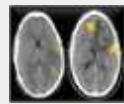
Speech



Translate



Recommender



Healthcare



Manufacturing



Finance



Molecular
Simulations



Weather
Forecasting



Seismic
Mapping



Creative &
Technical



Knowledge
Workers

CONSUMER INTERNET & INDUSTRY APPLICATIONS

SCIENTIFIC APPLICATIONS

VIRTUAL GRAPHICS

APPS & FRAMEWORKS



python™



TensorFlow



mxnet



Chainer



ONNX

RAPIDS

PYTORCH

Amber
NAMD

+600
Applications

CATIA



AUTODESK
3DS MAX

Ps



Windows 10

CUDA-X & NVIDIA SDKs

MACHINE LEARNING

cuDF

cuML

cuGRAPH

DEEP LEARNING

cuDNN

CUTLASS

TensorRT

HPC

OpenACC

cuFFT

VIRTUAL GPU

vDWS

vPC

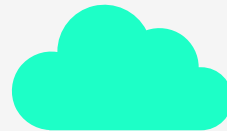
vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

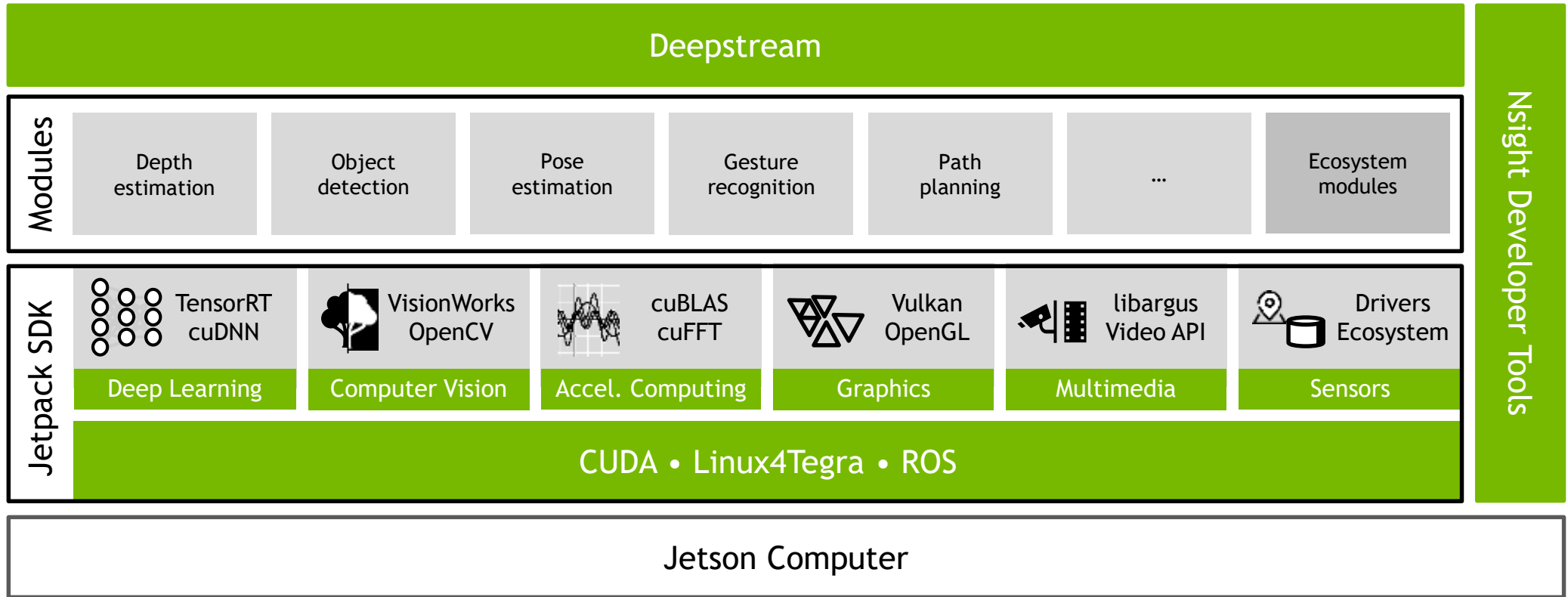
TESLA GPUs & SYSTEMS



TESLA (server) GPU

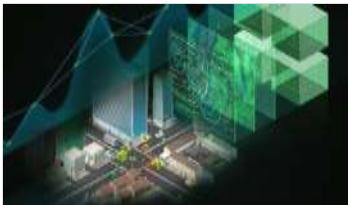


JETSON SOFTWARE



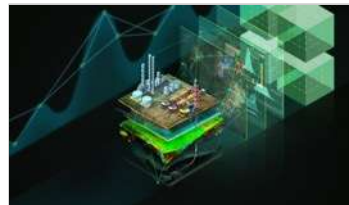
GPU-ACCELERATED DATA SCIENCE

Use Cases in Every Industry



CONSUMER INTERNET

- Ad Personalization
- Click Through Rate Optimization
- Churn Reduction



OIL & GAS

- Sensor Data Tag Mapping
- Anomaly Detection
- Robust Fault Prediction



FINANCIAL SERVICES

- Claim Fraud
- Customer Service Chatbots/Routing
- Risk Evaluation



MANUFACTURING

- Remaining Useful Life Estimation
- Failure Prediction
- Demand Forecasting



HEALTHCARE

- Improve Clinical Care
- Drive Operational Efficiency
- Speed Up Drug Discovery



TELECOM

- Detect Network/Security Anomalies
- Forecasting Network Performance
- Network Resource Optimization (SON)



RETAIL

- Supply Chain & Inventory Management
- Price Management / Markdown Optimization
- Promotion Prioritization And Ad Targeting

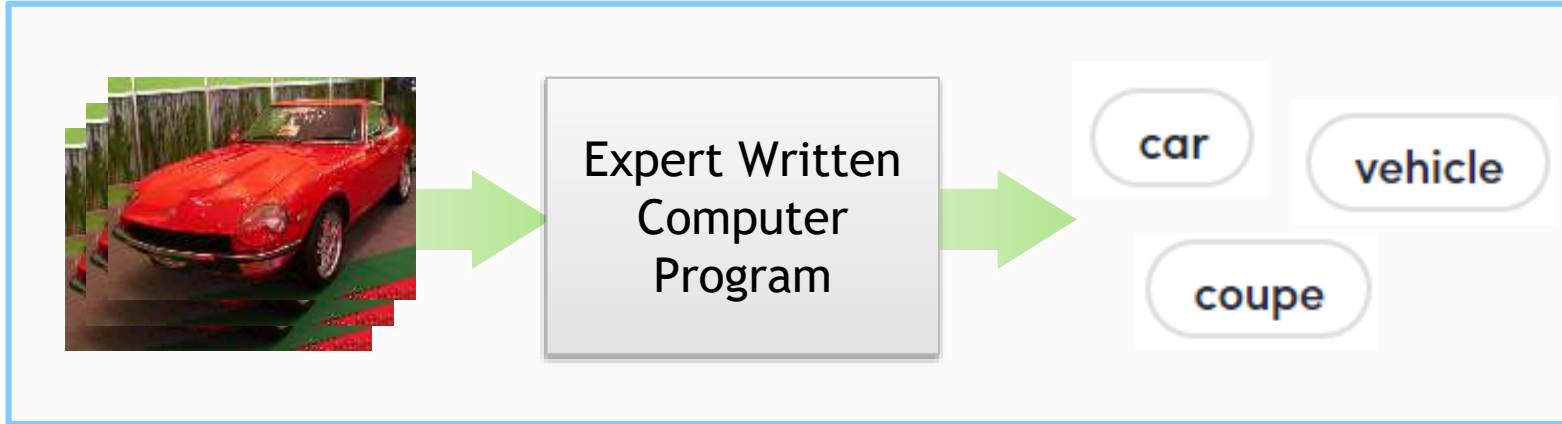


AUTOMOTIVE

- Personalization & Intelligent Customer Interactions
- Connected Vehicle Predictive Maintenance
- Forecasting, Demand, & Capacity Planning

