Diversity and Inclusion

We strive to host inclusive, accessible events that enable all individuals to engage fully.

If you have a concern, if you need an accommodation for a disability, or if you would like to learn more, please talk to one of our Microsoft staff on-site and we'll do our best to assist you.

Thank you.

Practice
Awareness

Exercise
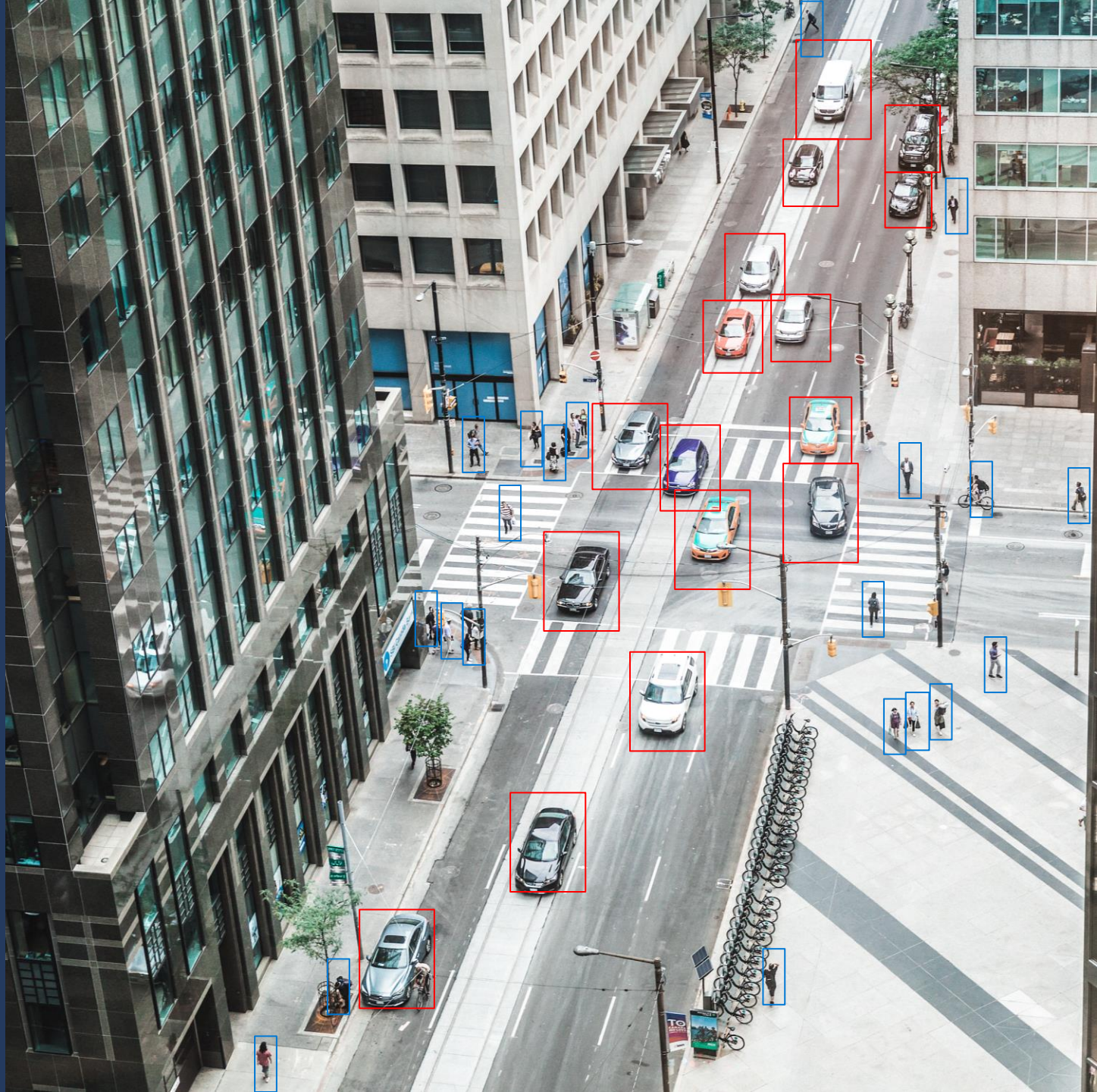**Curiosity**

Demonstrate
Courage

GPU Accelerated IoT Workloads at the Edge

Paul DeCarlo
Principal Cloud Advocate
@pjdecarlo

Presentation available @
