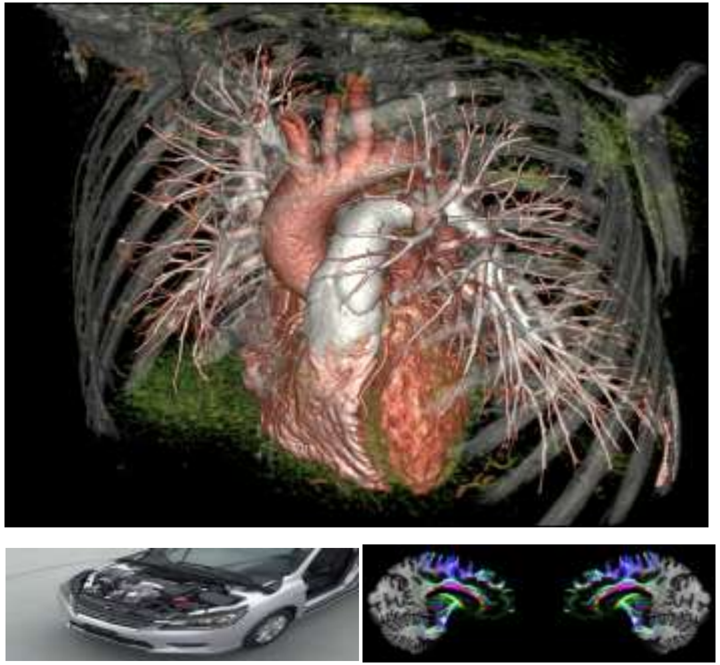
NVIDIA ACCELERATED COMPUTING

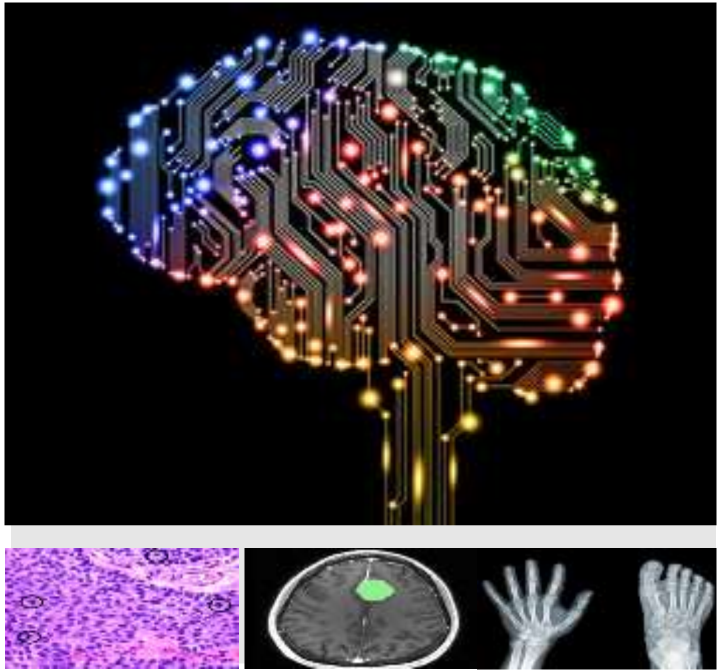
Uli Knechtel, September 2019 uknechtel@nvidia.com

NVIDIA “The AI Computing Company”