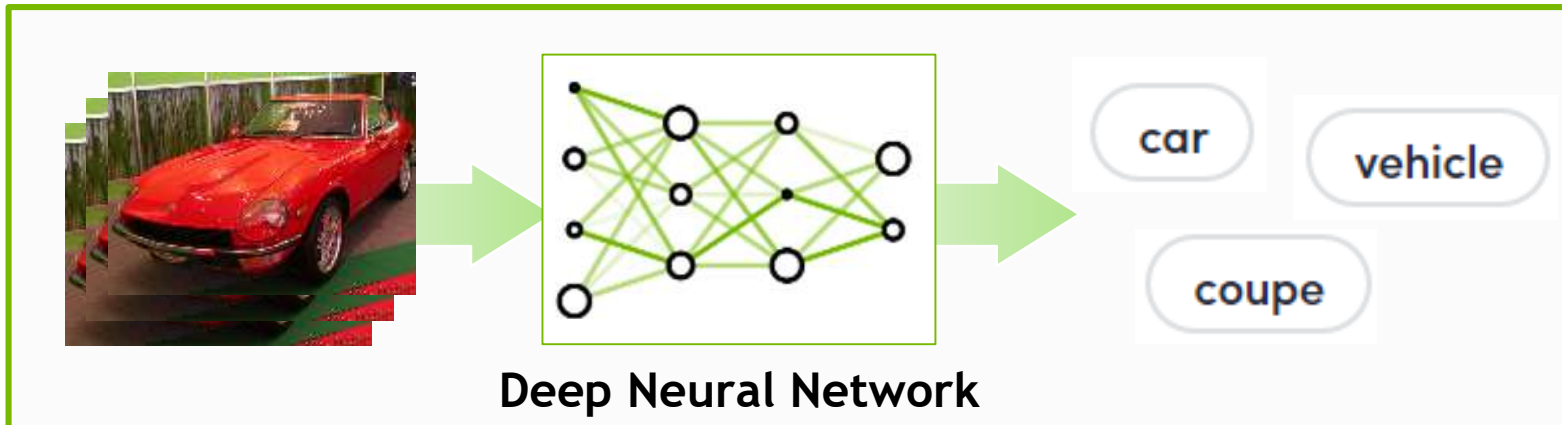
DEEP LEARNING - A NEW COMPUTING MODEL

Algorithms that Learn from Examples



Traditional Approach

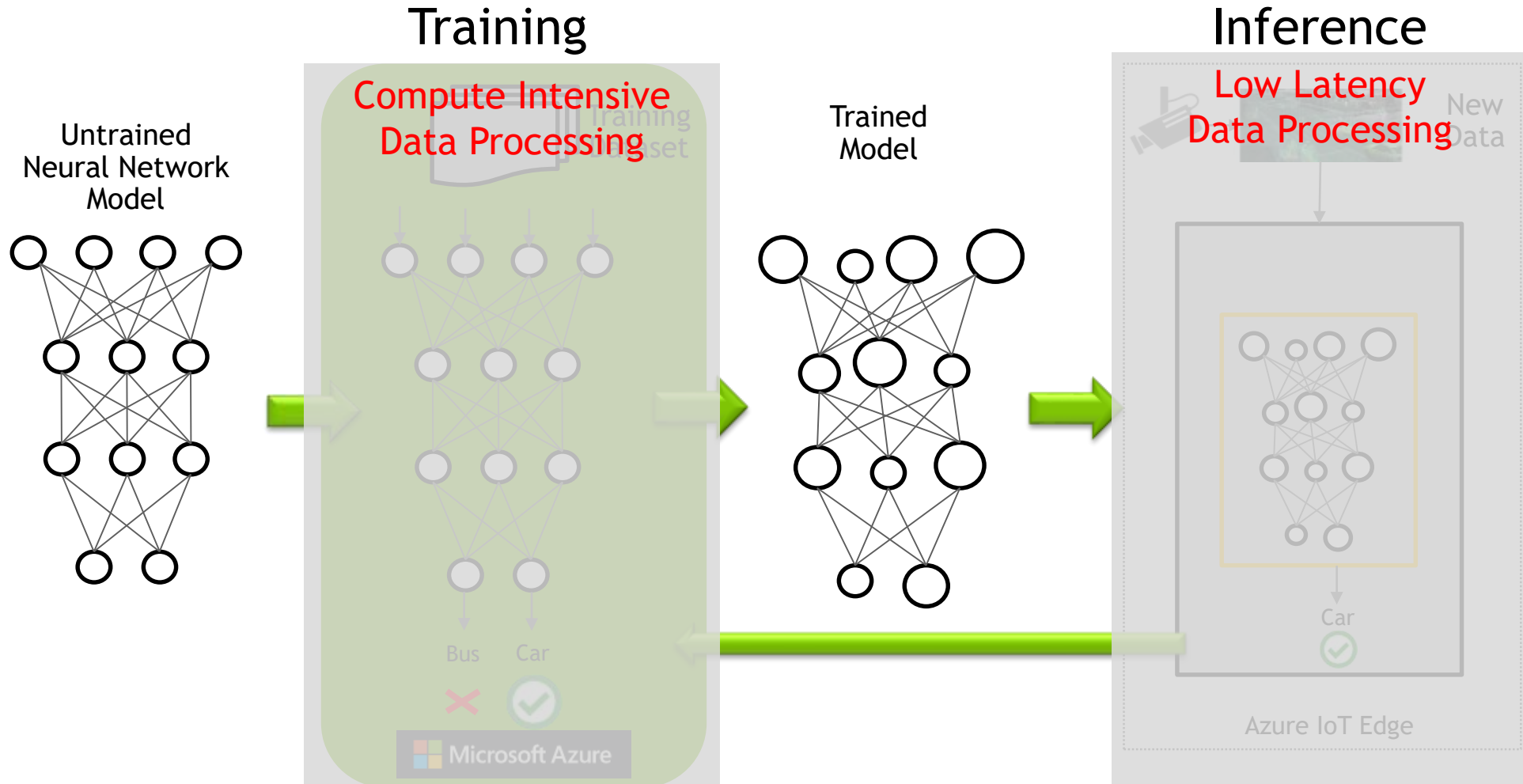
- Requires domain experts
- Time consuming
- Error prone
- Not scalable to new problems