aka.ms/intelligentedgeholdeck

Demo

YoloV3 IoT Edge Module Running on Jetson Xavier

YoloV3 IoT Edge Module Running on Jetson Xavier

Agenda

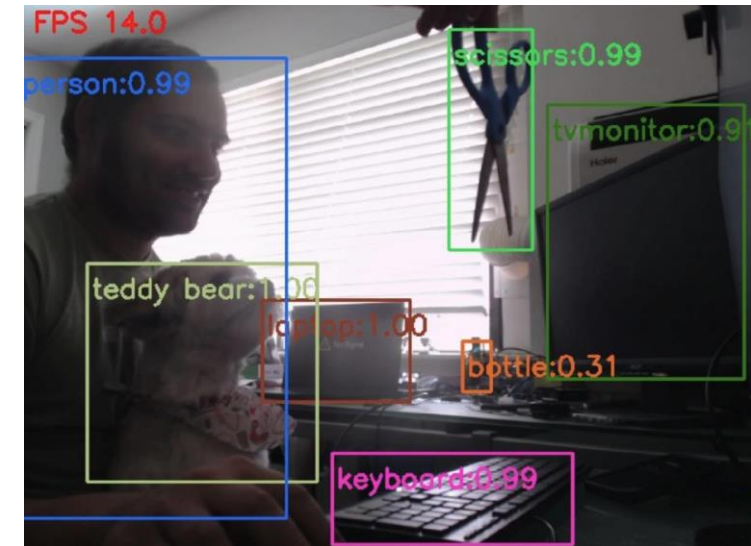Introduce Azure IoT Edge

Introduce Nvidia Jetson Devices

Developing IoT Edge Modules with GPU acceleration
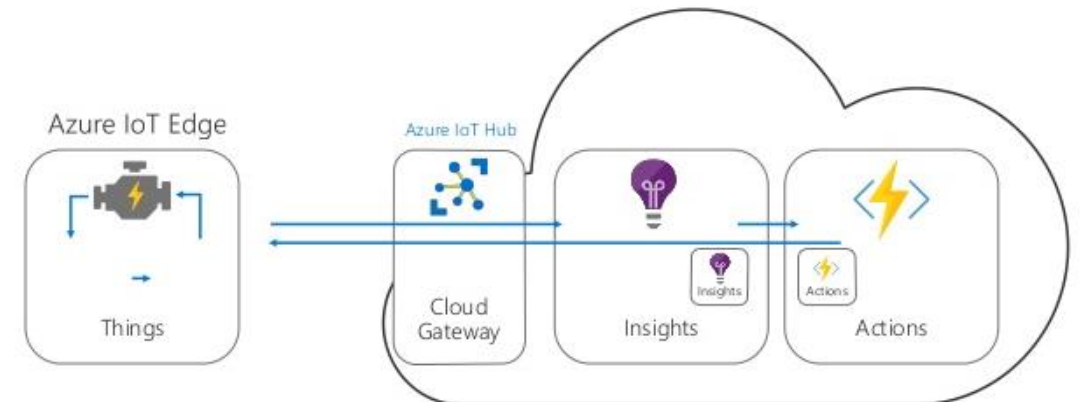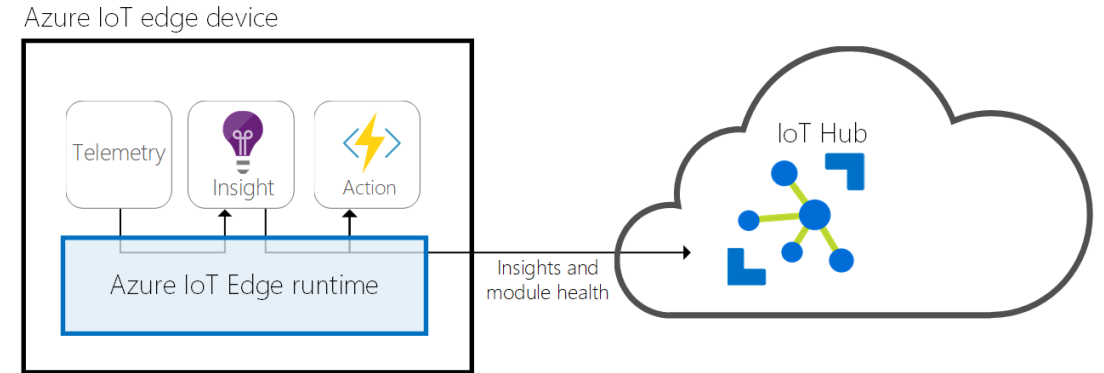
Intelligent Edge Hands-On-Lab

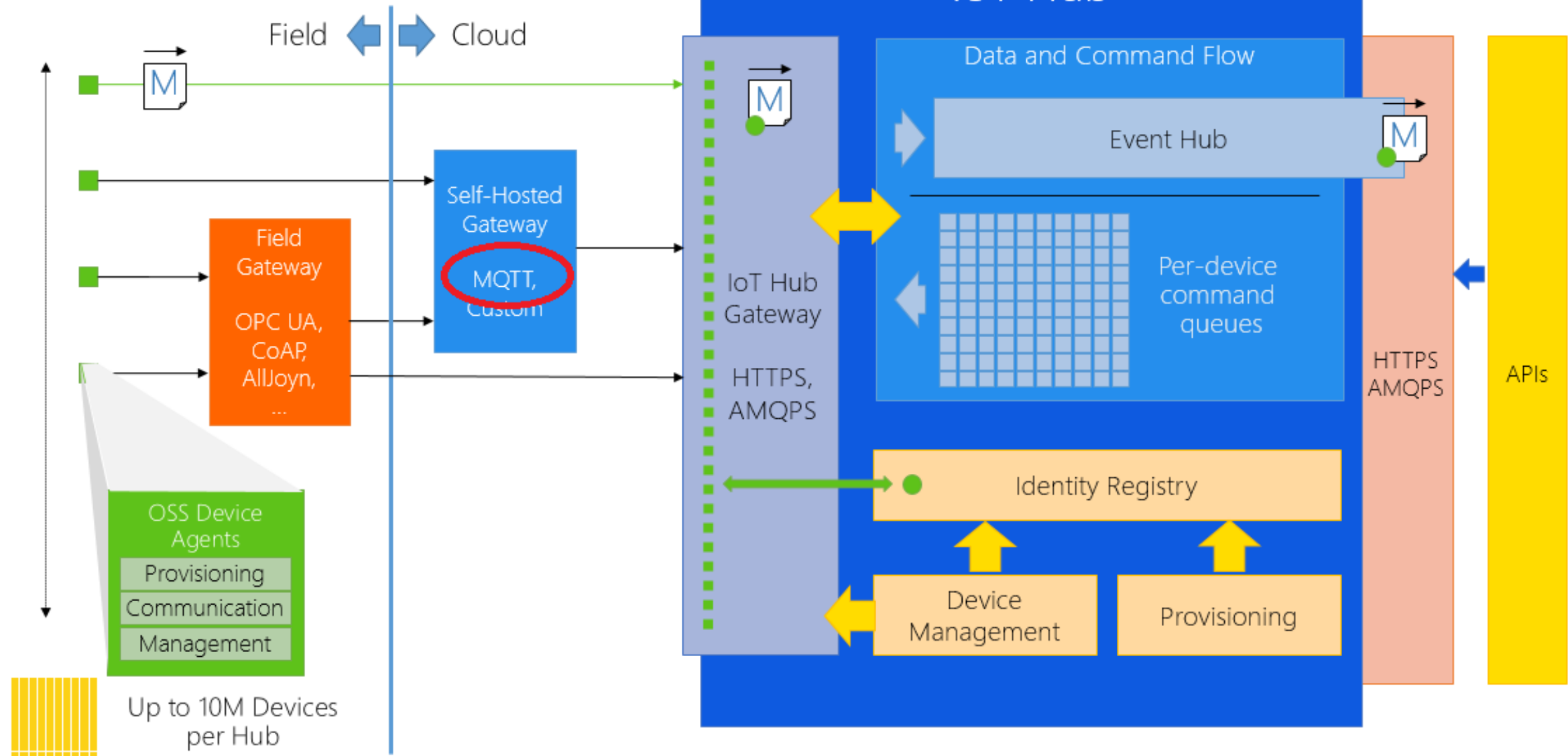Additional Resources

Azure IoT Edge

- Built on container technology as 'modules'