Computer Graphics

GPU Computing



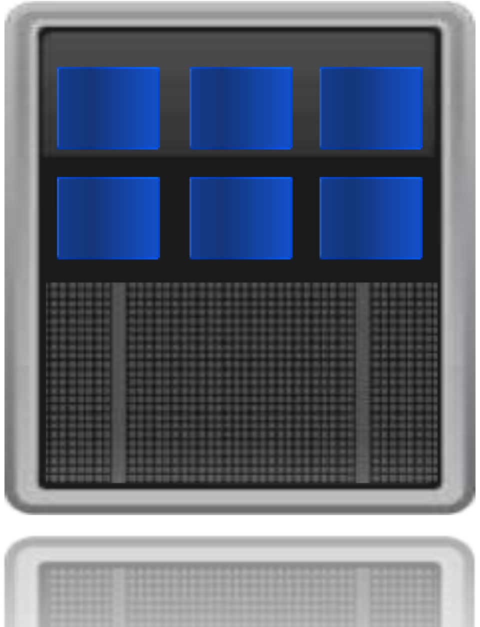
Artificial Intelligence

ACCELERATED COMPUTING

Focus on Performance, Energy Efficiency and Throughput

CPU

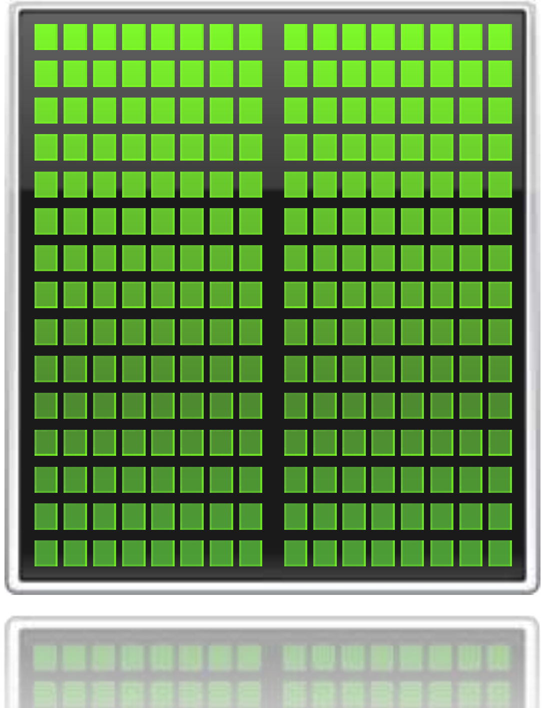
Optimized for
Serial Tasks



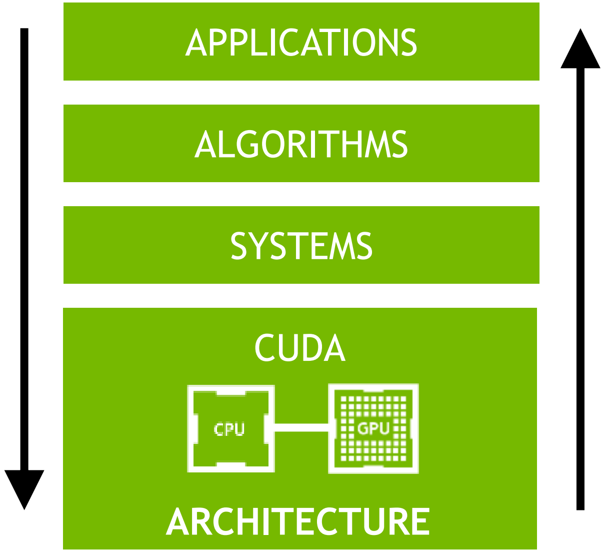
+

GPU Accelerator

Optimized for
Parallel Tasks



RISE OF GPU COMPUTING



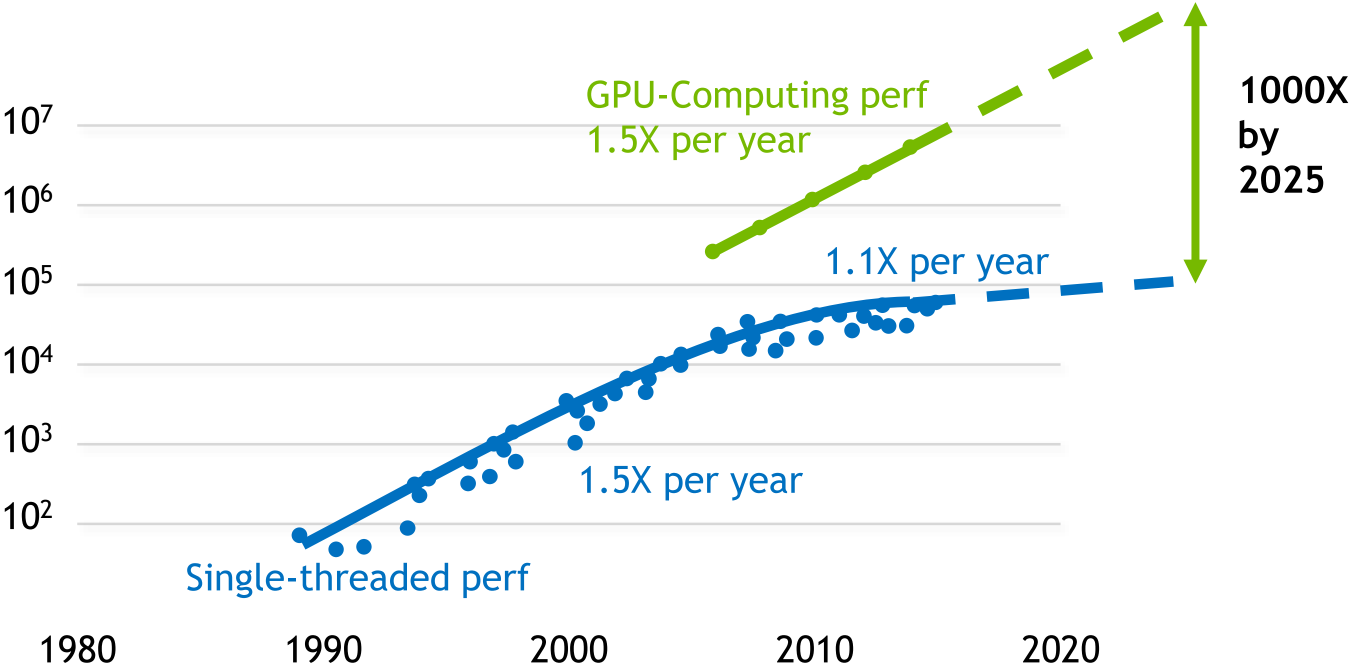
APPLICATIONS

ALGORITHMS

SYSTEMS

CUDA

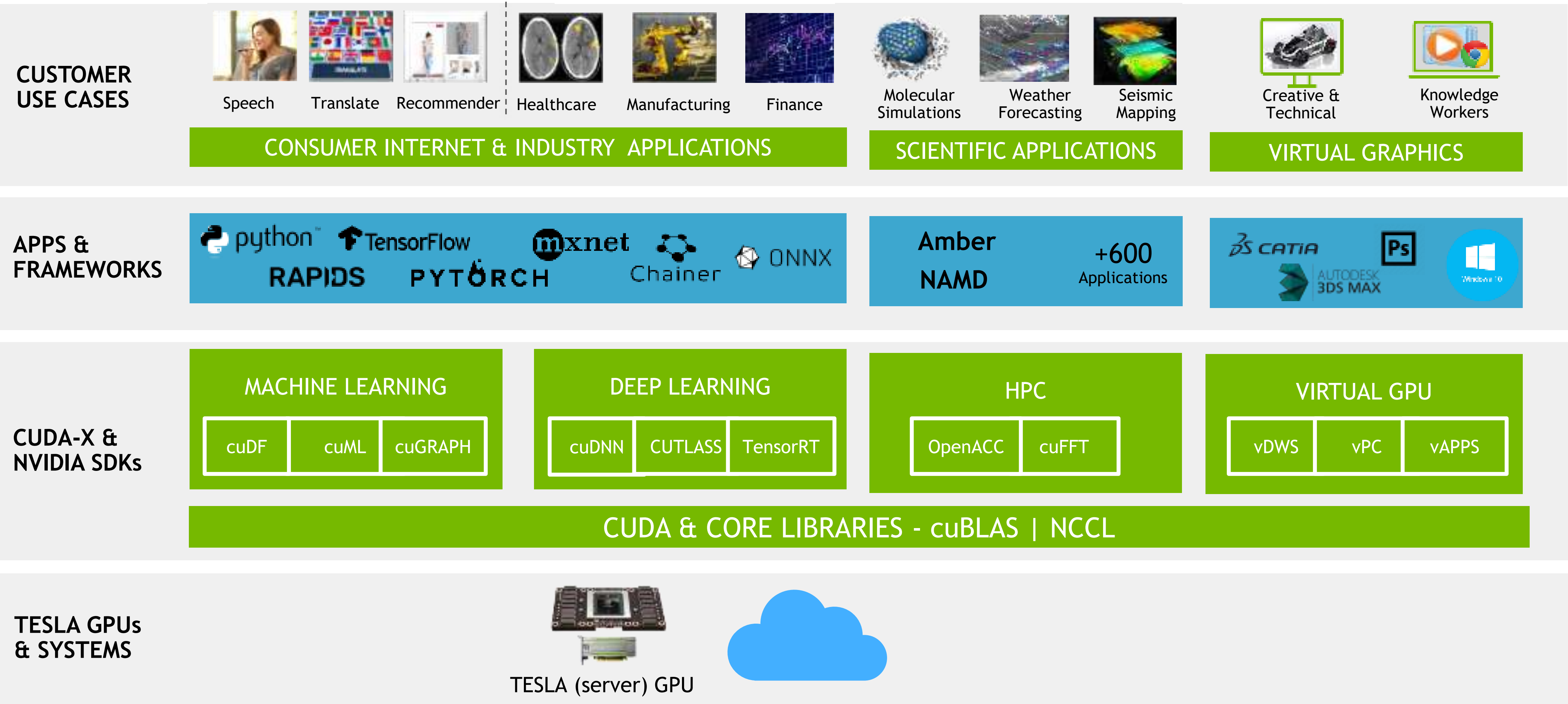
ARCHITECTURE



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

CONSUMER INTERNET & INDUSTRY APPLICATIONS

Molecular Simulations | Weather Forecasting | Seismic Mapping

SCIENTIFIC APPLICATIONS

Creative & Technical | Knowledge Workers

VIRTUAL GRAPHICS

APPS & FRAMEWORKS

python | TensorFlow | mxnet | Chainer | ONNX | RAPIDS | PYTORCH

Amber | NAMD | +600 Applications

CATIA | Ps | AUTODESK 3DS MAX | Windows 10

CUDA-X & NVIDIA SDKs

MACHINE LEARNING: cuDF | cuML | cuGRAPH

DEEP LEARNING: cuDNN | CUTLASS | TensorRT

HPC: OpenACC | cuFFT

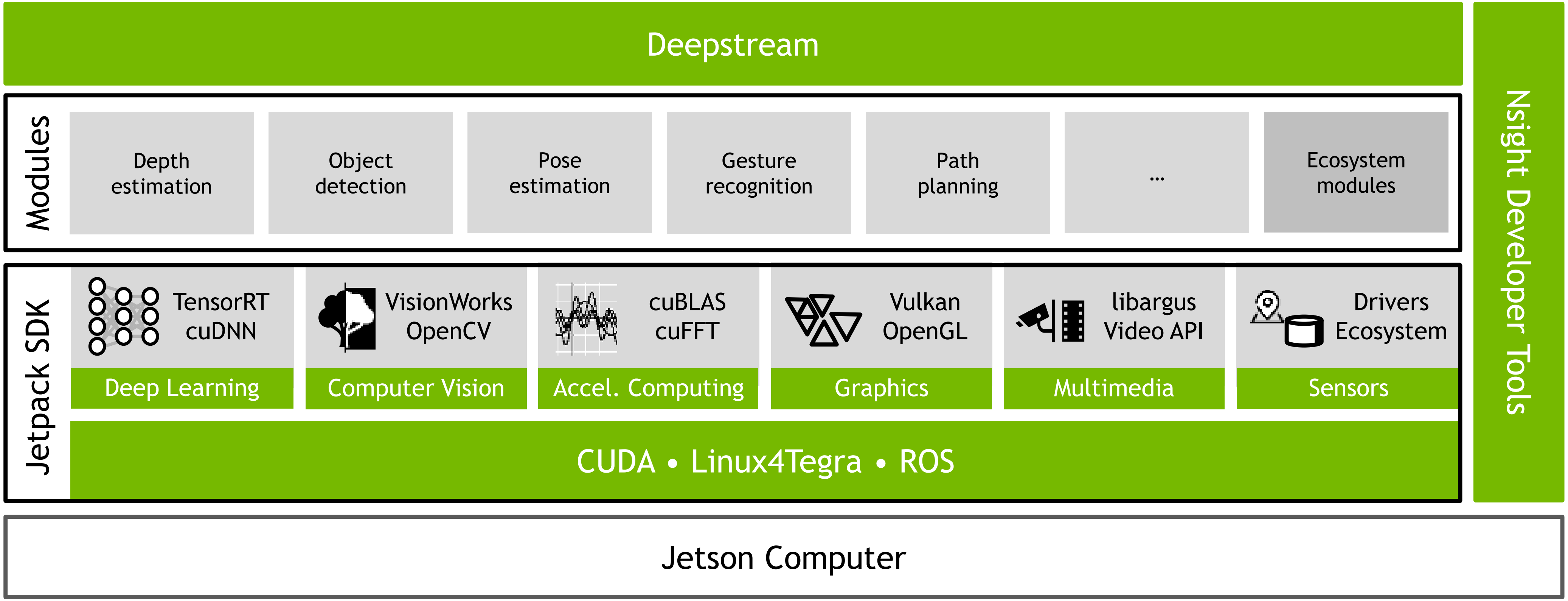
VIRTUAL GPU: vDWS | vPC | vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

TESLA GPUs & SYSTEMS

TESLA (server) GPU

EDGE COMPUTING: JETSON SOFTWARE



Deepstream

Nsight Developer Tools

Modules

Depth estimation

Object detection

Pose estimation

Gesture recognition

Path planning

...

Ecosystem modules

Jetpack SDK

TensorRT cuDNN

Deep Learning

VisionWorks OpenCV

Computer Vision

cuBLAS cuFFT

Accel. Computing

Vulkan OpenGL

Graphics

libargus Video API

Multimedia

Drivers Ecosystem

Sensors

CUDA • Linux4Tegra • ROS

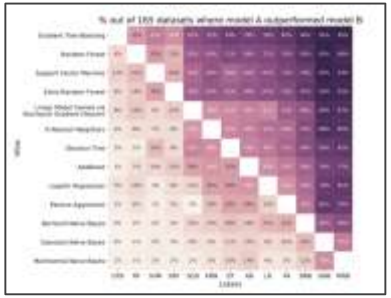
Jetson Computer

Jetson software: developer.nvidia.com/jetson

NVIDIA GPU USE CASES ON AZURE

HPC

Deep Learning



Machine Learning

Virtual Graphics

Automotive/Manufacturing

Oil & Gas

Financial Services

Life Science/Healthcare

Government

Retail (IVA*)

*Intelligent Video Analytics

GPU-ACCELERATED APPLICATIONS

600+ Applications from ISVs

COMP. FINANCE

16 apps

Including:
- O-Quant Options Pricing
- MUREX
- MISYS

CLIMATE & WEATHER

4 apps

Including:
- Cosmos
- Gales
- WRF

DATA SCI. & ANALYTICS

27 apps

Including:
- MapD
- Kinetica
- Graphistry

DEEP LEARNING

36 apps

Including:
- Caffe2
- MXNet
- Tensorflow

FEDERAL & DEFENSE

15 apps

Including:
- ArcGIS Pro
- EVNI
- SocetGXP

MFG, CAD, & CAE

129 apps

Including:
- Ansys Fluent
- Abaqus SIMULIA
- AutoCAD
- CST Studio Suite

MEDIA & ENT.

148 apps

Including:
- DaVinci Resolve
- Premiere Pro CC
- Redshift Renderer

MEDICAL IMAGING

20 apps

Including:
- Gaussian
- VASP
- AMBER
- HOOMD-Blue
- GAMESS

OIL & GAS

19 apps

Including:
- RTM
- SPECFEM 3D

RESEARCH: HER AND SC

126 apps

Including:
- Amber
- MILC
- NAMD
- Relion
- VASP

SAFETY & SECURITY

24 apps

Including:
- Cyllance
- FaceControl
- Syndex Pro

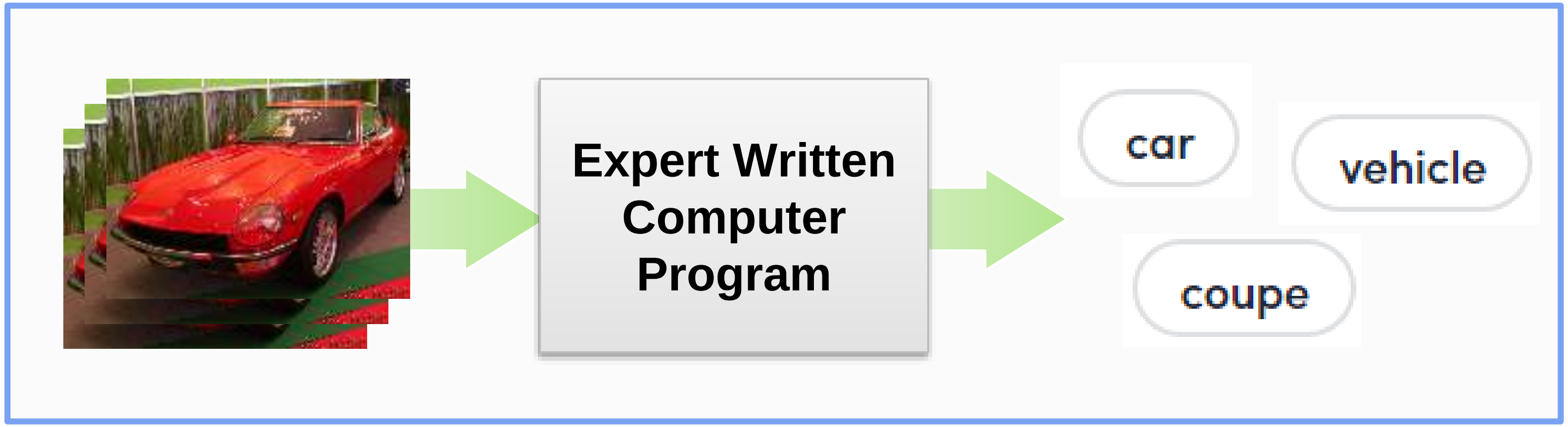
TOOLS & MGMT.

16 apps

Including:
- Bright Cluster Manager
- HPCtoolkit
- Vampir

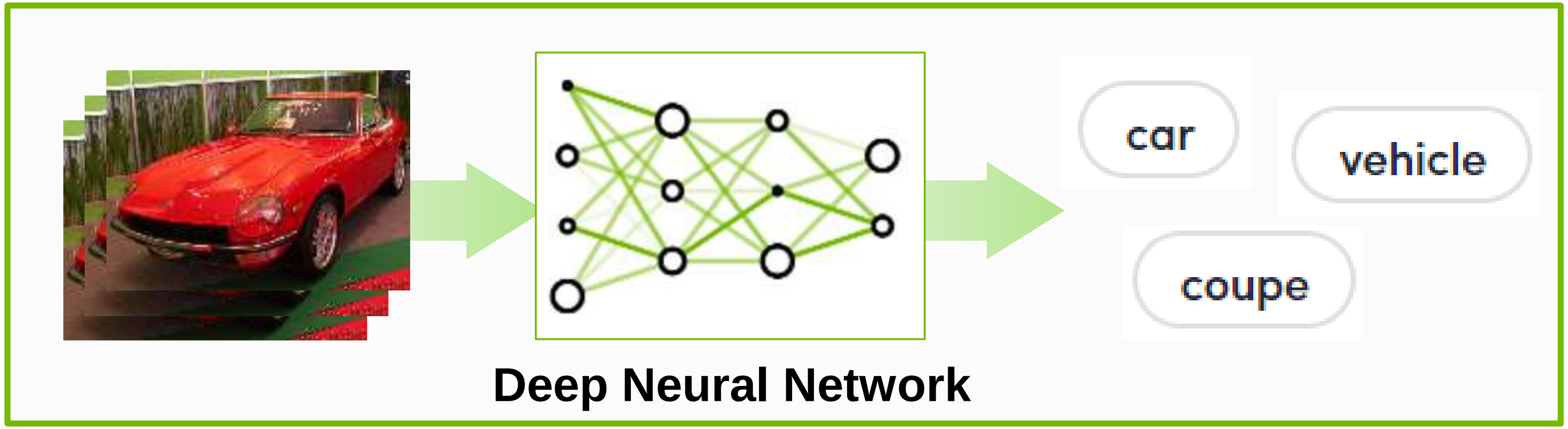
DEEP LEARNING - A NEW COMPUTING MODEL

Algorithms that Learn from Examples



Traditional Approach

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



Deep Neural Network

Deep Learning Approach

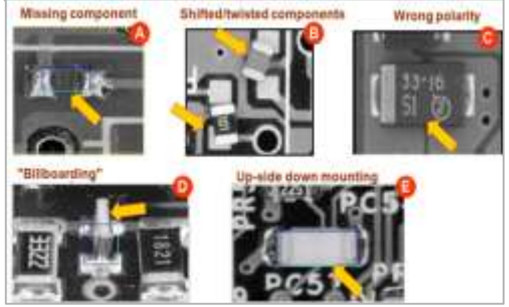
- ✓ Learn from data
- ✓ Easily to extend
- ✓ Speedup with GPUs

