Deploying AI and Machine Learning on Layerscape

Ravi Malhotra

Strategic Marketing Manager

May 2019 | AMF-SOL-T3526

SECURE CONNECTIONS
FOR A SMARTER WORLD

Agenda

- What is AI?
- Examples of AI usage in Industrial
- Breakdown of an Edge AI Application
- Layerscape support for AI
- Mapping AI use-cases to Layerscape
- Deploying AI with EdgeScale

Defining Common Terms

- Artificial intelligence (AI)
  - A computer performs tasks considered heretofore to require human intelligence
- Machine learning (ML)
  - Key term is learning: input data teaches the model how to function
  - Learning is typically supervised (the model is trained using input and the correct output)
    - Application of the trained model is called inferencing
  - But learning may be unsupervised (e.g., cluster analysis)
- Neural network (NN)
  - A class of ML algorithms
- Deep learning
  - ML using a big neural net

Similar AI Tasks Have Important Differences

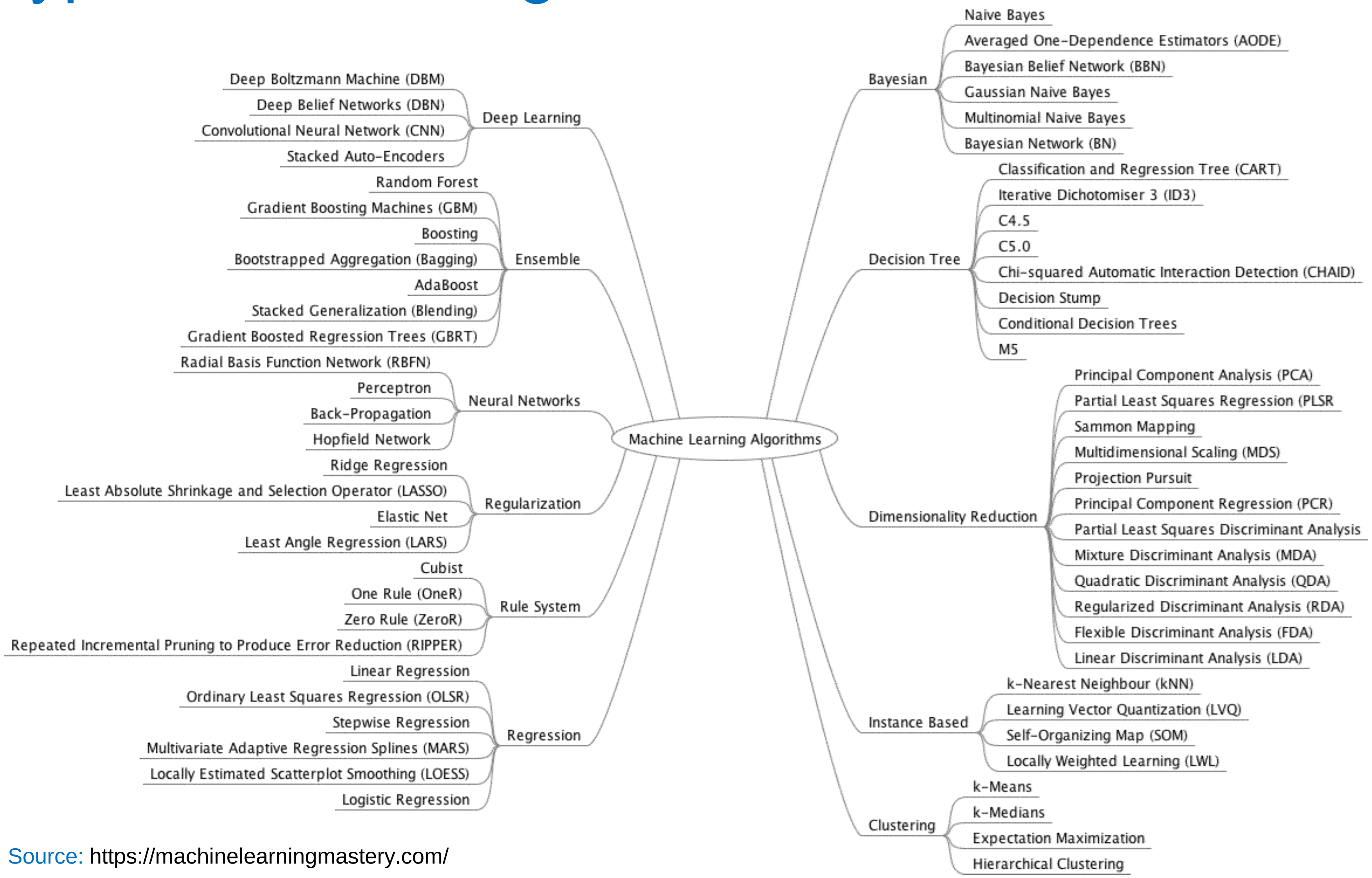
- ADAS
  - Identifies pedestrians, cars, signs, lane markings, obstacles, etc.
  - Regardless of who a pedestrian is, it won’t run him over