- Modules support Python, NodeJS, .Net Core, Java, & C

- Low-latency AMQP / MQTT data transport

- Operate in offline / intermittent network conditions

- Supports Linux X64 | ARM32/64 , Windows X64
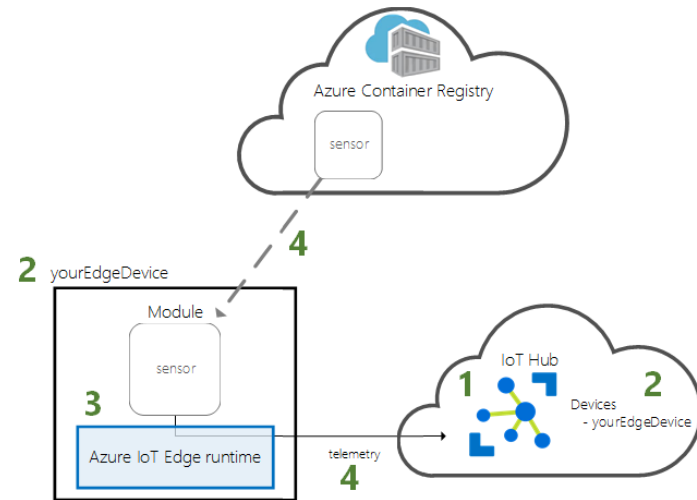
OSS and available @ https://github.com/Azure/iotedge

System Modules

edge-agent:

Deployment & Container orchestration
Ensures module uptime

edge-hub:

Communication to/from Azure IoT Hub
Inter-module communication
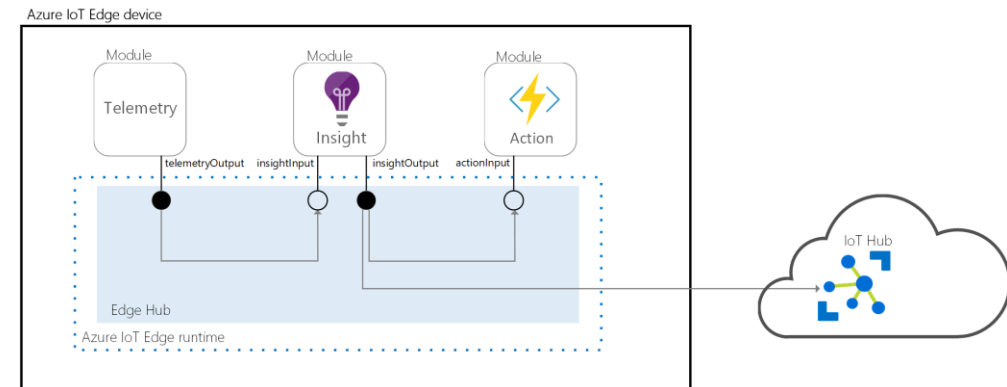
Azure Services as Edge Modules

Azure Stream Analytics
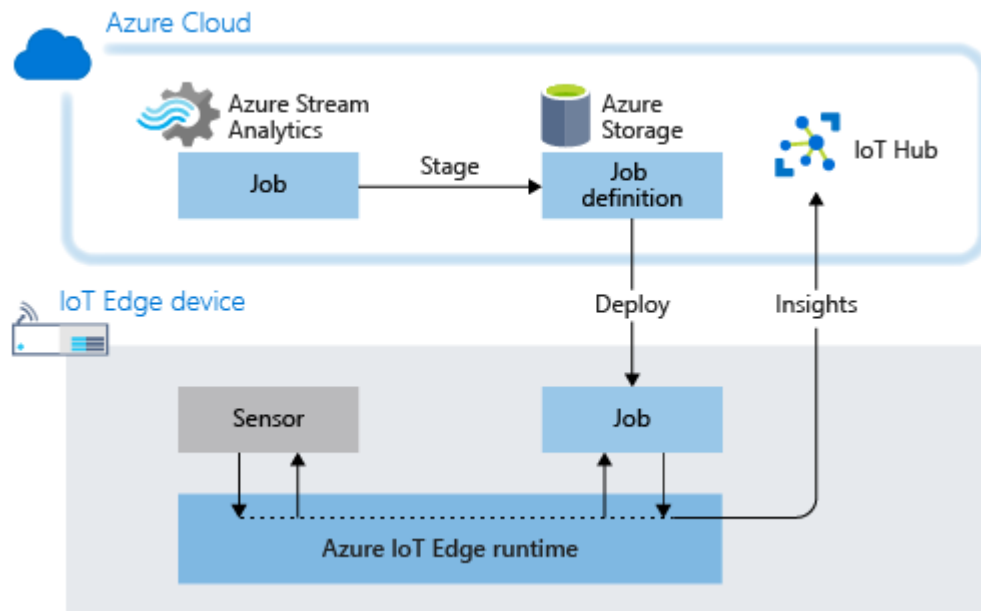
Azure Functions
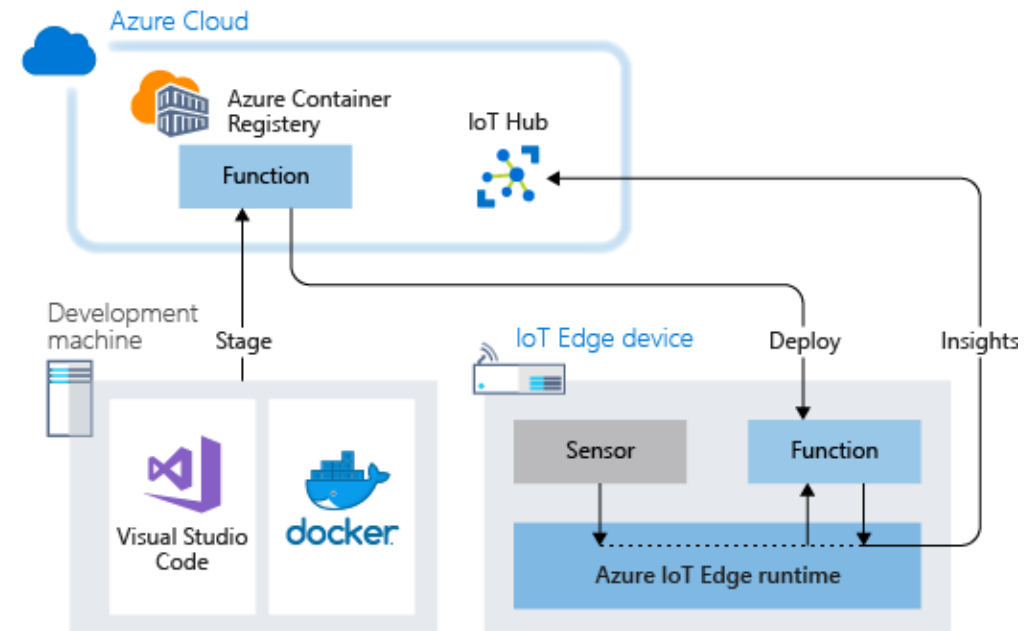
AI/ML Modules

- Cognitive Services including:
  - Text Analytics - Key Phrase Extraction, Language Detection, Sentiment Analysis
  - Face – Detection, Verification, and Emotion
  - Computer Vision – Text Recognition, custom vision module allows to finds whatever you train it to find