Deep Learning Approach

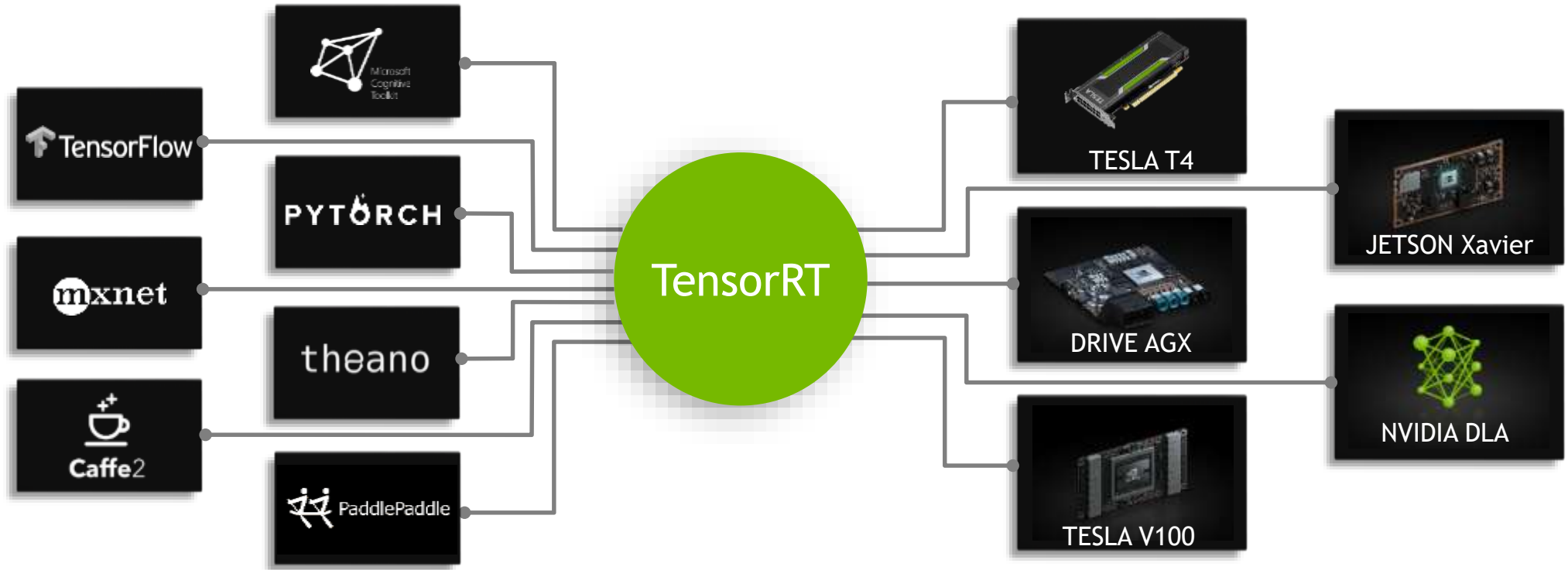
- ✓ Learn from data
- ✓ Easily to extend
- ✓ Speedup with GPUs