INTELLIGENT VIDEO ANALYTICS (IVA) FOR EFFICIENCY AND SAFETY

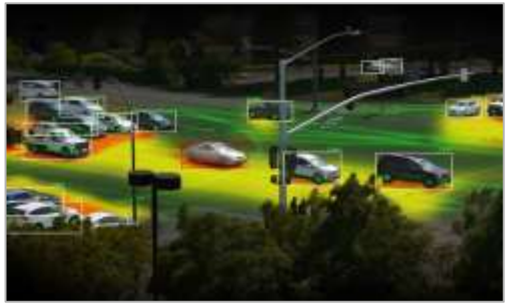
Access Control

Public Transit

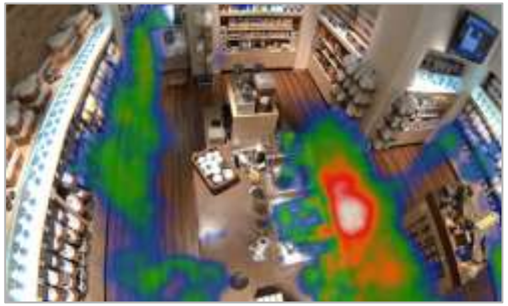
Industrial Inspection



Traffic Engineering



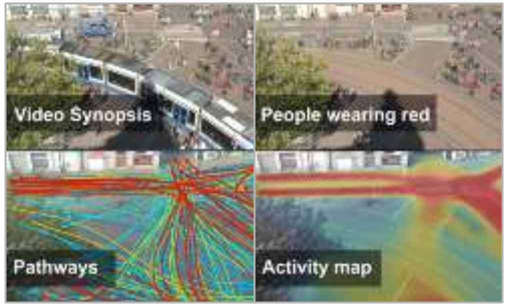
Retail Analytics



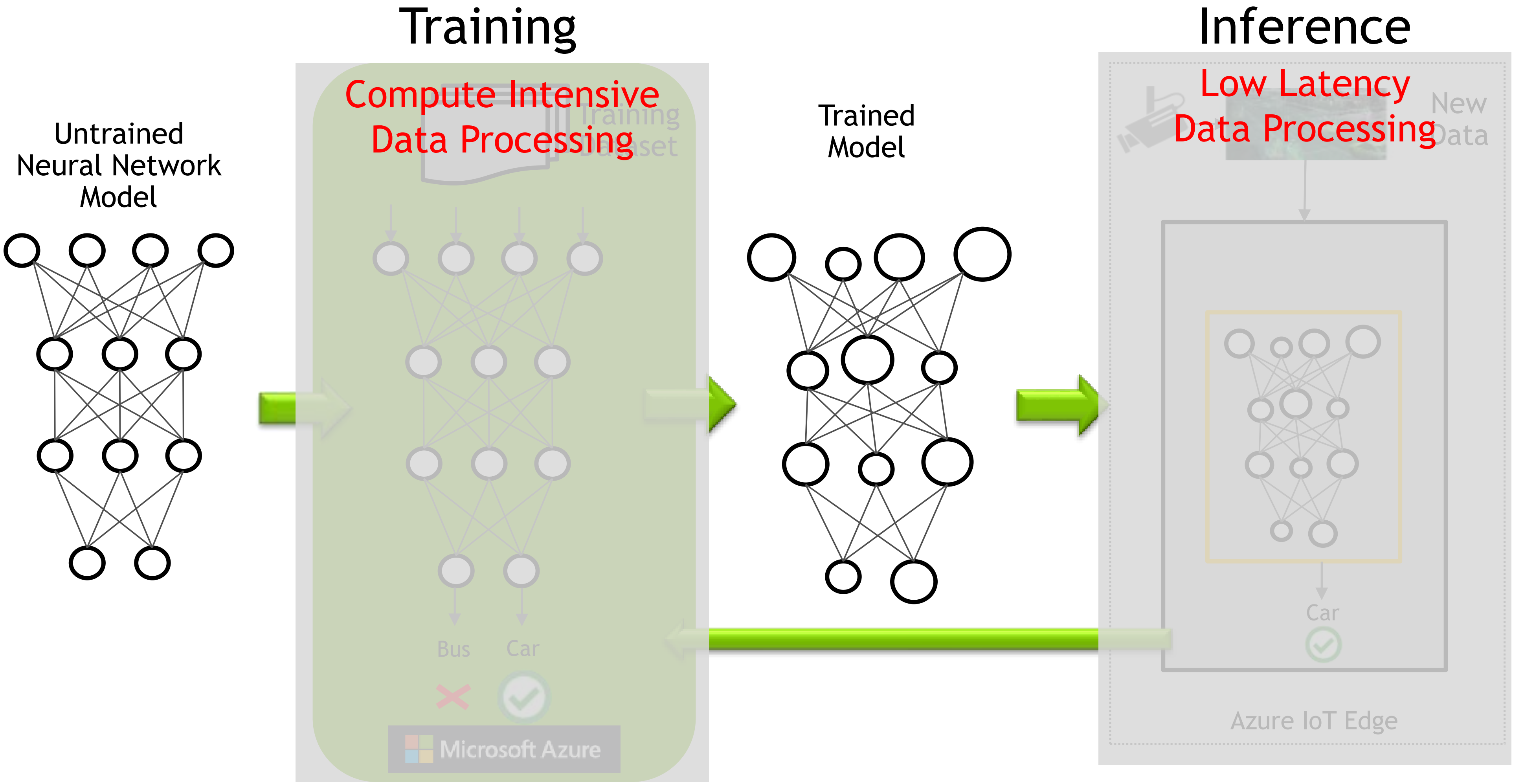
Logistics

Critical Infrastructure

Public Safety



DEEP LEARNING AND IOT EDGE



Untrained Neural Network Model

Training

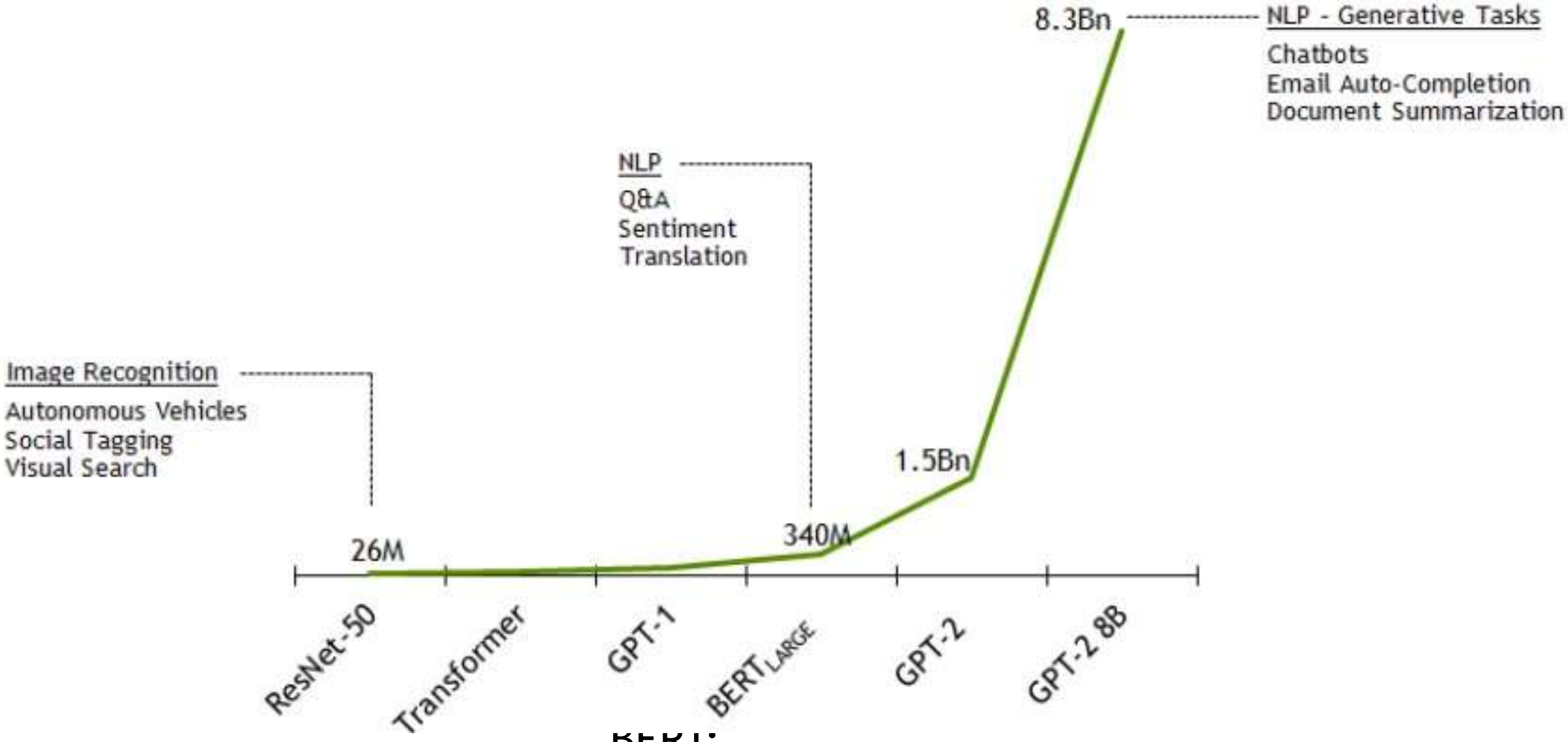
Compute Intensive Data Processing

Trained Model

Inference

Low Latency Data Processing

NATURAL LANGUAGE PROCESSING (NLP)



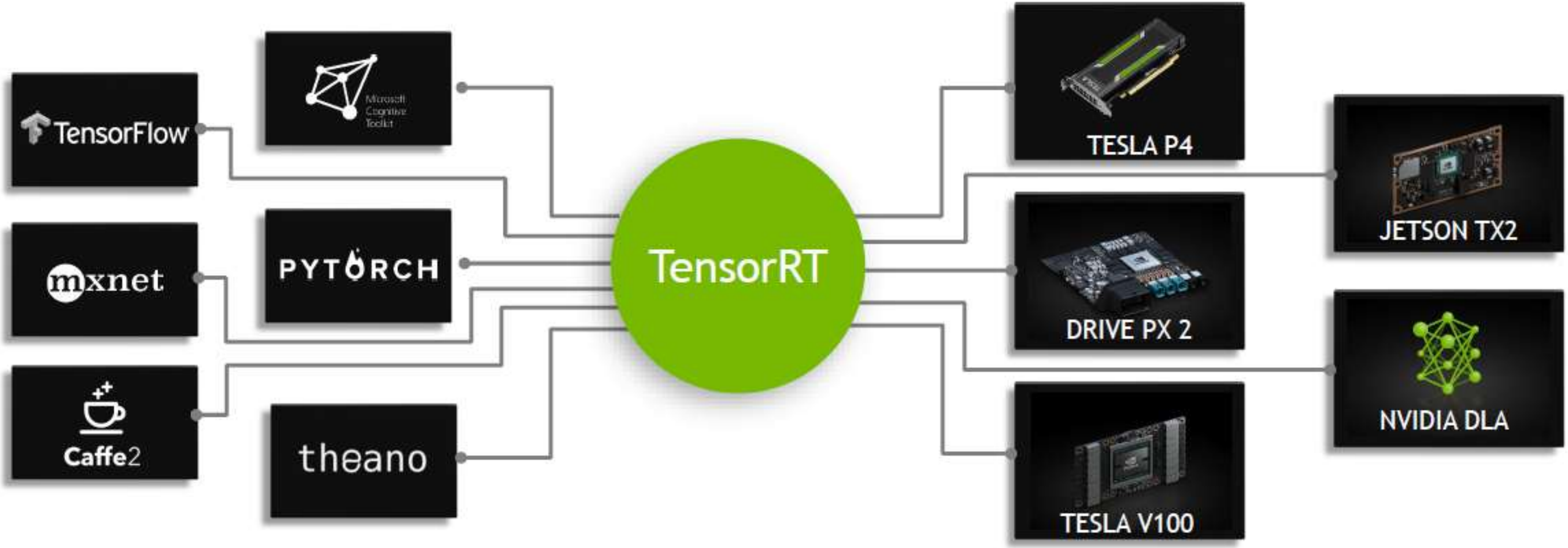
BERT:
Bidirectional Encoder Representations from Transformers

https://youtu.be/Wxi_fbQxCM0

NVIDIA AI INFERENCE PLATFORM

NVIDIA TENSORRT

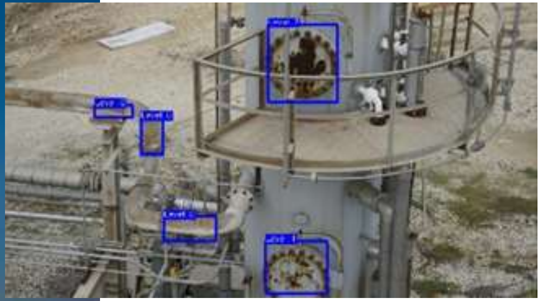
From Every Framework, Optimized For Each Target Platform