- Azure Machine Learning at the Edge
  - Export AML models to run a Edge modules

IoT Edge Deployment Process

Publish Deployment to IoT Hub for targeted devices

Iot Edge Deployment created
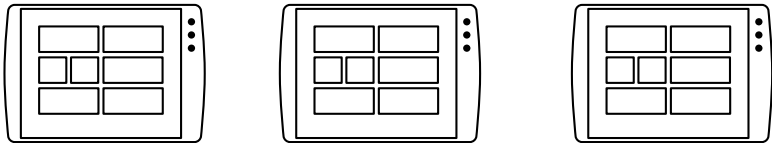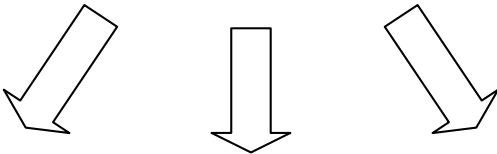
Iot Edge Devices matching target criteria in respective DeviceTwin receive updated deployment configuration

IoT Edge Devices   IoT Edge Deployments

IoT Edge Deployments

An IoT Edge deployment configures a target set of IoT Edge devices to run a set of IoT Edge modules. Each deployment continuous condition it meets. Each IoT Hub supports up to 10 deployments. Learn more about IoT Edge deployment.
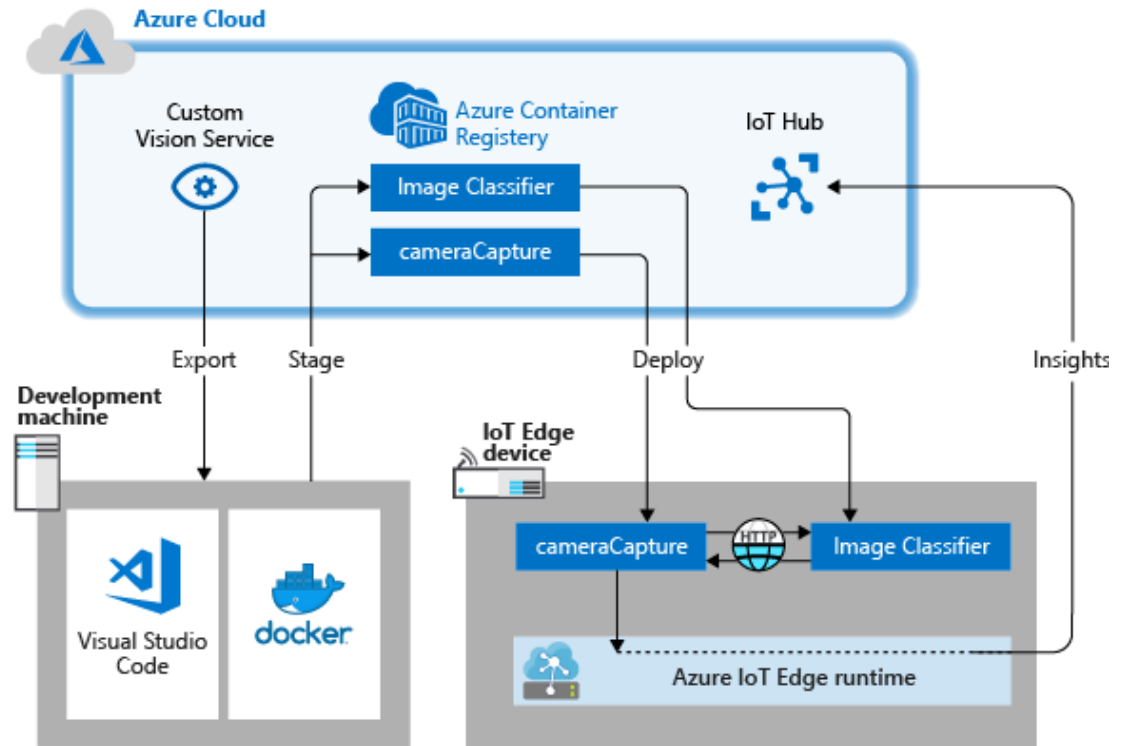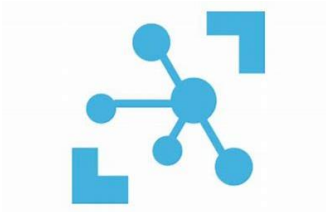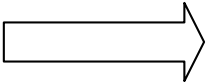
ID	TARGET CONDITION
vsts-deployment-cpu	tags.environment='dev' and tags.platform='cpu'

Target Condition   Deployment Labels

The target condition defines the devices ta

* Target Condition ⓘ

tags.environment='dev' and tags.platform='cpu'