DEEP LEARNING AND IOT EDGE



NVIDIA TensorRT

From Every Framework, Optimized For Each Target Platform



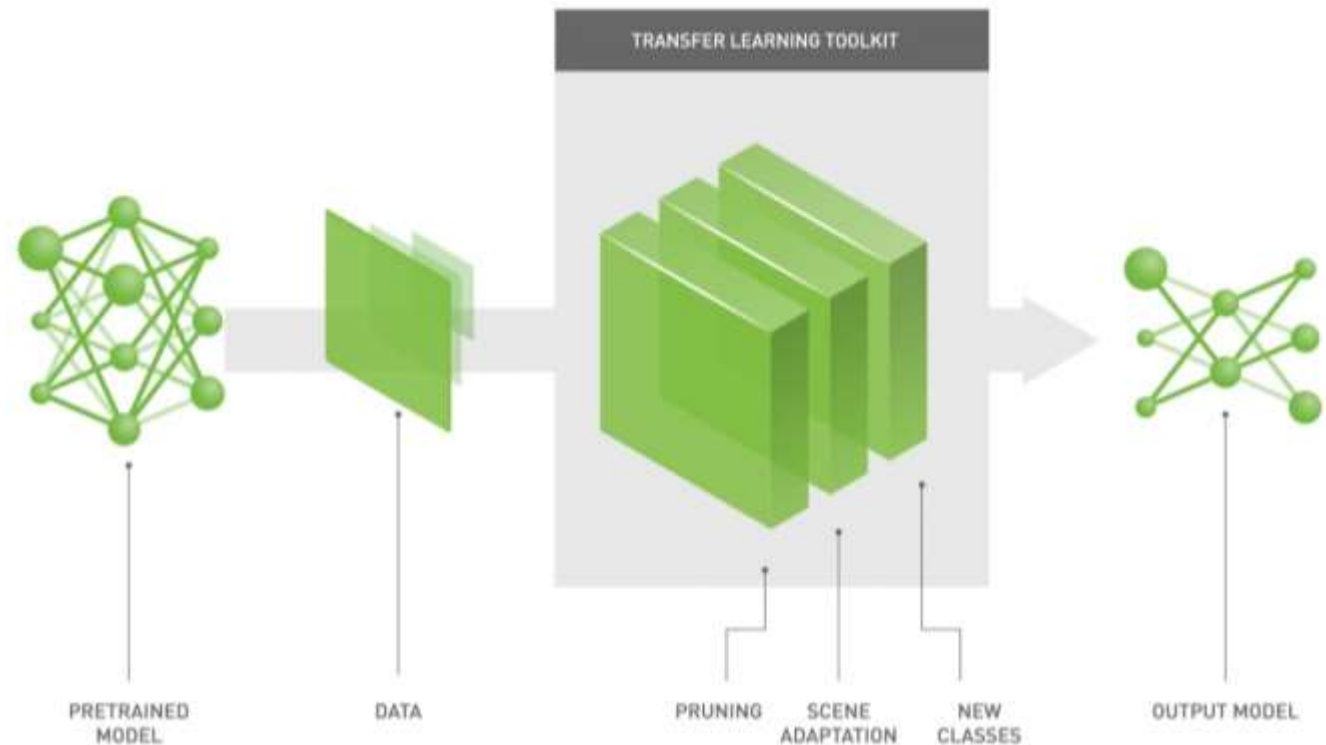
TRANSFER LEARNING TOOLKIT

High level SDK for tuning of domain specific DNNs

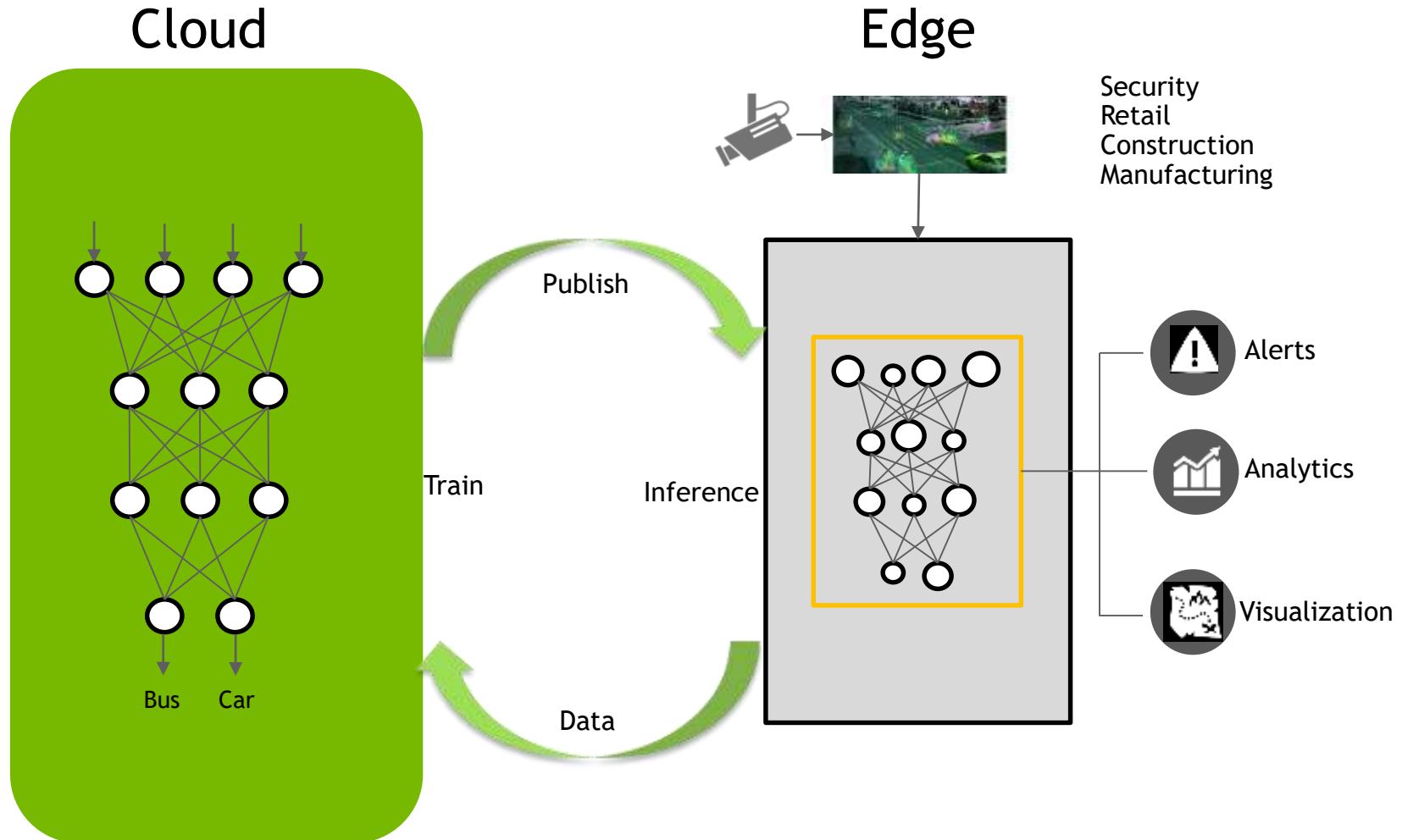
Faster and efficient deep learning training workflow