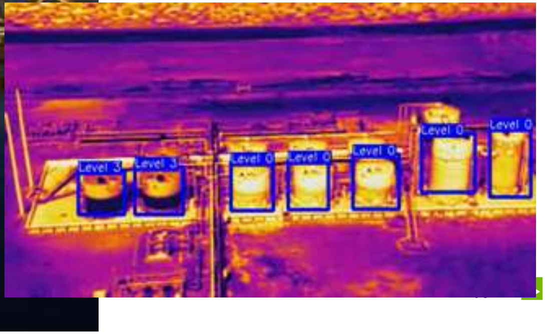


REAL-TIME EDGE COMPUTING REQUIRED

Drivers for AI Inference in Edge Environments

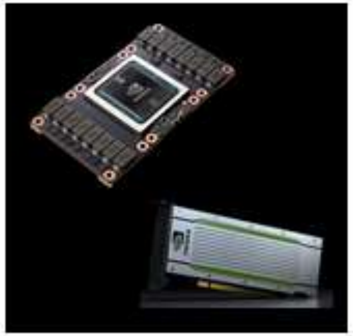
1. Low latency
2. Bandwidth constraints
3. Data sovereignty





CLOUD TO EDGE COMPUTING

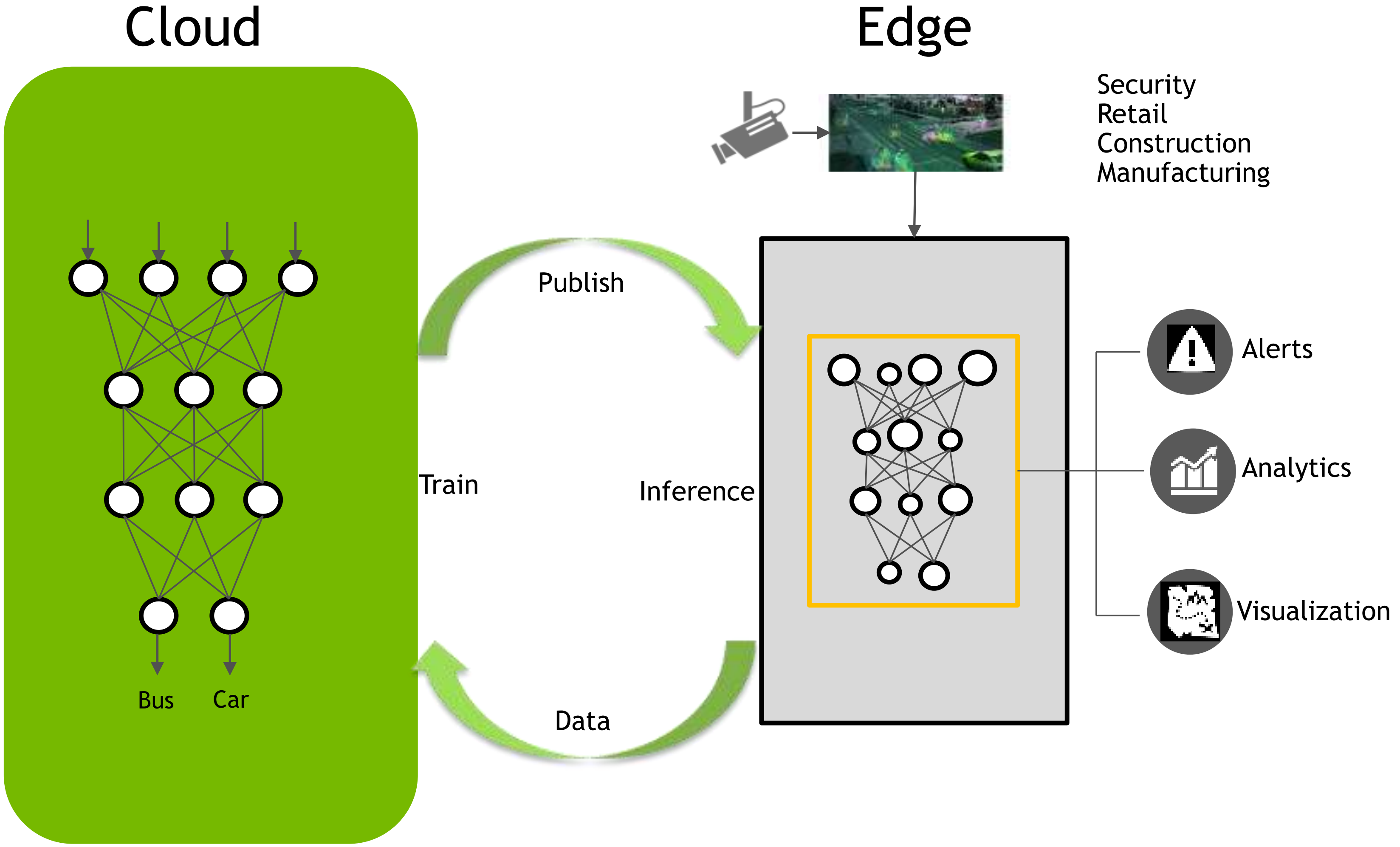
EGX is a scalable edge computing platform

Data Center	Edge Micro Data Centers	Edge Servers	Edge Miniservers	Edge Microservers	Devices
V100/T4	EGX				AGX
					

High Performance Servers — Mainstream Compute Servers — Nano-GPU

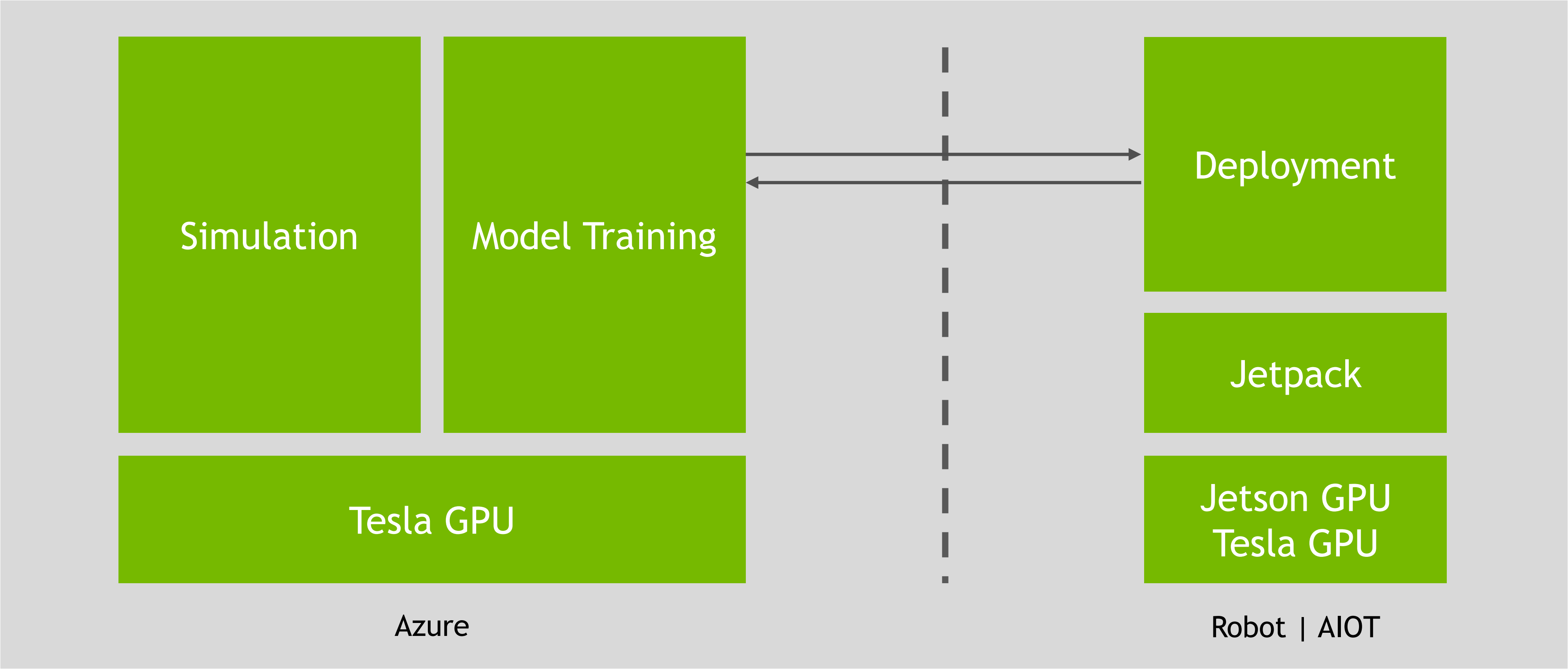
10,000 TOPS — 320 TOPS — 0.5 TOPS

DEEP LEARNING AND IOT EDGE



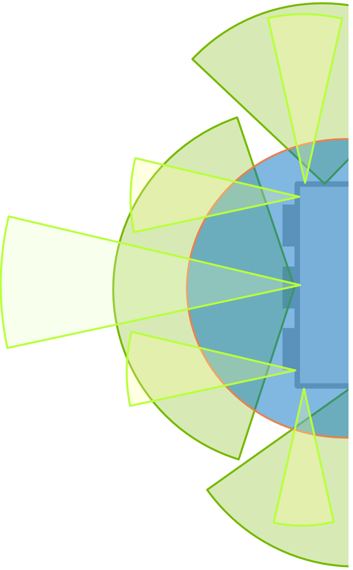
NVIDIA AI PLATFORM

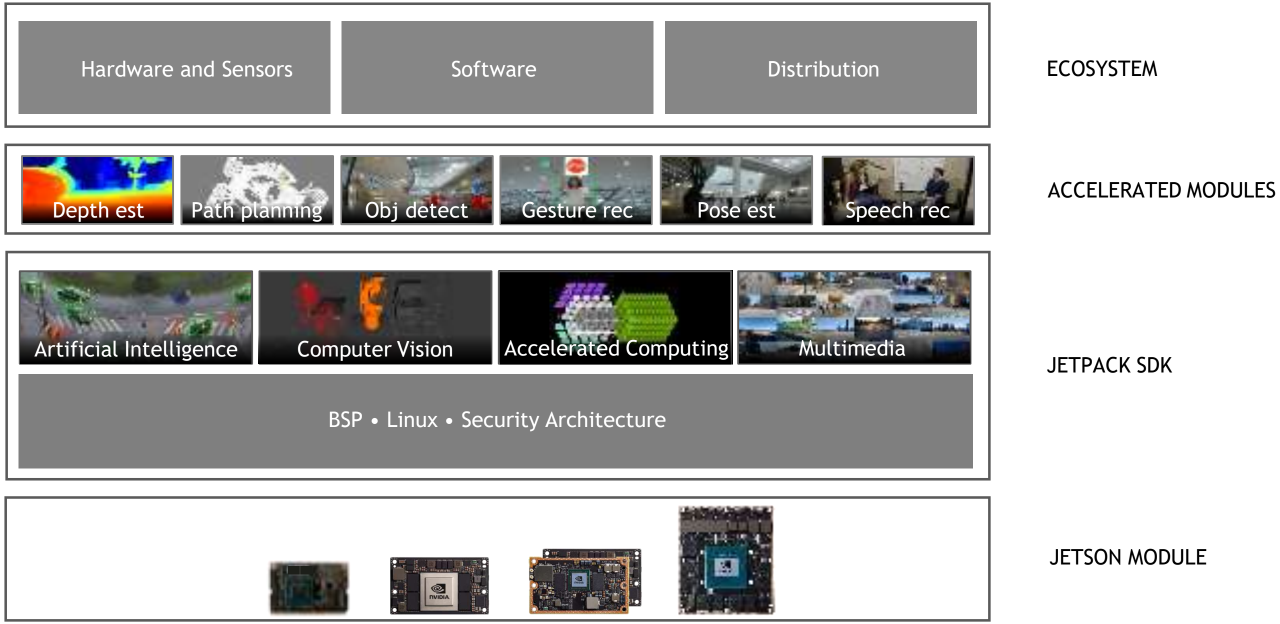
From data center to machines



NVIDIA JETSON
SOFTWARE-DEFINED AUTONOMOUS MACHINES

Powerful and efficient AI, CV, HPC | Rich Software Development Platform
Open Platform | 250K Developers