Show devices   Reset target conditio

Targeted   Applied   Reporting Success

🔍 Filter devices

DEVICE ID

cse-paul-device

erisch

Total Devices: 2

DevOps Process with IoT Edge
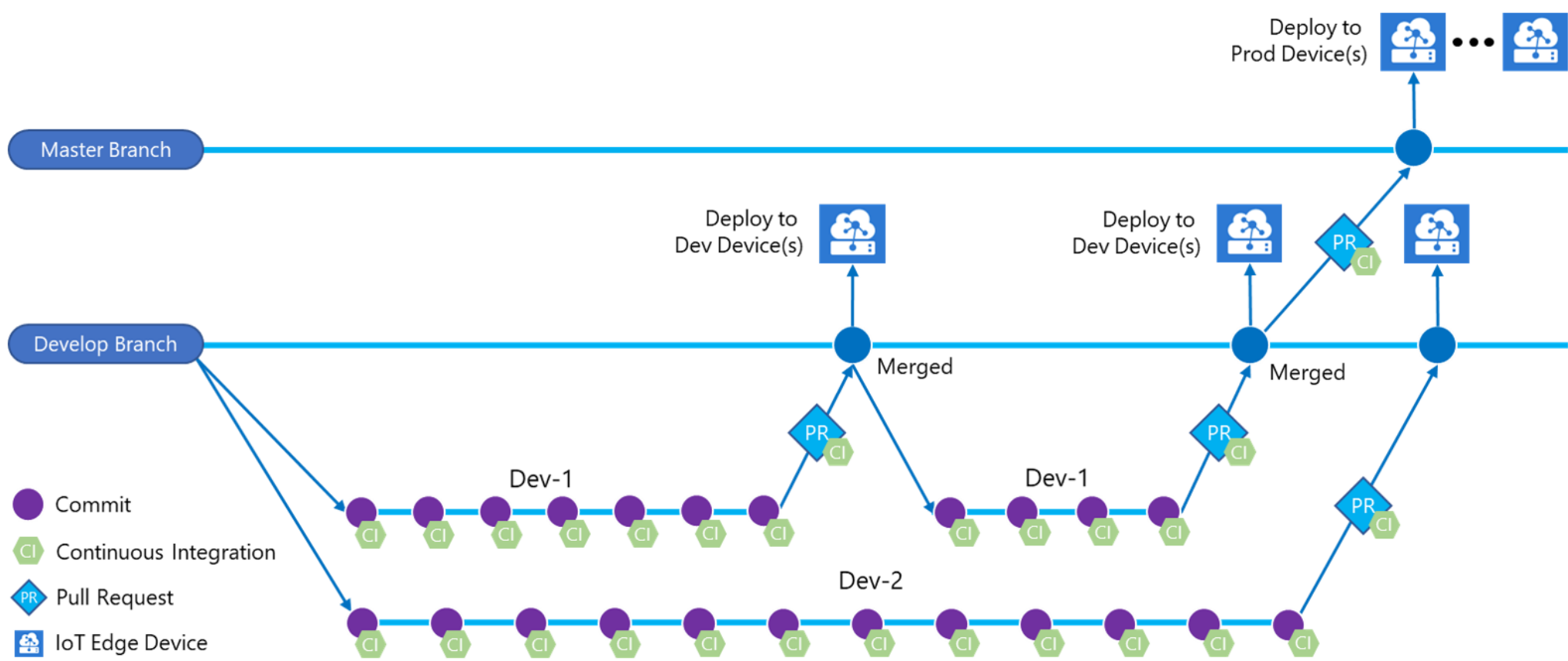
Edge intelligence enabled with Azure IoT Edge
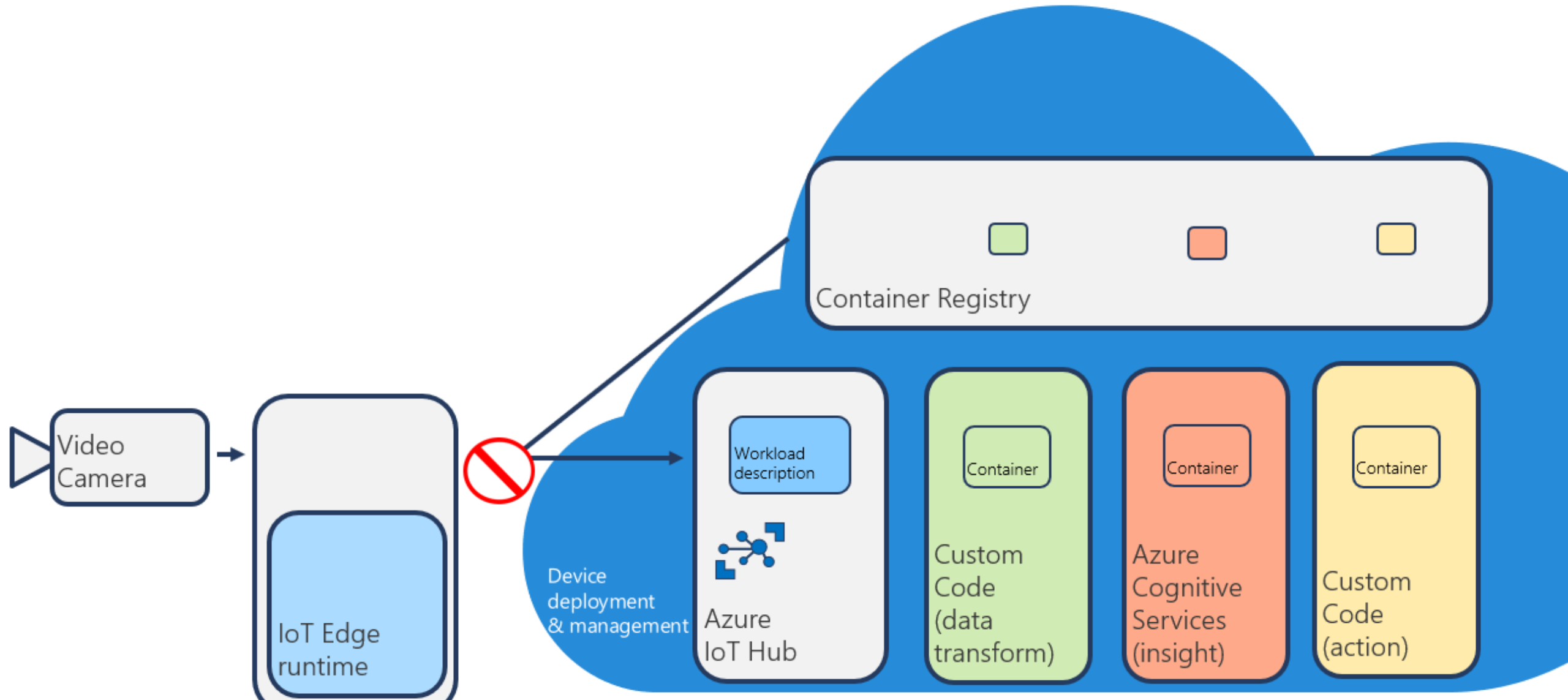
THE JETSON FAMILY
From AI at the Edge to Autonomous Machines

JETSON NANO

5–10W
0.5 TFLOPS (FP16)
45mm x 70mm

JETSON TX1   **JETSON TX2 4GB**

7–15W
1–1.3 TFLOPS (FP16)
50mm x 87mm

JETSON TX2 8GB | Industrial

7–15W
1.3 TFLOPS (FP16)
50mm x 87mm

JETSON AGX XAVIER

10–30W
11 TFLOPS (FP16) | 32 TOPS (INT8)
100mm x 87mm

AI at the Edge

Fully Autonomous Machines

Multiple Devices — Same Software

Compare NVIDIA Jetson Module Specifications

	Jetson Nano	Jetson TX2 Series			Jetson AGX Xavier Series	
		TX2 4GB	TX2	TX2i	Jetson AGX Xavier 8GB	Jetson AGX Xavier
GPU	NVIDIA Maxwell™ architecture with 128 NVIDIA CUDA® cores	NVIDIA Pascal™ architecture with 256 NVIDIA CUDA cores			NVIDIA Volta™ architecture with 384 NVIDIA CUDA cores and 48 Tensor cores	NVIDIA Volta™ architecture with 512 NVIDIA CUDA cores and 64 Tensor cores
	0.5 TFLOPs (FP16)	1.3 TFLOPS (FP16)			5.5 TFLOPS (FP16) 11.1 TOPS (INT8)	11 TFLOPS (FP16) 22 TOPS (INT8)
CPU	Quad-core ARM® Cortex®-A57 MPCore processor	Dual-core Denver 2 64-bit CPU and quad-core ARM A57 complex			6-core Carmel ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3	8-core Carmel ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3