Leverages prior investment in pretrained models

Output optimized and ready for deployment



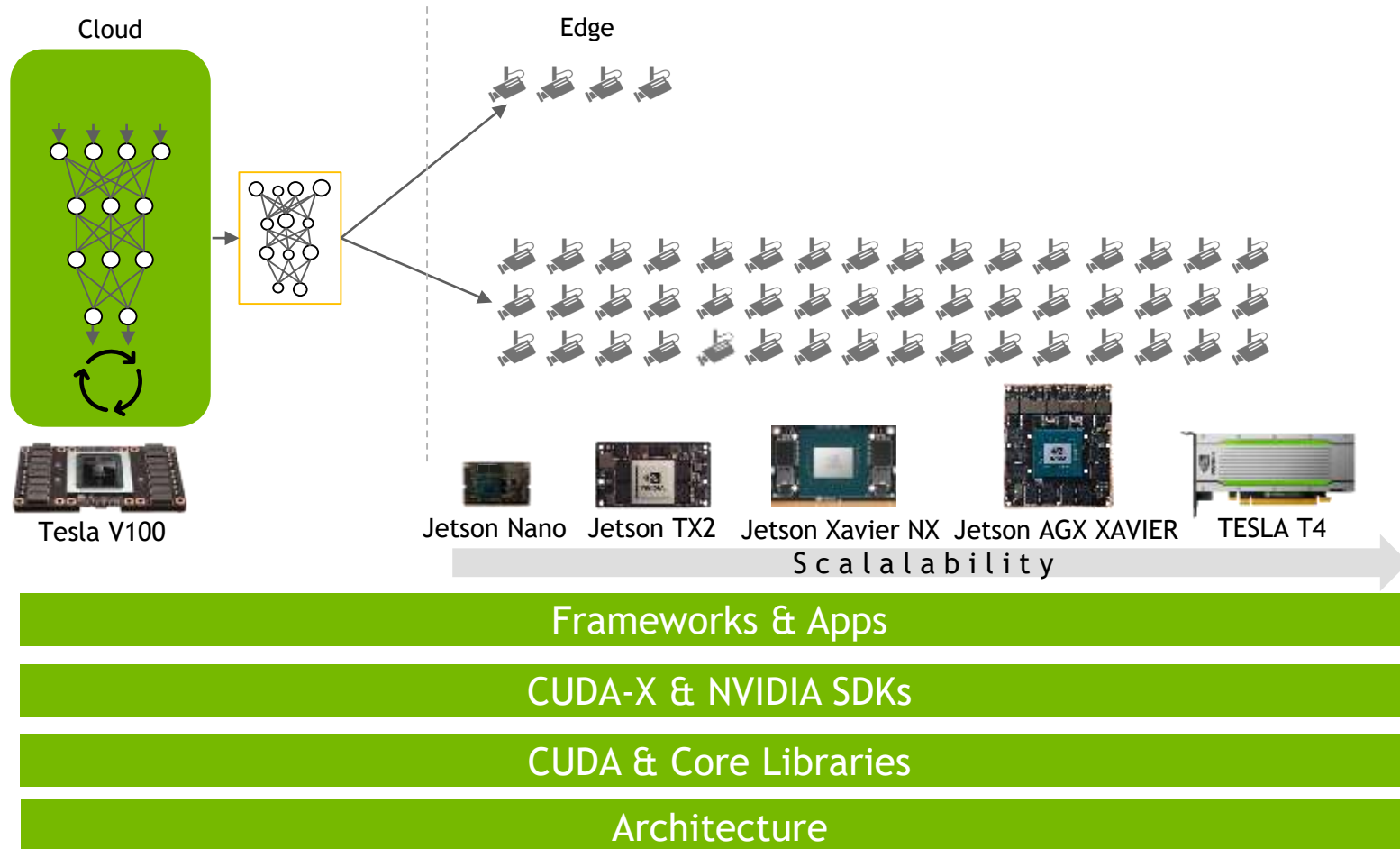
DEEP LEARNING AND IOT EDGE





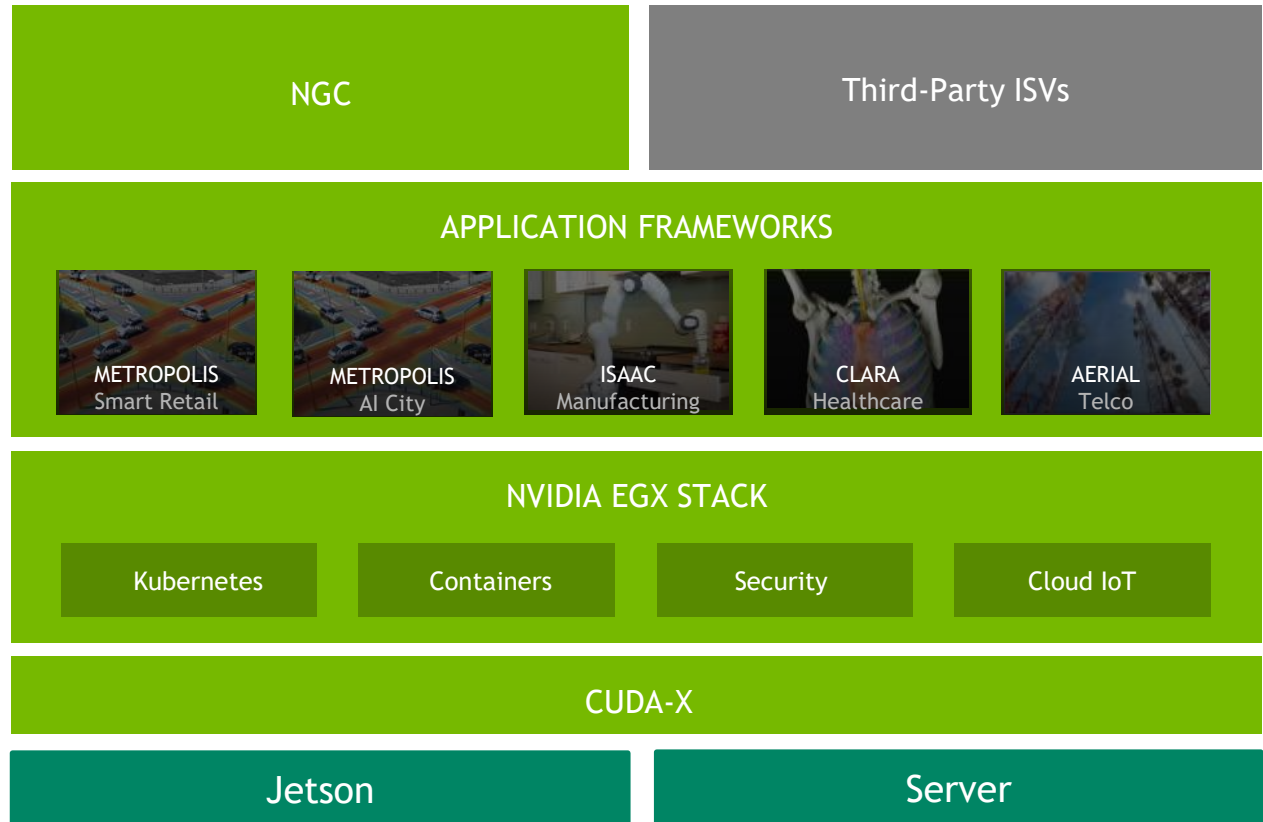
WORKFLOW SUPPORT

Development and scalable Deployment

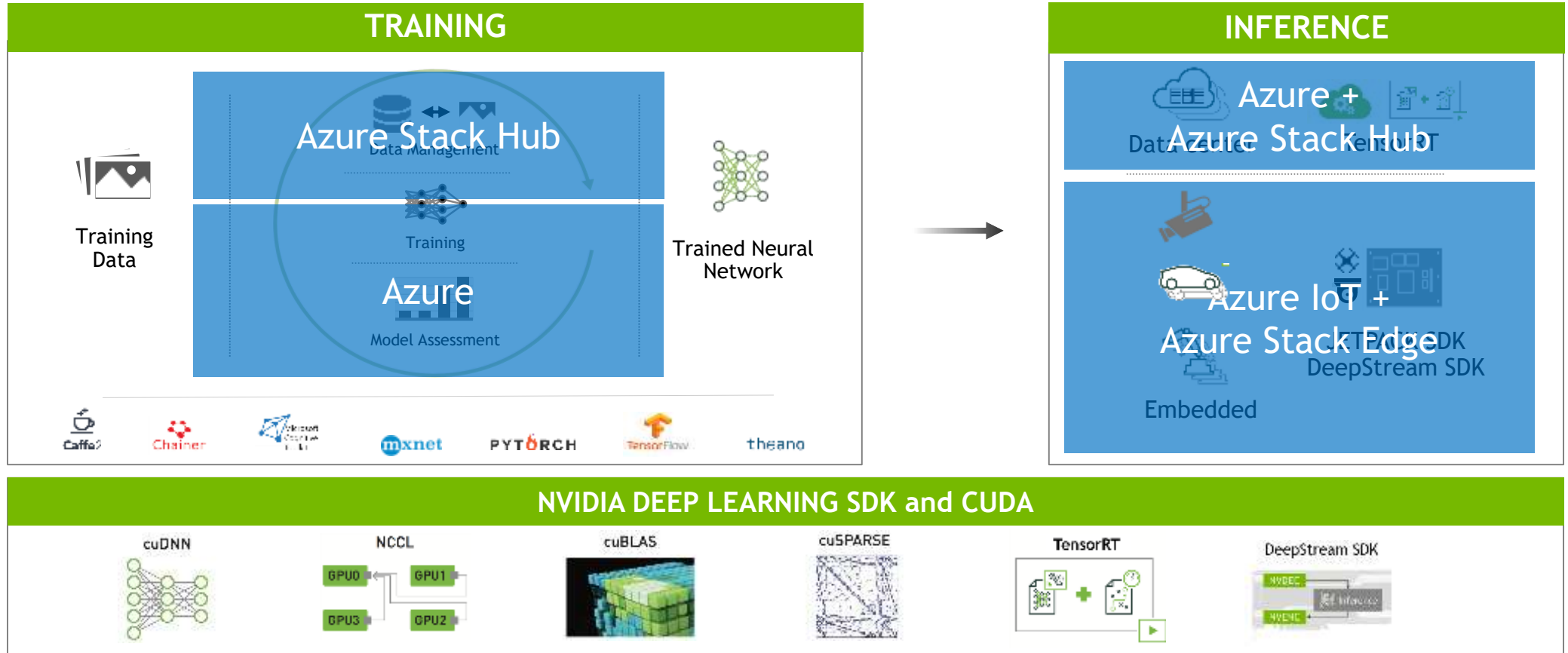


SINGLE PLATFORM FOR SERVICES AND 5G RAN

- ▶ Powered by NVIDIA CUDA GPU
- ▶ EGX Stack - supported by Azure
- ▶ Vertical Industry SDKs
- ▶ Commercially off the shelf (COTS)
- ▶ Scale from 2W to 2 Petaflops



NVIDIA DEEP LEARNING SOFTWARE PLATFORM









Conversation: Inactive

WHAT PROBLEM ARE YOU SOLVING?