ECOSYSTEM

Hardware and Sensors | Software | Distribution

ACCELERATED MODULES

Depth est | Path planning | Obj detect | Gesture rec | Pose est | Speech rec

JETPACK SDK

Artificial Intelligence | Computer Vision | Accelerated Computing | Multimedia

BSP • Linux • Security Architecture

JETSON MODULE

THE JETSON FAMILY

From AI at the Edge to Autonomous Machines

JETSON NANO
5 - 10W
0.5 TFLOPS (FP16)
45mm x 70mm
$129

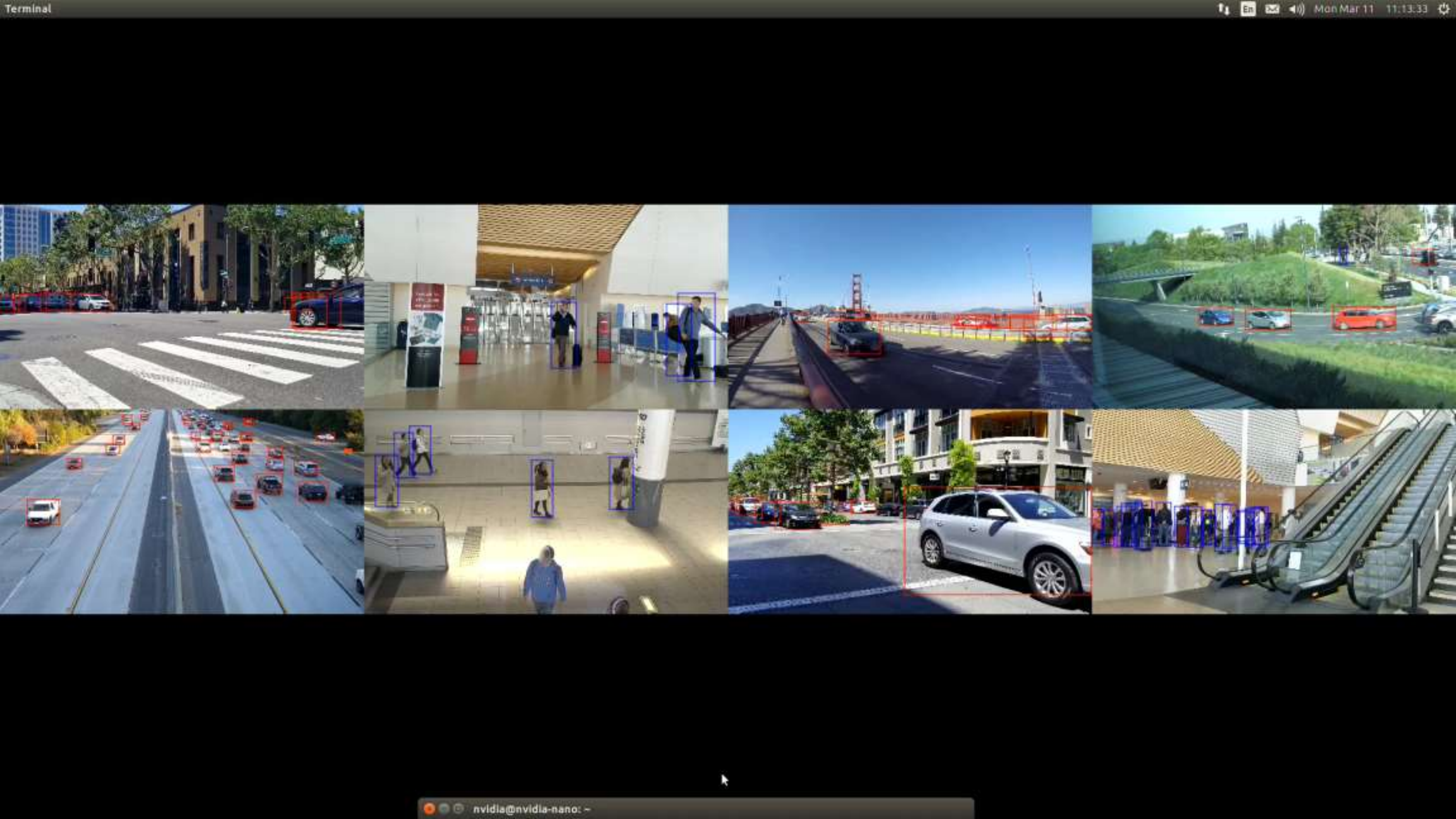
JETSON TX1 → **JETSON TX2 4 GB**
7 - 15W
1 – 1.3 TFLOPS (FP16)
50mm x 87mm
$299

JETSON TX2 8GB | Industrial
7 – 15W
1.3 TFLOPS (FP16)
50mm x 87mm
$399 - $749

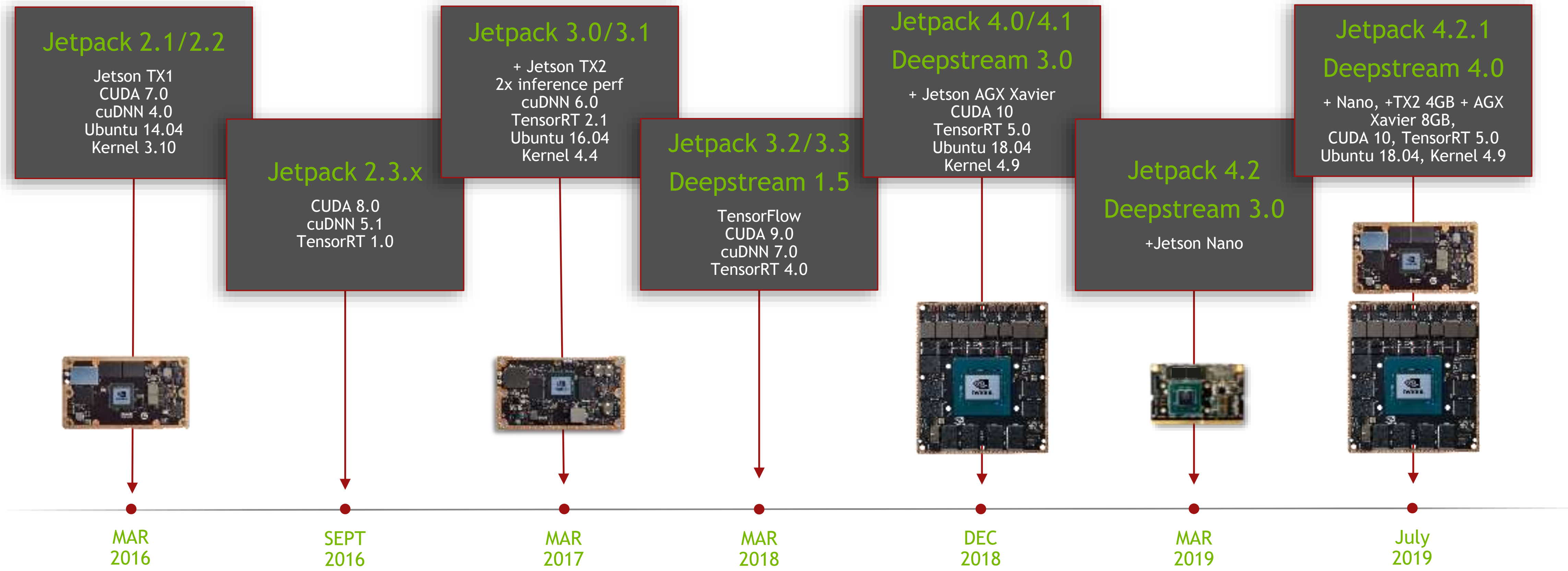
JETSON AGX XAVIER
10 – 30W
10 TFLOPS (FP16) | 32 TOPS (INT8)
100mm x 87mm
$1099

AI at the edge

Fully autonomous machines


Terminal
En
Mon Mar 11
11:13:33
nvidia@nvidia-nano: ~

CONTINUOUS SOFTWARE INVESTMENT



Jetpack 2.1/2.2

Jetson TX1
CUDA 7.0
cuDNN 4.0
Ubuntu 14.04
Kernel 3.10

MAR 2016

Jetpack 2.3.x

CUDA 8.0
cuDNN 5.1
TensorRT 1.0

SEPT 2016

Jetpack 3.0/3.1

+ Jetson TX2
2x inference perf
cuDNN 6.0
TensorRT 2.1
Ubuntu 16.04
Kernel 4.4

MAR 2017

Jetpack 3.2/3.3
Deepstream 1.5

TensorFlow
CUDA 9.0
cuDNN 7.0
TensorRT 4.0

MAR 2018

Jetpack 4.0/4.1
Deepstream 3.0

+ Jetson AGX Xavier
CUDA 10
TensorRT 5.0
Ubuntu 18.04
Kernel 4.9

DEC 2018

Jetpack 4.2
Deepstream 3.0

+Jetson Nano

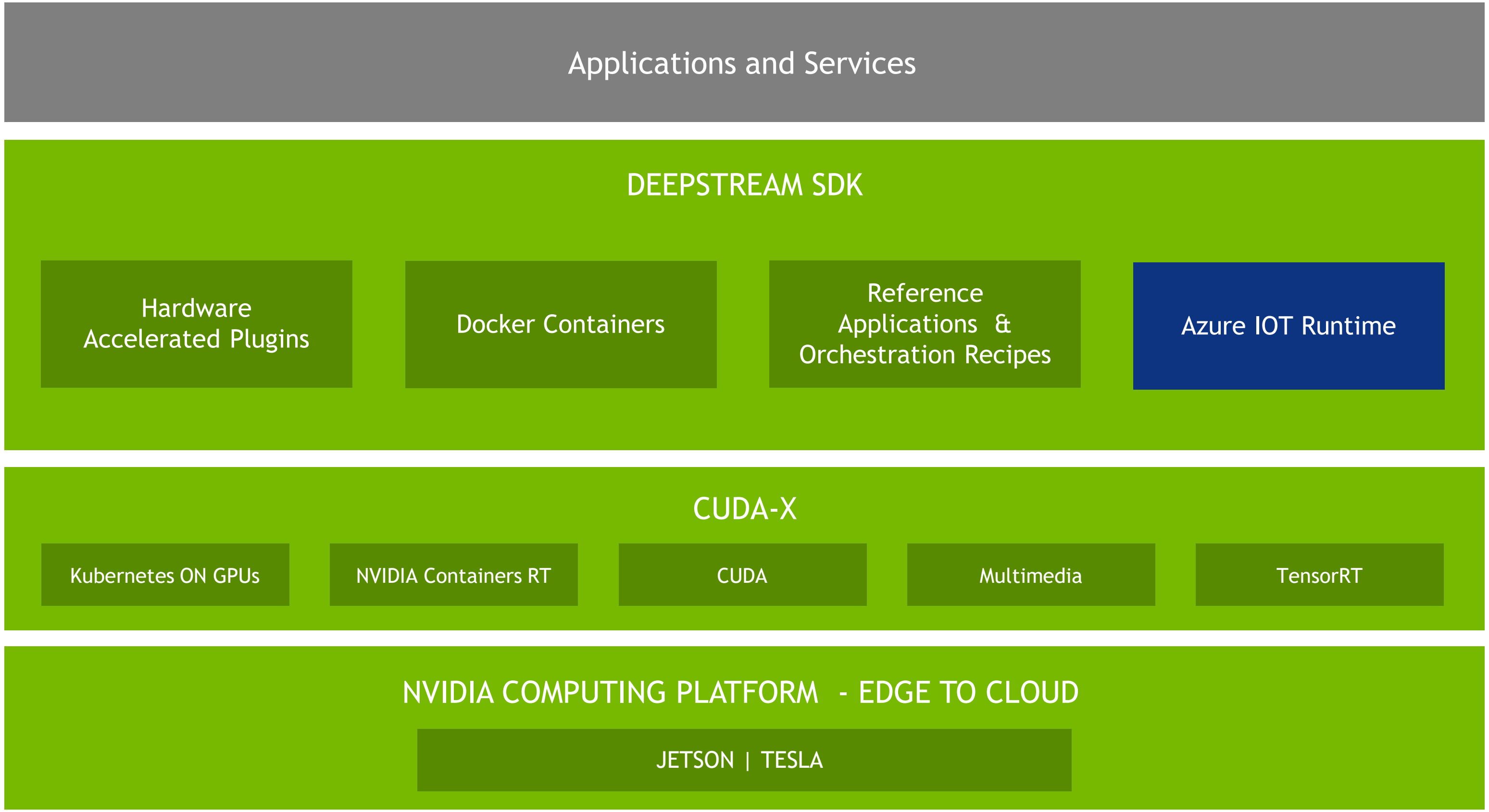
MAR 2019

Jetpack 4.2.1
Deepstream 4.0

+ Nano, +TX2 4GB + AGX Xavier 8GB,
CUDA 10, TensorRT 5.0
Ubuntu 18.04, Kernel 4.9

July 2019

WHAT IS DEEPSTREAM?