JETSON SOFTWARE

DeepStream SDK			Isaac Robotics Engine			

Modules

Depth Estimation	Object Detection	Pose Estimation	Gesture Recognition	Path Planning	Autonomous Navigation	Ecosystem Modules

JetPack SDK

TensorRT cuDNN	VisionWorks OpenCV	cuBLAS cuFFT	Vulkan OpenGL	libargus Video API	Drivers Ecosystem
Deep Learning	Computer Vision	Accel. Computing	Graphics	Multimedia	Sensors

CUDA-X • Linux for Tegra • ROS

Nsight Developer Tools

Jetson Nano   Jetson TX1/TX2   Jetson AGX Xavier

developer.nvidia.com/jetpack

JETPACK 4.2

Available Now For Jetson
developer.nvidia.com/jetpack

Package Versions

L4T BSP	32.1
Linux Kernel	4.9.140
Vulkan	1.1.1
OpenGL	4.6
OpenGL-ES	3.2.5
EGL	1.5
GLX	1.4
X11 ABI	24
Wayland	1.14
L4T Multimedia API	32.1
Argus Camera API	0.97
GStreamer	1.14.1
Nsight Systems	2019.3
Nsight Graphics	2018.7
Nsight Compute	1.0
Jetson GPIO	1.0
Jetson OS	Ubuntu 18.04
Host OS	Ubuntu 16.04 / 18.04

CUDA	10.0.166
cuDNN	7.3.1.28
TensorRT	5.0.6.3
VisionWorks	1.6
OpenCV	3.3.1
NPP	10.0

Install TensorFlow, PyTorch, Caffe, Caffe2, MXNet, ROS, and other GPU-accelerated libraries

NVIDIA TensorRT
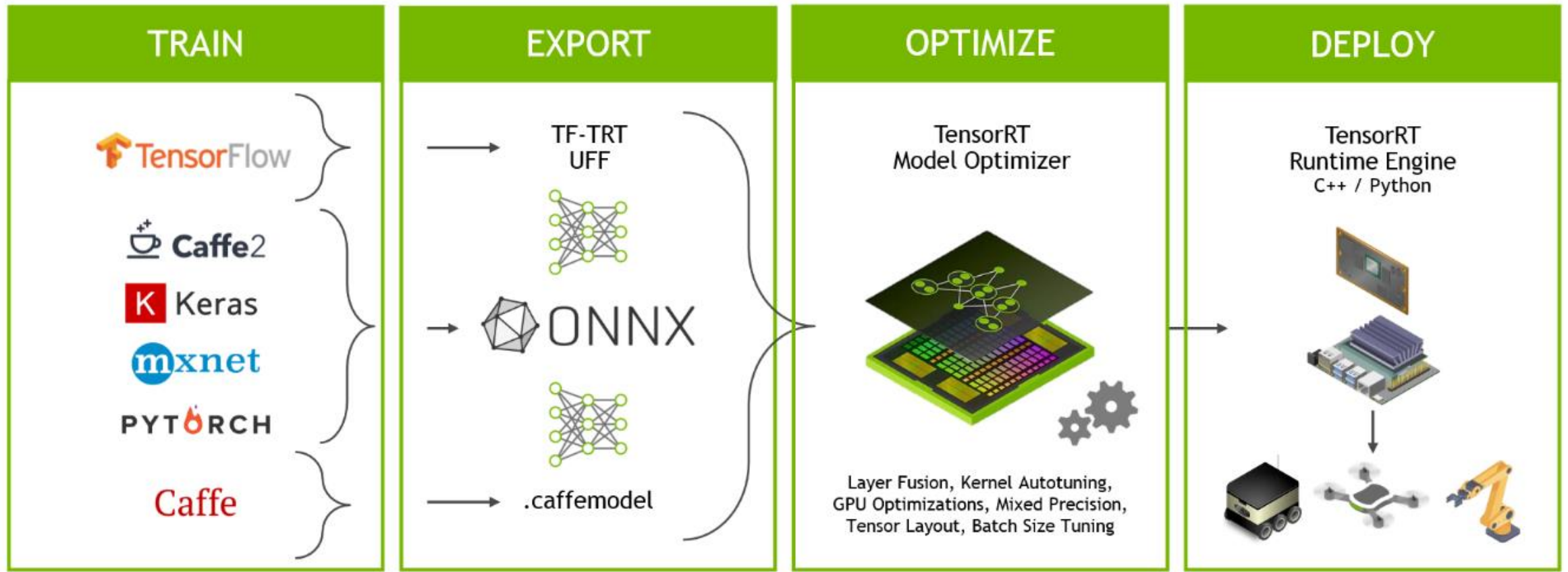
JETSON NANO RUNS MODERN AI

Inference

Img/sec

ResNet-50	36
Inception-v4	11
VGG-19	10
SSD Mobilenet-v2 (300x300)	39
SSD Mobilenet-v2 (480x272)	27
SSD Mobilenet-v2 (960x544)	8
Tiny YOLO	25
U-Net	18
Super Resolution	15
OpenPose	14