Defining the AI/DL task

INPUTS	BUSINESS QUESTIONS	AI / DL TASK	EXAMPLE OUTPUTS		
			HEALTHCARE	RETAIL	FINANCE
 Text Data  Images  Video  Audio	Is “it” present or not?	Detection	Cancer Detection	Targeted Ads	Cybersecurity
	What type of thing is “it”?	Classification	Image Classification	Basket Analysis	Credit Scoring
	To what extent is “it” present?	Segmentation	Tumor Size / Shape Analysis	Build 360° Customer View	Credit Risk Analysis
	What is the likely outcome ?	Prediction	Survivability Prediction	Sentiment & Behavior Recognition	Fraud Detection
	What will likely satisfy the objective?	Recommendations	Therapy Recommendation	Recommendation Engine	Algorithmic Trading

INTELLIGENT VIDEO ANALYTICS (IVA) FOR EFFICIENCY AND SAFETY

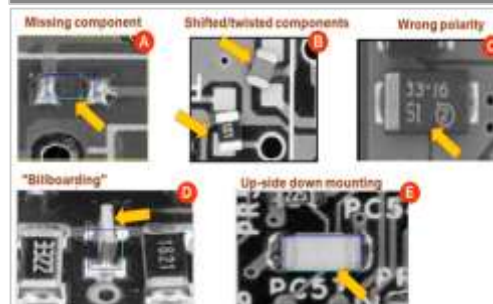
Access Control



Public Transit



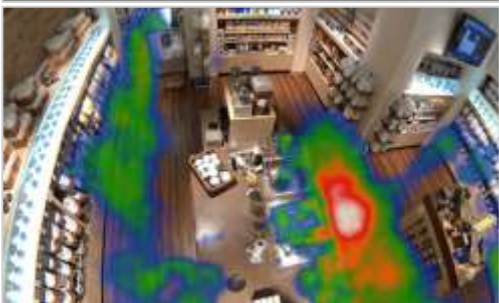
Industrial Inspection



Traffic Engineering



Retail Analytics



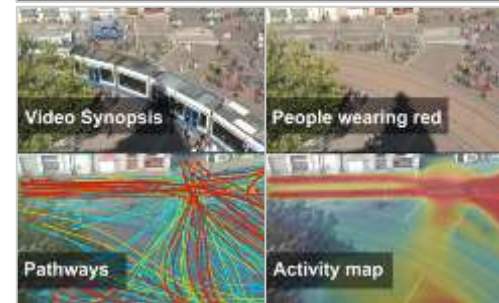
Logistics



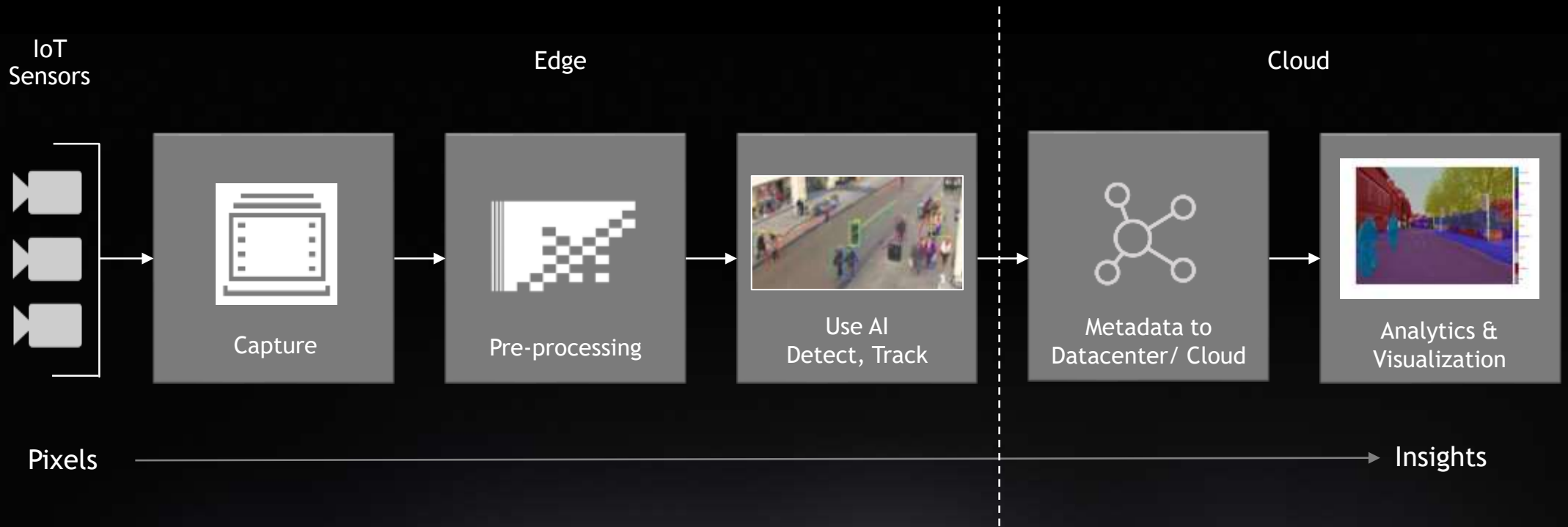
Critical Infrastructure



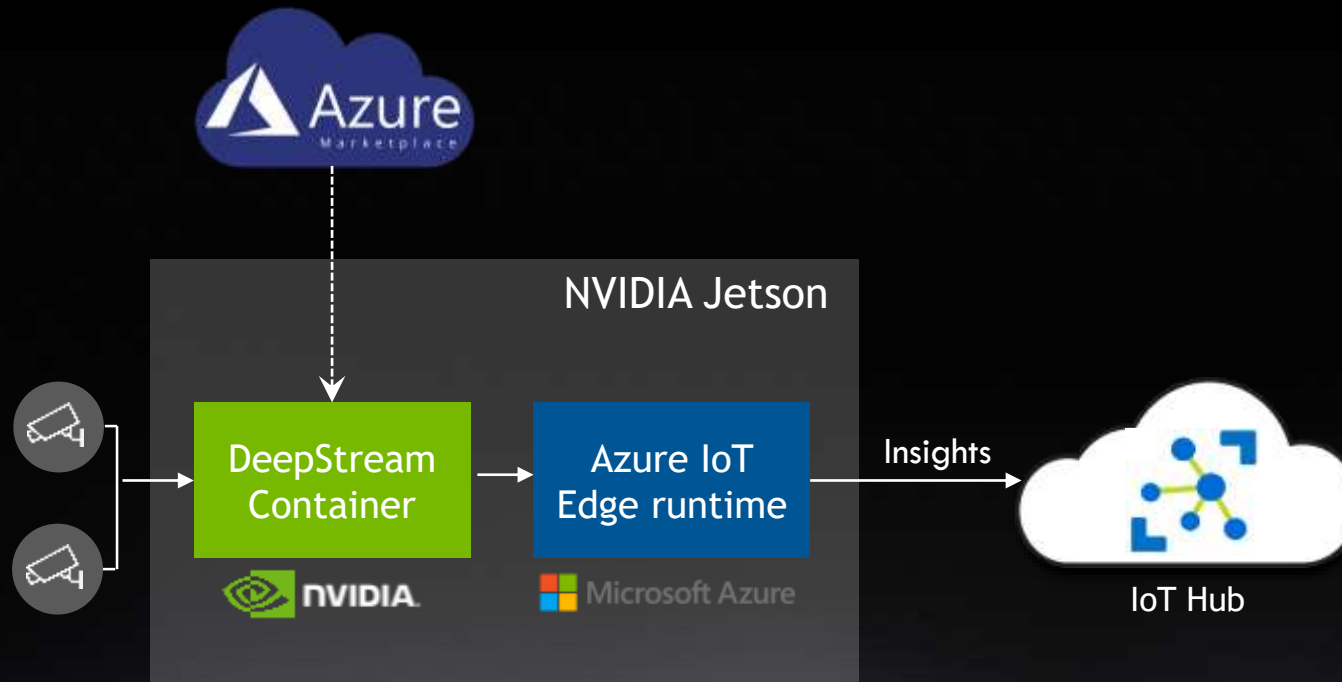
Public Safety



IVA APPLICATION WORKFLOW



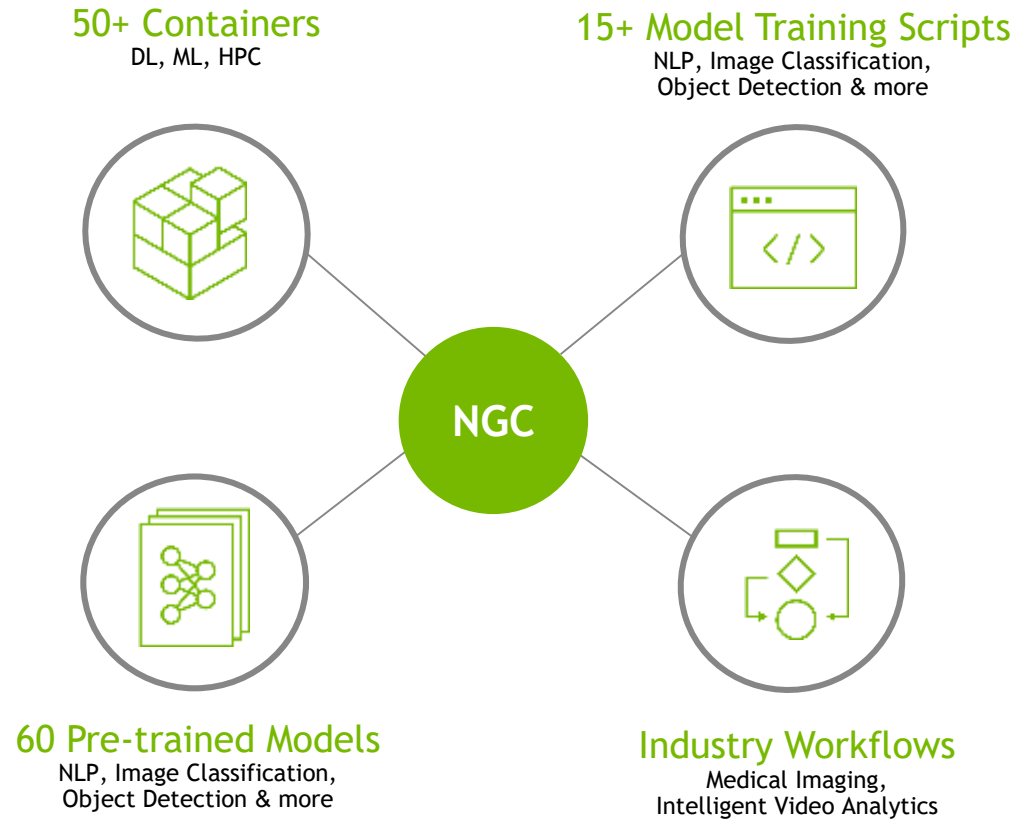
DEEPSTREAM WITH IOT EDGE RUNTIME



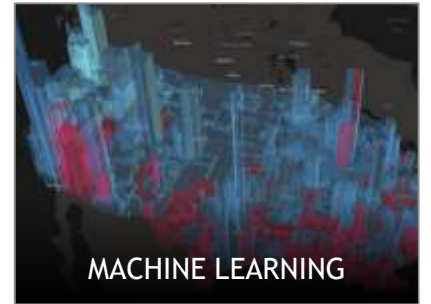
Connect to Azure IoT Hub through Azure IoT edge runtime