Applications and Services

DEEPSTREAM SDK

Hardware Accelerated Plugins

Docker Containers

Reference Applications & Orchestration Recipes

Azure IOT Runtime

CUDA-X

Kubernetes ON GPUs

NVIDIA Containers RT

CUDA

Multimedia

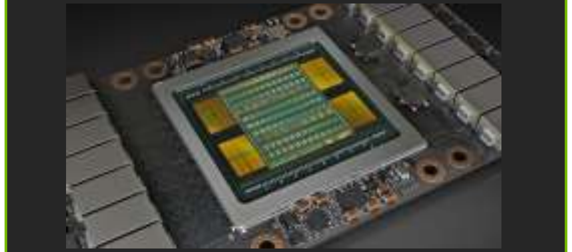
TensorRT

NVIDIA COMPUTING PLATFORM - EDGE TO CLOUD

JETSON | TESLA

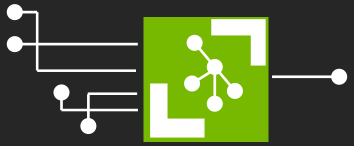
DEEPSTREAM 4.0 KEY FEATURES

UNIFIED SDK , ALL PLATFORMS



Portability from Jetson Nano to T4

TURNKEY IoT INTEGRATION



Microsoft Azure IoT Hub*

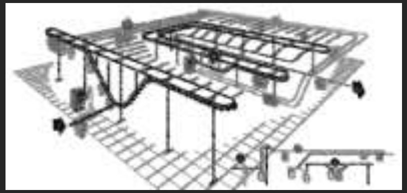
DOCKER CONTAINERS ON NGC

Easy to scale and maintain

MONOCHROME AND JPEG

Enabling Industrial Inspection

SUPPORT FOR IMAGE SEGMENTATION



Enabling Retail & Supply Chain Solutions

PLUGIN SOURCES

Inference

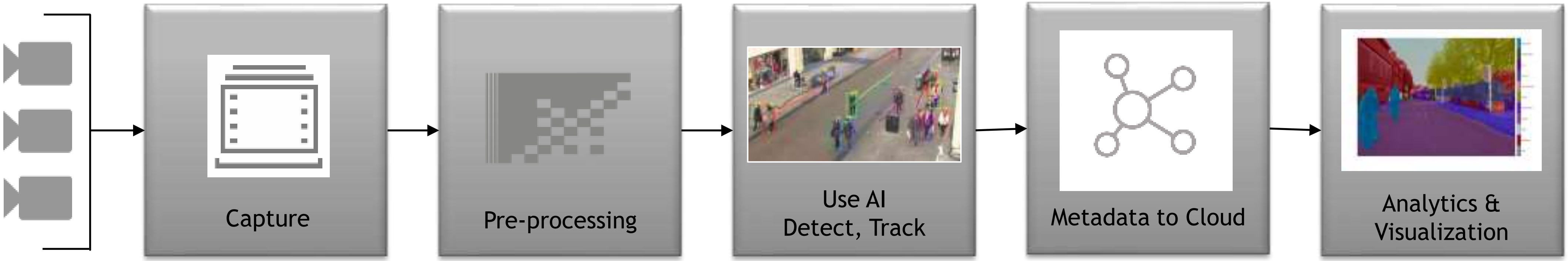
Decode

Messaging

Greater control for your use case

*Containers on Azure Marketplace coming soon. Available on NGC now

IVA APPLICATION WORKFLOW



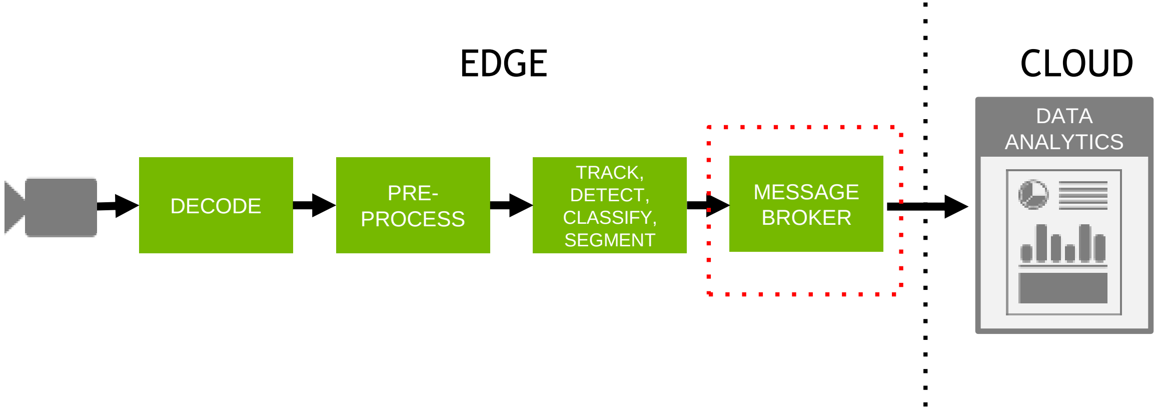
Pixels → Insights

SMART CITIES: INTELLIGENT TRAFFIC SYSTEM

USE CASE

Need to generate actionable insights from 1000s of cameras

SOLUTION

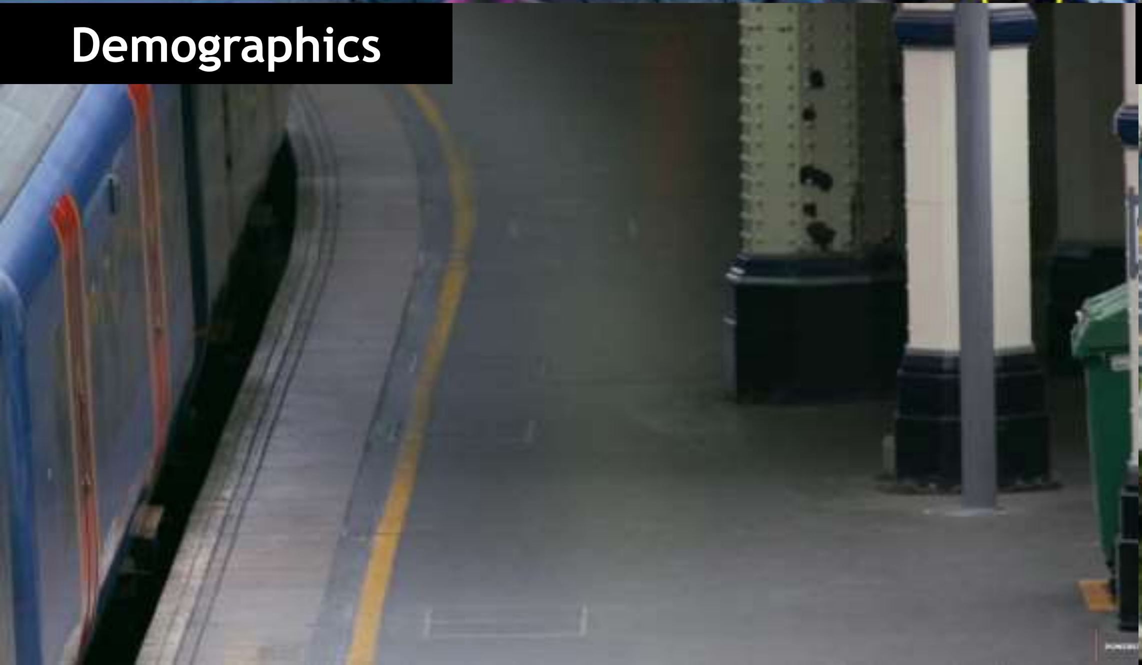


DeepStream offers the ability to seamlessly connect from edge to cloud using the message broker plugin