| TensorFlow | PyTorch | MxNet | TensorFlow | TensorFlow | TensorFlow | Darknet | Caffe | PyTorch | Caffe |
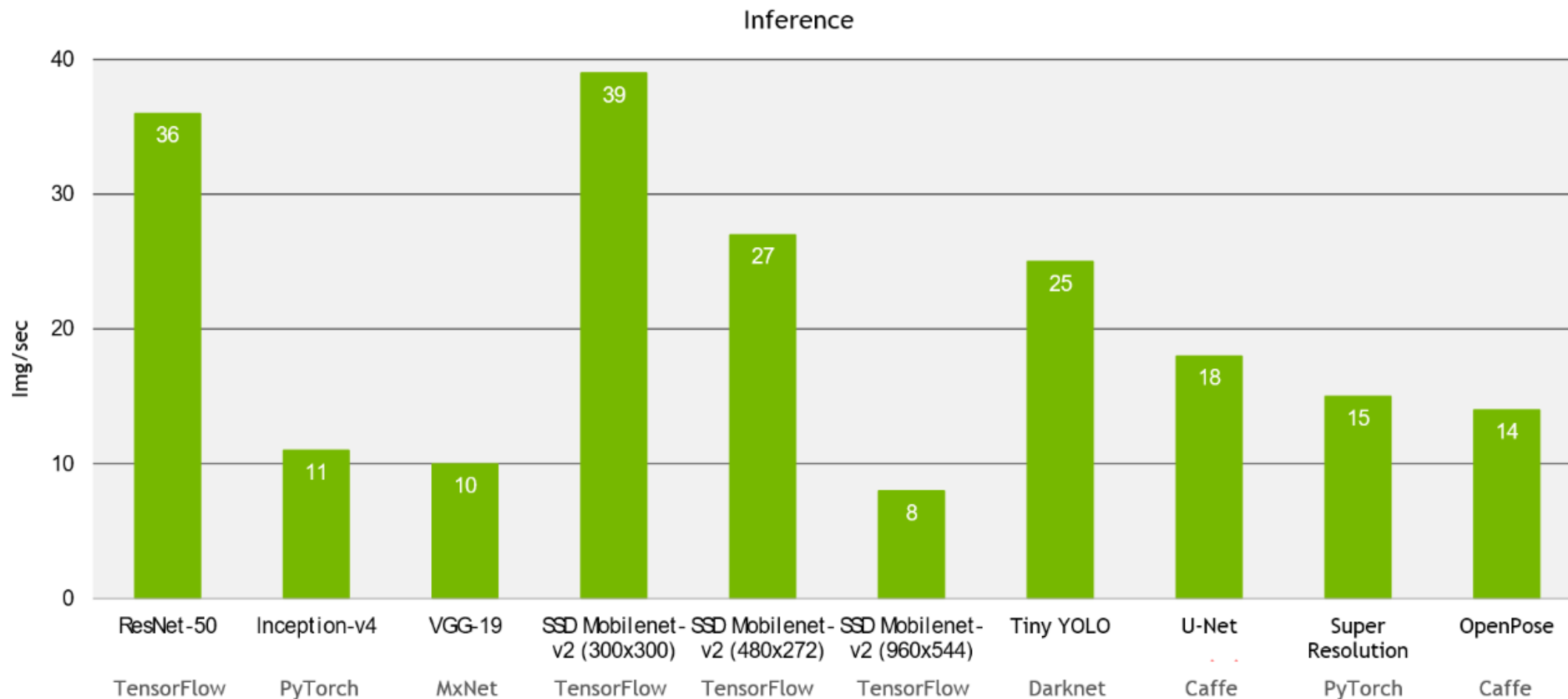
developer.nvidia.com/embedded/jetson-nano-dl-inference-benchmarks

Achieving Real-Time Performance

· Number of 1080p/30 FPS streams processed on various platforms.

NVIDIA Products	H.264	H.265
Jetson Nano	8	8
Jetson TX1	8	8
Jetson TX2	14	14
Jetson AGX Xavier	32	49
T4	35	68

Data measured using deepstream-app from DeepStream SDK 4.0

JETSON NANO COMPUTE MODULE

PROCESSOR	
CPU	64-bit Quad-core ARM A57 @ 1.43GHz
GPU	128-core NVIDIA Maxwell @ 921MHz
Memory	4GB 64-bit LPDDR4 @ 1600MHz | 25.6GB/s
Video Encoder	4Kp30 | (4x) 1080p30 | (2x) 1080p60
Video Decoder	4Kp60 | (2x) 4Kp30 | (8x) 1080p30 | (4x) 1080p60

INTERFACES	
USB	USB 3.0 | (3x) USB 2.0
Camera	12 lanes MIPI CSI-2 (up to 4 cameras)
Display	HDMI | DP | eDP | DSI
Networking	Gigabit Ethernet
PCIe	PCIe Gen2 x1/x2/x4
Storage	16GB eMMC 5.1
Other I/O	(4x) I2C | (2x) SPI | (3x) UART | (2x) I2S | GPIO
Power	5V DC, 5W | 10W
Size	45x70mm, 260-pin SODIMM connector
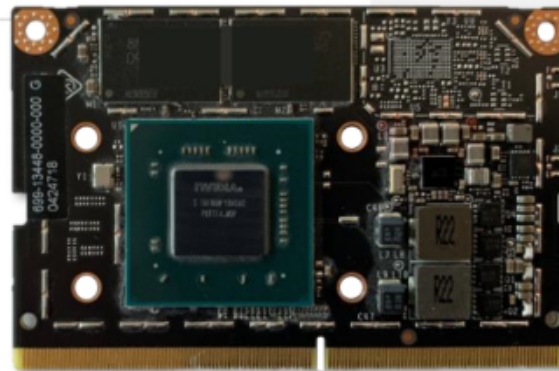
Production module
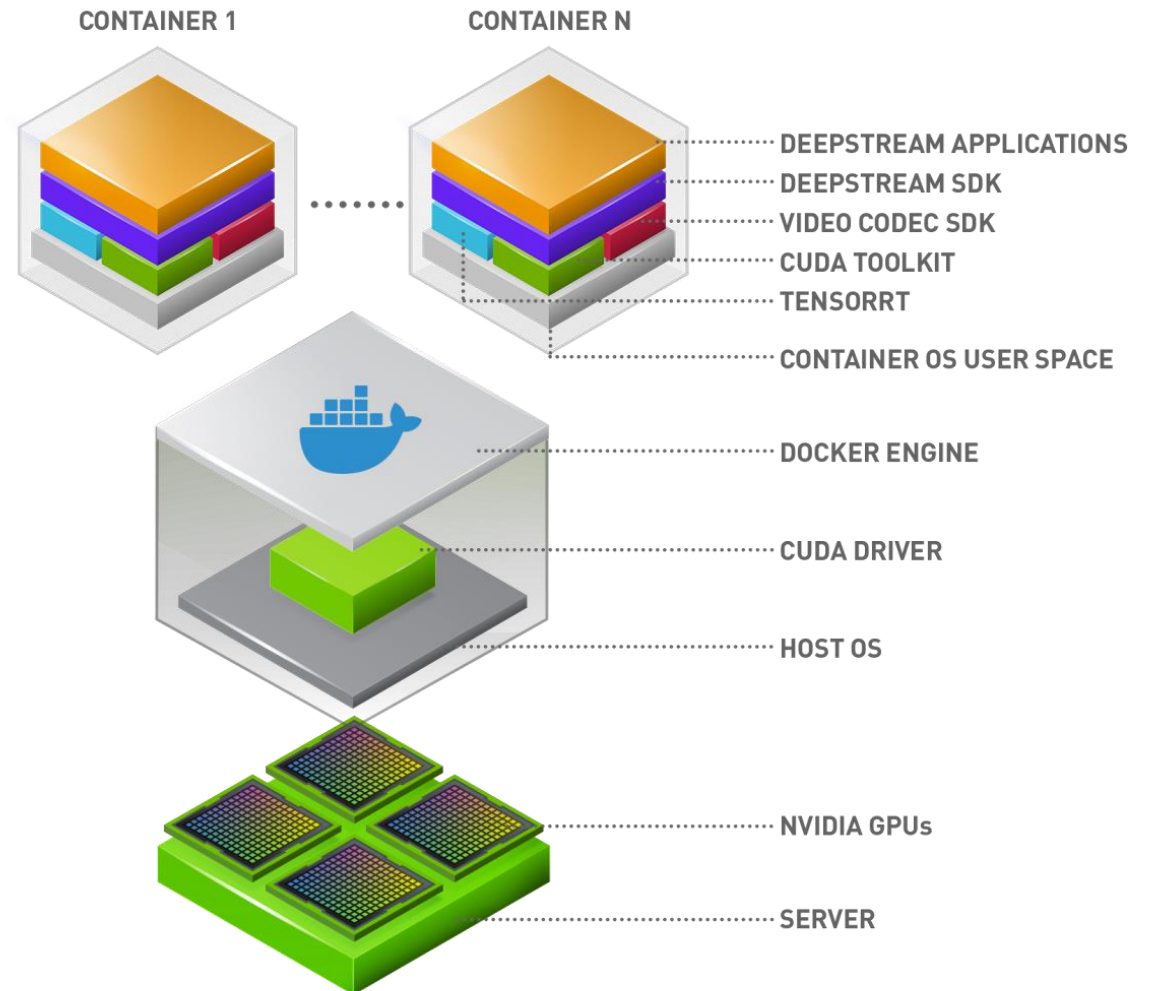available June 2019