NGC: GPU-OPTIMIZED SOFTWARE HUB

Simplifying DL, ML, and HPC Workflows



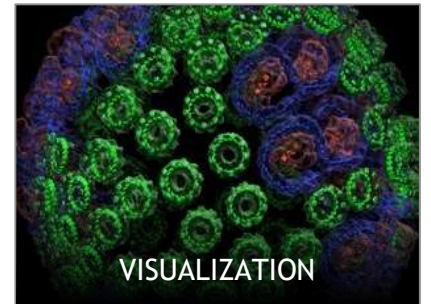
TensorFlow | PyTorch | more



RAPIDS | H2O | more



NAMD | GROMACS | more



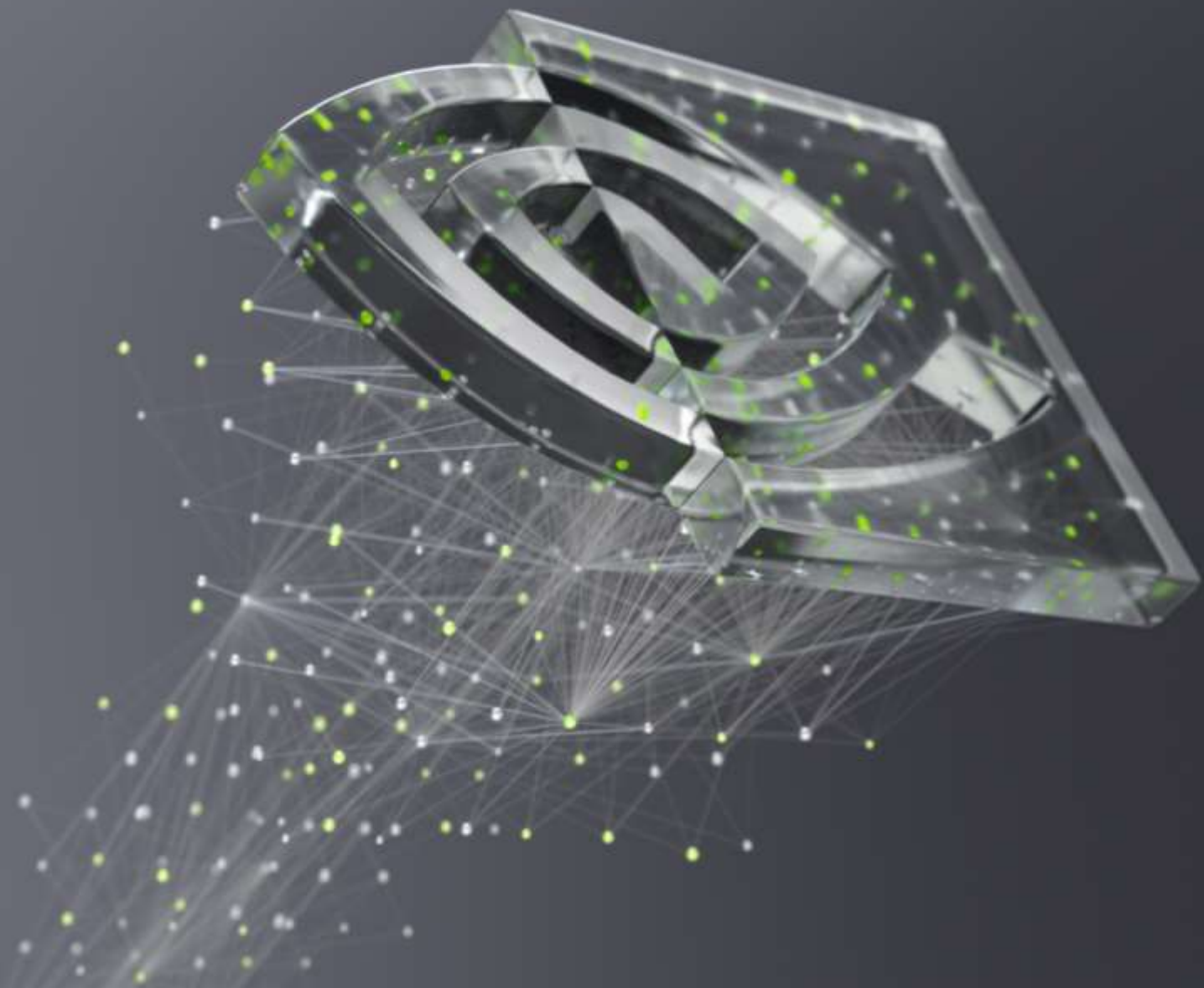
ParaView | Index | more

KEY TAKEAWAYS

NVIDIA is the Industry's Most Advanced AI Computing Platform

NVIDIA EGX is AI-Optimized for Industry IoT Use Cases

Processes from solution development to production roll-out tight into Microsoft platforms and services

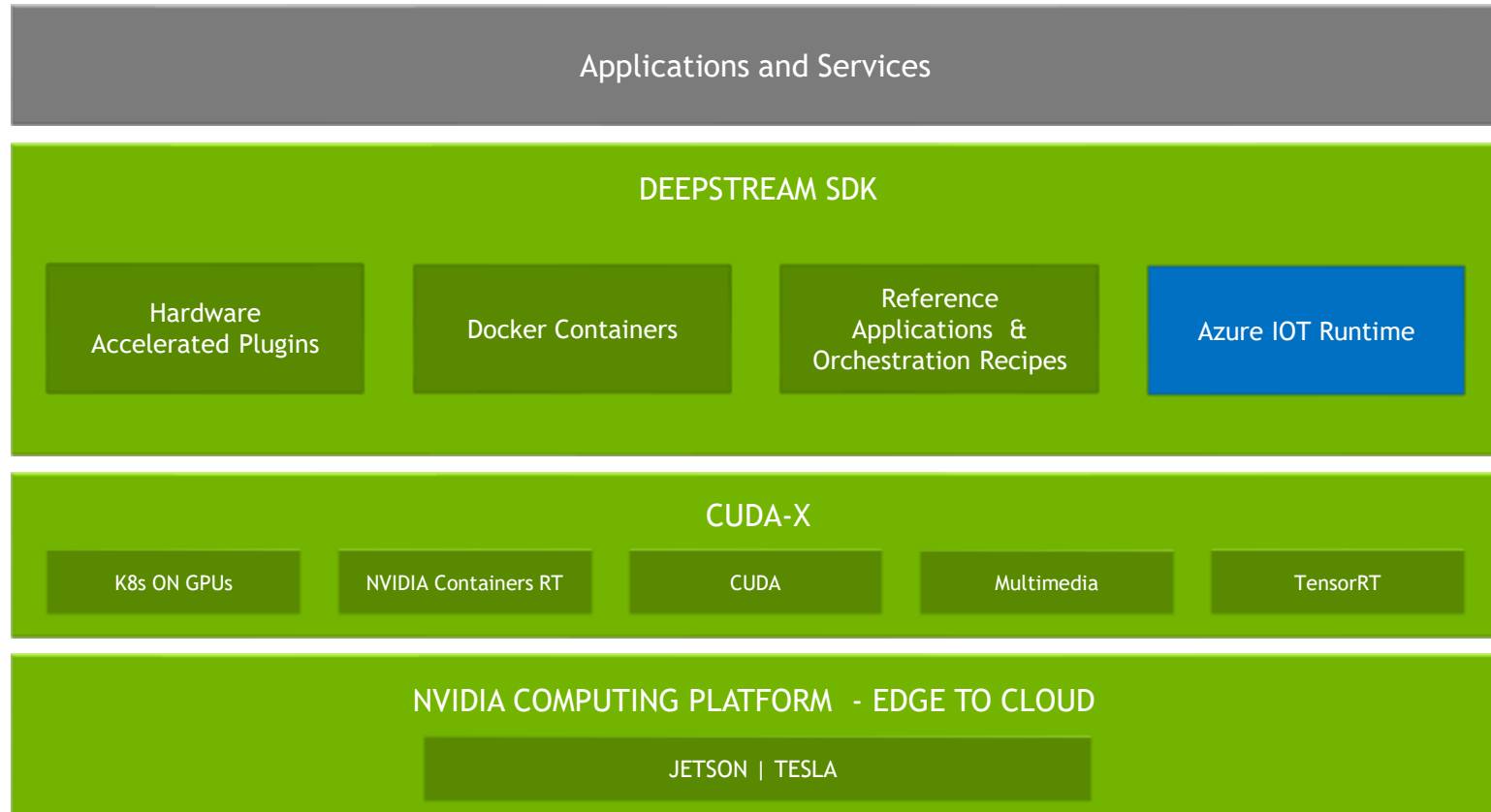


nvidia.



BACKUP

WHAT IS DEEPSTREAM?



JETSON ECOSYSTEM

DISTRIBUTION



SOFTWARE



HW AND SENSORS