METROPOLIS

People Tracking

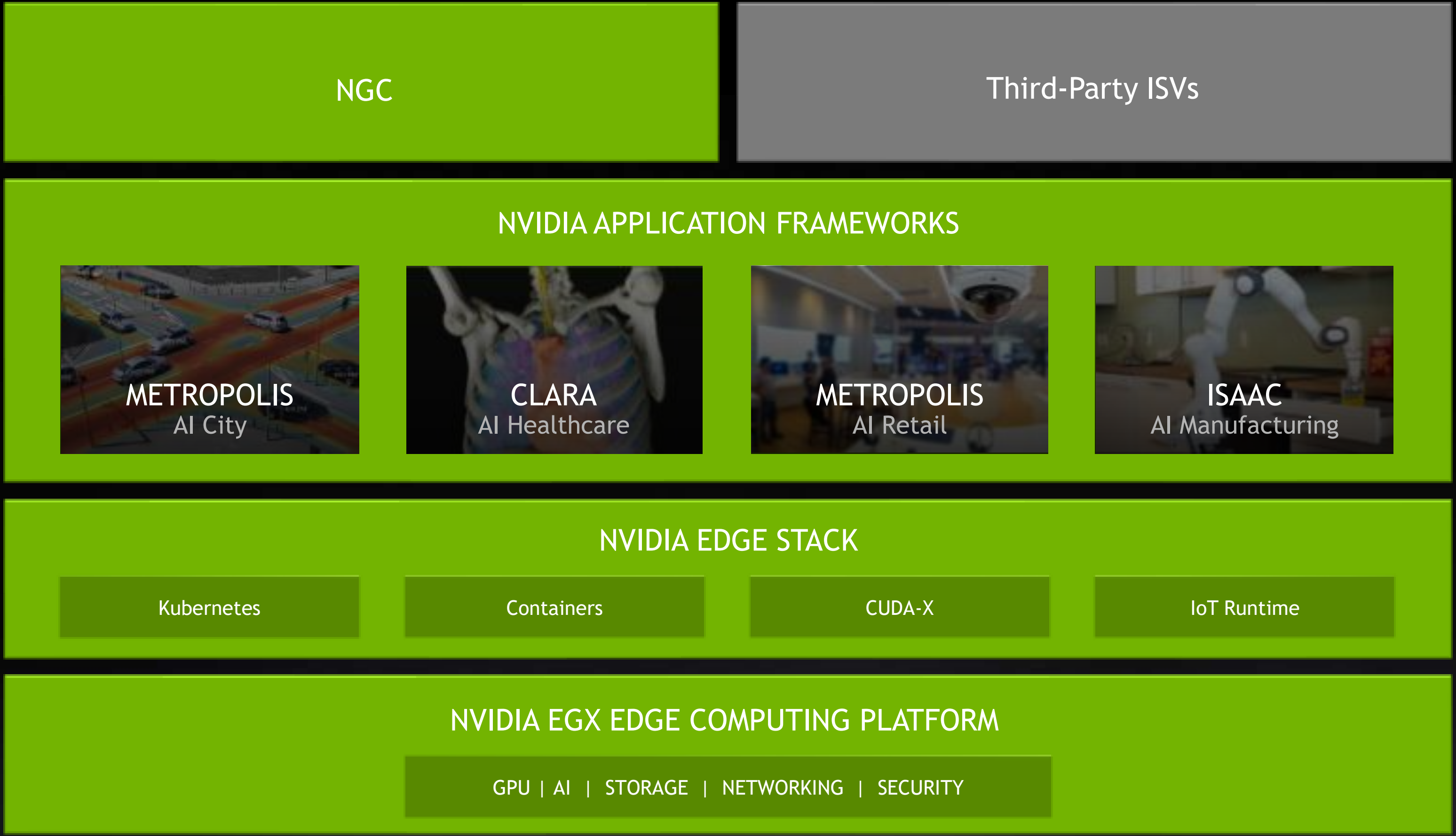
People Tracking



Demographics

Traffic Analytics

NVIDIA EGX EDGE COMPUTING



NGC

Third-Party ISVs

NVIDIA APPLICATION FRAMEWORKS

METROPOLIS
AI City

CLARA
AI Healthcare

METROPOLIS
AI Retail

ISAAC
AI Manufacturing

NVIDIA EDGE STACK

Kubernetes

Containers

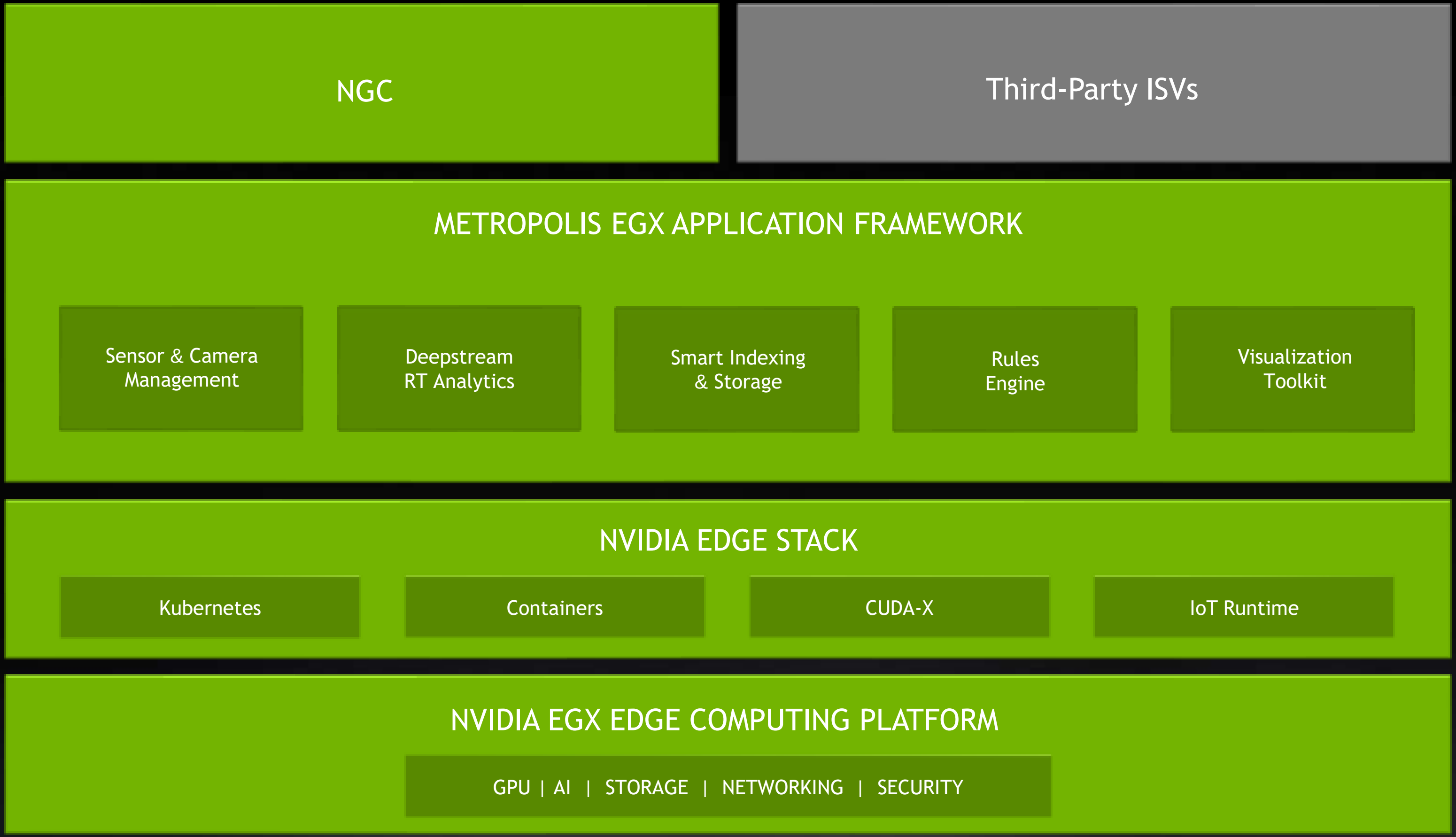
CUDA-X

IoT Runtime

NVIDIA EGX EDGE COMPUTING PLATFORM

GPU | AI | STORAGE | NETWORKING | SECURITY

METROPOLIS EGX OPEN AI CITY PLATFORM



NGC

Third-Party ISVs

METROPOLIS EGX APPLICATION FRAMEWORK

Sensor & Camera Management

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

SMART CITY TARGET AUDIENCES

LAW ENFORCEMENT

Police Departments
Chief Information Officer

AIRPORTS & MASS TRANSIT

Chief Information Officer
Chief Security Offer

SCHOOLS & UNIVERSITIES

Campus Security Authority
Emergency Mgmt Centers
Chief Information Officer

CASINOS & GAMING

Chief Security Officer Casino
Operations and IT

SMART CITY SOLUTIONS

METROPOLIS ISVs	LAW ENFORCEMENT	AIRPORT/MASS TRANSIT	CASINO & GAMING
AnyVision	Face Recognition & Watch Lists	Access Control & Watch Lists	Customer Service
Athena Security	Weapon Detection	Weapon Detection	Weapon Detection
Briefcam	Crowd Management & Crowd Safety	Crowd/Queue Management & Retail Analytics	Retail Analytics & Queue Management
Deepvision	Vehicle/Pedestrian Analysis	Vehicle/Pedestrian Analysis	Vehicle/Pedestrian Analysis
Irvine Sensors	Pedestrian/Left Baggage	Pedestrian/Left Baggage	Game Analytics
openALPR	License Plate Recognition	License Plate Recognition	License Plate Recognition
Vintra	Video Search	Video Search	Video Search

TOP AI RETAIL USE CASES

LOSS PREVENTION

Ticket Switching
Mis-scanning
Employee Theft
Security

STORE ANALYTICS

Heat Mapping
Demographic Analysis
Shopper/Employee Tracking
Stockout
Customer Engagement

AUTONOMOUS SHOPPING

Autonomous Checkout
Nano Stores
Smart Cabinets

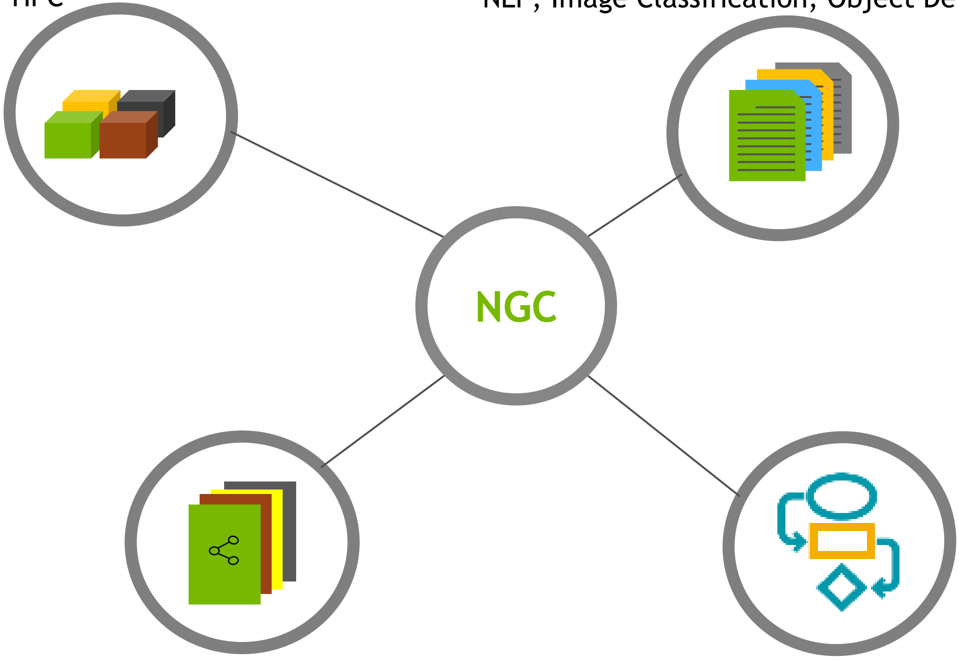
LOSS PREVENTION SOLUTIONS

METROPOLIS ISVs	LOSS PREVENTION	SECURITY & SURVEILLANCE
AnyVision	✓	✓
Briefcam	✓	✓
Everseen	✓	
Third Eye Labs	✓	
Malong	✓	
Sunrise Technology	✓	
Ntechlab	✓	✓

NGC - NVIDIA GPU CLOUD

NGC: GPU-OPTIMIZED SOFTWARE HUB

Simplifying DL, ML and HPC Workflows



50+ Containers

DL, ML, HPC

15+ Model Training Scripts

NLP, Image Classification, Object Detection & more

60 Pre-trained Models

NLP, Image Classification, Object Detection & more

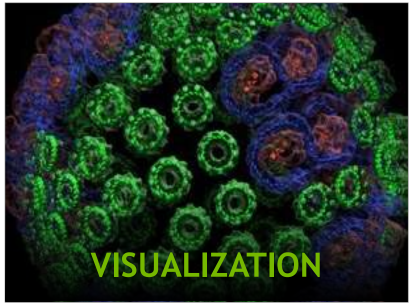
Industry Workflows

Medical Imaging, Intelligent Video Analytics

TensorFlow | PyTorch | more

RAPIDS | H2O | more

NAMD | GROMACS | more



ParaView | IndeX | more

NVIDIA DEEP LEARNING INSTITUTE

Online self-paced labs and instructor-led workshops on deep learning and accelerated computing

www.nvidia.com/dli

Talk to Microsoft or NVIDIA (Uli) and ask for hands-on instructor-led Deep Learning Institute (DLI)

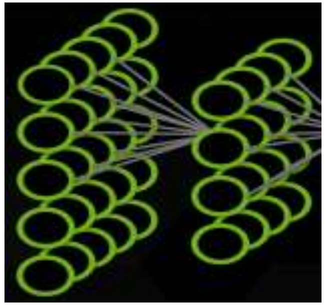
Caffe2

Microsoft Microsoft Cognitive Toolkit

mxnet

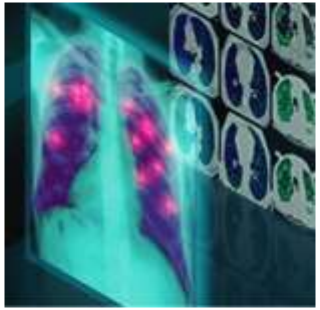
TensorFlow

PYTORCH



Fundamentals

Autonomous Vehicles

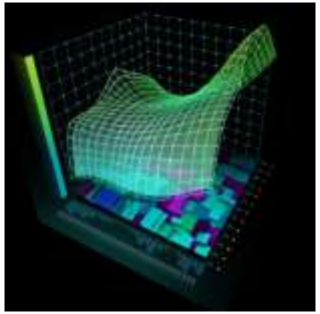


Healthcare

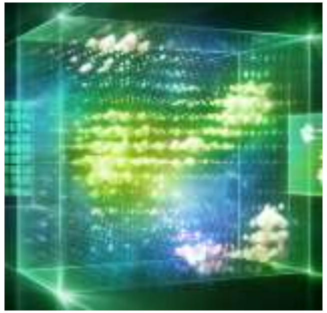
Intelligent Video Analytics

Robotics

Game Development & Digital Content



Finance



Accelerated Computing

Virtual Reality

IMPORTANT SOURCES

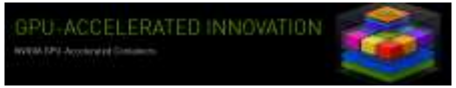
Contact

Uli Knechtel, uknechtel@nvidia.com
+49 162 1034441

Developer portal

NVIDIA. DEVELOPER developer.nvidia.com/

NGC (NVIDIA GPU Cloud)



www.nvidia.com/ngc

NVIDIA®

BACKUP

TESLA V100 TENSOR CORE GPU

World’s Most Powerful Data Center GPU

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s | 300GB/s NVLink

AZURE GPU-ACCELERATED PLATFORMS

Platform	All VMs	Windows Virtual Desktop	Virtual Desktop Workstation	Batch AI	Databricks	Azure ML	HDInsight	AKS, ACI
GPU Support	NDv2 - V100* ND - P40 NCv3 - V100 NCv2 - P100 NCv1 - K80 NVv2 – M60	NC, NV series	All N-Series except NCv3	NC, ND series	NC series	NCv2, ND and NDv2	NCv2, ND and NDv2	All N-Series
Use-case	Infra for all use-cases	Remote apps on Cloud	Proviz apps, 3D graphics	AI training, scheduling, hybrid	ML, DL/AI, Big Data, Spark on Cloud	Inferencing (ONNX runtime) ML, Data pipeline (RAPIDS)	Hadoop/ Big Data on Cloud	Containers, Orchestration for HPC, DL, ML and Visualization