Nvidia Docker

- CUDA Driver resides as part of the Host OS

- At runtime, nvidia-docker mounts CUDA Driver and adds devices to container

- Allows container to leverage GPU capabilities
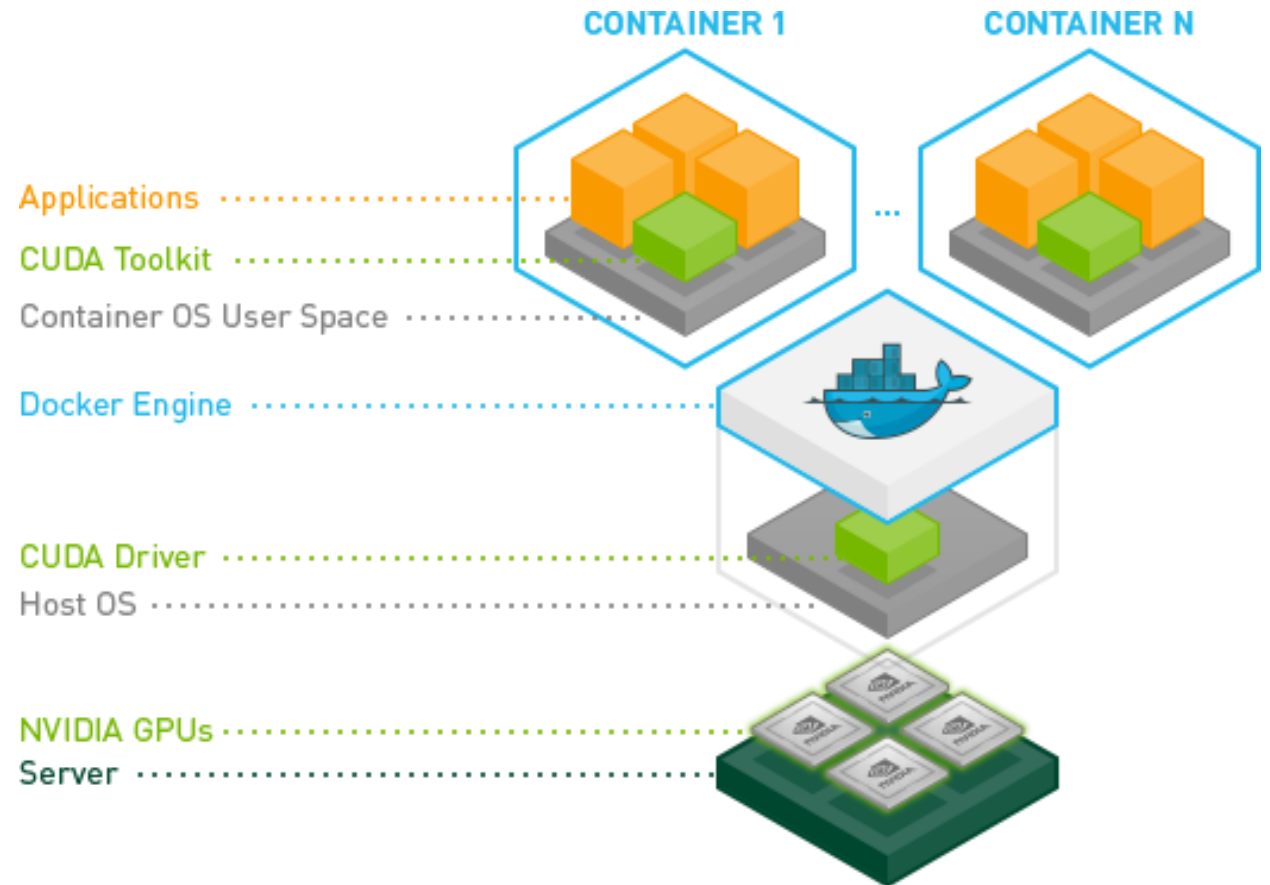
Deepstream Module

- Works on same principal as Nvidia Docker

- Contains Deepstream Application which is driven by a config file

- Supports messaging to Azure IoT Hub from Deepstream App

CONTAINER 1   CONTAINER N

DEEPSTREAM APPLICATIONS
DEEPSTREAM SDK
VIDEO CODEC SDK
CUDA TOOLKIT
TENSORRT
CONTAINER OS USER SPACE

DOCKER ENGINE

CUDA DRIVER

HOST OS

NVIDIA GPUs

SERVER

Demo

Nvidia Deepstream on Jetson Xavier

Video

Vision IQ Solution by the Marsden Group

Further Learning

Additional Resources

Onnx-runtime-iotedge:

https://github.com/Azure-Samples/onnxruntime-iot-edge

- Use onnx models with webcam input on Nano with IoT Edge

Ian Davis's blog:

https://codepyre.com/

- Contains info on building custom RootFS / Tensorflow

Paul DeCarlo's Dev.to articles:

https://dev.to/toolboc

- Articles related to IoT Edge Development on Nvidia devices

This powerpoint presentation:

http://aka.ms/IntelligentEdgeHOLdeck

Video recording of this presentation:

http://aka.ms/IntelligentEdgeHOLvideo

Hardware Checklist

Bill of Materials

Jetson Nano DevKit - https://amzn.to/2WFE5zF
Jetson Nano Case + Fan - https://amzn.to/2ZI2ki9
Jetson Nano 4V 2A barrel adapter - https://amzn.to/32DFsTq
Micro Jumper - https://amzn.to/2SoyPQa
128GB MicroSD card - https://amzn.to/32zj0ep
Edimax WiFi Adapter - https://amzn.to/2SuDXT8