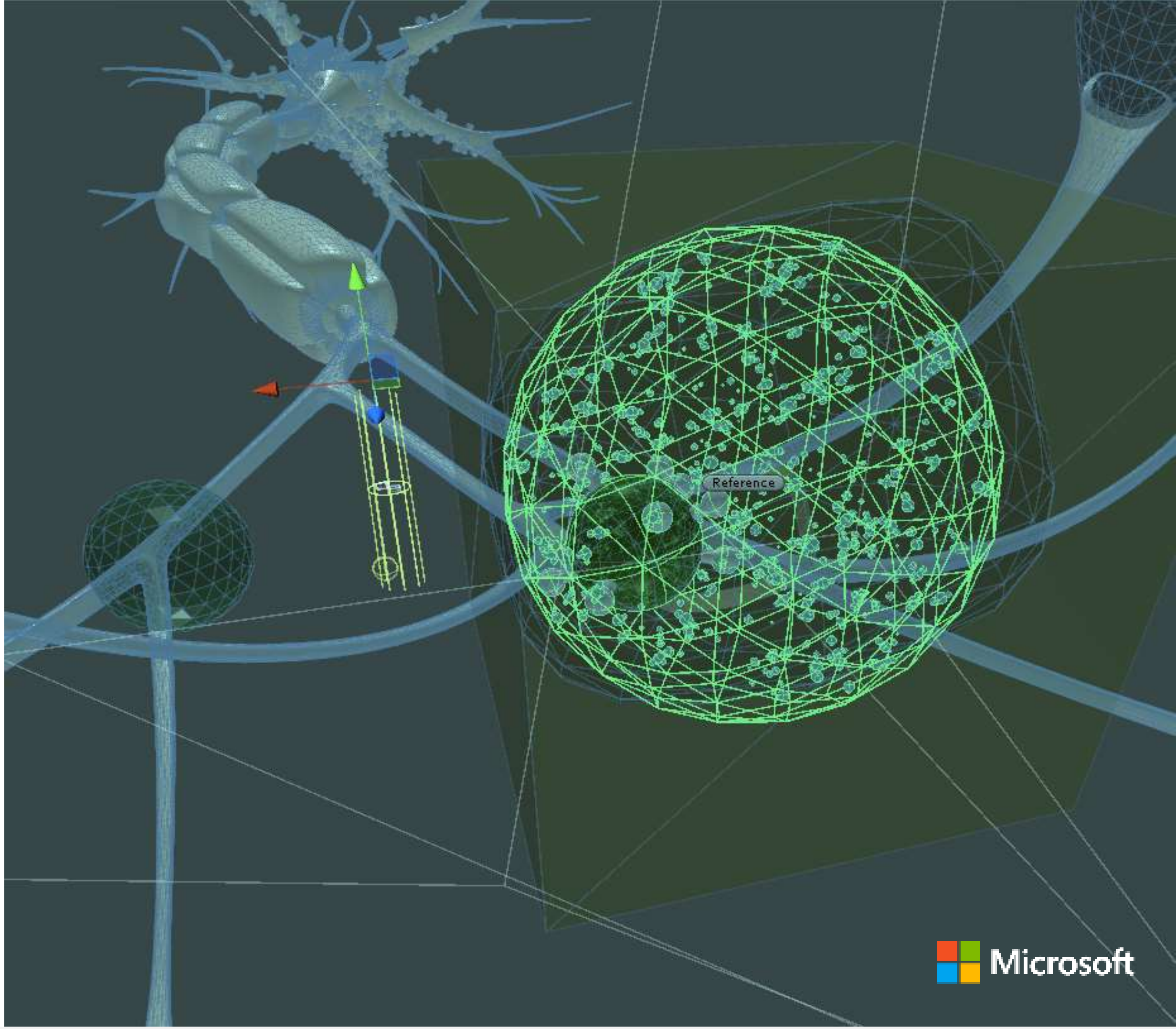
* is SXM2

	JETSON NANO	JETSON TX2 SERIES (TX2, TX2 4GB AND TX2i*)	JETSON AGX XAVIER SERIES AGX XAVIER 8GB AND AGX XAVIER	
GPU	128 Core Maxwell 0.5 TFLOPs (FP16)	256 Core Pascal TX2 & TX2 4GB 1.33 TFLOPS (FP16) TX2i 1.26 TFLOPS (FP16)	384 Core Volta + NVDLA 5.5 TFLOPS (FP16) 11.1 TOPS (INT8)	512 Core Volta + NVDLA 11 TFLOPS (FP16) 22 TOPS (INT8)
CPU	Quad-core ARM A57 (1.5 GHz)	Dual-core Denver and Quad-core A57 2GHz (2x) 2MB L2	6-core Carmel ARM CPU 1.3GHz (3x) 2MB L2 + 4MB L3	8-core Carmel ARM CPU 2.26GHz (4x) 2MB L2 + 4MB L3
DLA	-	-	4.1 TFLOPS (FP16) 8.2 TOPS (INT8)	5 TFLOPS (FP16) 10 TOPS (INT8)
Memory	4 GB 64-bit LPDDR4 29.8 GB/s	128-bit LPDDR4 TX2 60 GB/s, TX2 4GB & TX2i 51 GB/s	8GB 256-bit LPDDR4x 1333MHz - 85 GB/s	16GB 256-bit LPDDR4x 2133MHz - 137 GB/s
Storage	16 GB eMMC 5.1	TX2 4GB16 GB eMMC 5.1 TX2 & TX2i 32 GB eMMC 5.1	32 GB eMMC 5.1	
Video Encode	1x4K @30|2x1080p @60 |4x1080p @30(HEVC)	1x4K @60|3x4K @30|4x1080p @60 | 8x1080p @30(HEVC)	2x4K @60|6x4K @30|9x1080p @60 | 14x1080p @ 30(HEVC)	4x4K @60|8x4K @30|16x1080p @60 | 32x1080p @30 (HEVC)
Video Decode	1x4K @60|2x4K @30 | 4x 1080p @60 |8x1080p @30 (HEVC)	2x4K @60|4x4K @30|7x1080p @60 | 14x1080p @30(HEVC)	2x4K @60|4x4K @30|12x1080p @60 | 24x1080p @30 (HEVC) | 16x 1080p @ 30 (H.264)	2x8K @30|6x4K @60|12x4K @30 | 26x1080p @60|52x1080p @30 (HEVC) | 30x1080p @30 (H.264)
Camera	12 lanes (3x4|4x2) MIPI CSI-2 D-PHY 1.1 lanes (1.5 Gbps)	12 lanes (3x4|6x2) MIPI CSI-2 D-PHY 1.2 lanes (2.5Gbps)	16 lanes (4x4|6x2|6x1) MIPI CSI-2 | 8 lanes SLVS-EC D-PHY 1.2 (2.5Gbps total up to 40 Gbps) C-PHY 1.1 (1.75Gsym/s total up to 64 Gbps)	16 lanes (4x4|6x2|6x1) MIPI CSI-2 | 8 lanes SLVS-EC D-PHY 1.2 (2.5Gbps total up to 40 Gbps) C-PHY 1.1 (2.5Gsym/s total up to 109 Gbps)
Power	5W | 10W	7.5W | 15W	10W | 20W	10W | 15W | 30W
Mechanical	69.6mm x 45mm 260 pin edge connector	87mm x 50mm 400 pin connector	100mm x 87mm 699 pin connector	
Software	Jetpack SDK - Unified software release across all Jetson products			

*i = for industrial environments

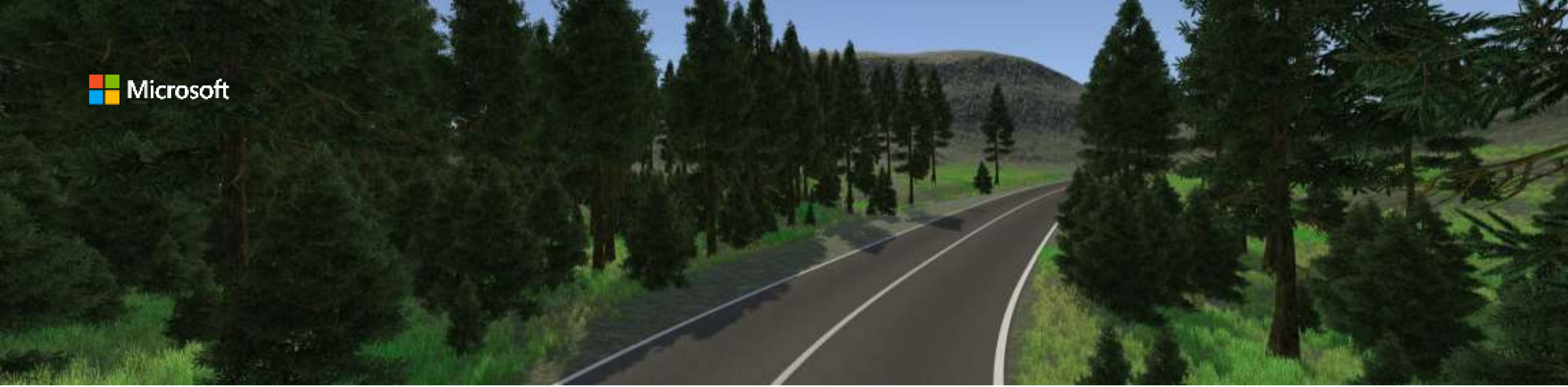
Cure for Alzheimer's and Parkinson's draws closer with neuron simulation boosted by cloud-based GPUs

Neurological disorders such as Alzheimer's and Parkinson's diseases afflict millions of people worldwide, yet no known cure is in sight. Biotech startup NeuroInitiative is working to change that by harnessing NVIDIA graphics processing units (GPUs) in Microsoft Azure to run neuron pathway simulations faster. With its high-performance computing simulation tool, NeuroInitiative is hopeful that it can cut today's 12-to-20-year drug development period in half.



NeuroInitiative
Biotech to Cure Disease

Products and Services	Organization Size	Industry	Country
Microsoft Azure Azure Storage Azure Virtual Machines NC-series Azure Virtual Network	10 employees	Health Provider	United States



Audi technology partner EFS uses deep learning to analyze roads for self-driving vehicles

Based in Gaimersheim, Germany, EFS is the number one partner of Audi in chassis development. It examines and helps implement future-looking technologies, including automated driving. As part of its research efforts, the company used Azure NC-series virtual machines powered by NVIDIA Tesla P100 GPUs to drive a deep learning AI solution that analyzes high-resolution two-dimensional images of roads. The purpose is to give self-driving vehicles a better understanding of those roads. EFS proved that the concept works, and the company can now move ahead with product development.

EFS ELEKTRONISCHE FAHRWERKSYSTEME

Products and Services	Organization Size	Industry	Country	Partner
Microsoft Azure Azure NC-series VMs Azure storage	422 employees	Professional Services	Germany	NVIDIA

Diagnostic services provider uses Azure Machine Learning with NVIDIA GPUs to help end preventable blindness

Diabetes is the leading cause of preventable blindness in the United States, but there was no easy way to diagnose diabetic vision damage through primary care providers. That’s why IRIS used Microsoft Azure to help create a platform that can identify diabetic retinopathy before patients suffer from vision loss. Using Azure Machine Learning Package for Computer Vision, the IRIS platform processes images quickly and accurately so doctors can share data with patients and other clinicians, better prevent diabetic blindness, and help reduce healthcare costs.

iris™
Intelligent Retinal Imaging Systems

Products and Services	Organization Size	Industry	Country
Microsoft Azure Azure Functions Azure Machine Learning Azure Service Bus Azure SQL Database	34 employees	Health Provider	United States

NVIDIA®