- Face recognition
  - Only identifies faces
  - Differentiates many people

- Machine inspection
  - Only knows widgets
  - Only classifies as good or bad

		Types of Items	
		Few	Many
Items Within A Type	Few	Machine Inspection	ADAS
	Many	Face Recognition	General Object Recognition

Many Types of AI/ML Algorithms Out There…



Machine Learning Algorithms

- Deep Learning
  - Deep Boltzmann Machine (DBM)
  - Deep Belief Networks (DBN)
  - Convolutional Neural Network (CNN)
  - Stacked Auto-Encoders
- Ensemble
  - Random Forest
  - Gradient Boosting Machines (GBM)
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization (Blending)
  - Gradient Boosted Regression Trees (GBRT)
- Neural Networks
  - Radial Basis Function Network (RBFN)
  - Perceptron
  - Back-Propagation
  - Hopfield Network
- Regularization
  - Ridge Regression
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Elastic Net
  - Least Angle Regression (LARS)
- Rule System
  - Cubist
  - One Rule (OneR)
  - Zero Rule (ZeroR)
  - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
- Regression
  - Linear Regression
  - Ordinary Least Squares Regression (OLSR)
  - Stepwise Regression
  - Multivariate Adaptive Regression Splines (MARS)
  - Locally Estimated Scatterplot Smoothing (LOESS)
  - Logistic Regression
- Bayesian
  - Naive Bayes
  - Averaged One-Dependence Estimators (AODE)
  - Bayesian Belief Network (BBN)
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Bayesian Network (BN)
- Decision Tree
  - Classification and Regression Tree (CART)
  - Iterative Dichotomiser 3 (ID3)
  - C4.5
  - C5.0
  - Chi-squared Automatic Interaction Detection (CHAID)
  - Decision Stump
  - Conditional Decision Trees
  - M5
- Dimensionality Reduction
  - Principal Component Analysis (PCA)
  - Partial Least Squares Regression (PLSR
  - Sammon Mapping
  - Multidimensional Scaling (MDS)
  - Projection Pursuit
  - Principal Component Regression (PCR)
  - Partial Least Squares Discriminant Analysis
  - Mixture Discriminant Analysis (MDA)
  - Quadratic Discriminant Analysis (QDA)
  - Regularized Discriminant Analysis (RDA)
  - Flexible Discriminant Analysis (FDA)
  - Linear Discriminant Analysis (LDA)
- Instance Based
  - k-Nearest Neighbour (kNN)
  - Learning Vector Quantization (LVQ)
  - Self-Organizing Map (SOM)
  - Locally Weighted Learning (LWL)
- Clustering
  - k-Means
  - k-Medians
  - Expectation Maximization
  - Hierarchical Clustering

Source: https://machinelearningmastery.com/

Why AI?

Faster than human analysis

Cooler under pressure

Analyzes more data than humanly possible

Better insights than man-made models

Reduces **cost**, increases **revenue**

Increases **safety**

AI Improves Quality

- Quality management reduces manufacturing cost
- High-quality products improve customer satisfaction
- Object-detection techniques can be adapted to visual quality inspection
- Other sensors (e.g., acoustic) can inspect in ways people cannot
- Technology for smart maintenance can be adapted to process monitoring (Quality 4.0)

Security and Surveillance

- Fire, theft, trespassing cost businesses
- AI is more attentive than human agents
- AI frees people to focus on addressing issues
- AI-based security can be lower cost and less discriminatory
- Typical approach is to identify and track people
- AI systems can learn on their own to identify anomalous behavior

Industrial Safety Examples

- Virtual-fencing of safety zones
- Recognize faces to enforce authorization policies
- Detect objects to enforce PPE policies
- Monitor operator attention with gaze detection
- Track and monitor equipment and vehicles

AI in Warehousing

- Physical inventory using object detection
- Pick & place robots (see recent Boston Dynamics robot)
- Received-goods inspection (crate damage)
- Security and surveillance
- HVAC control (e.g., DeepMind and data centers)

Robotics

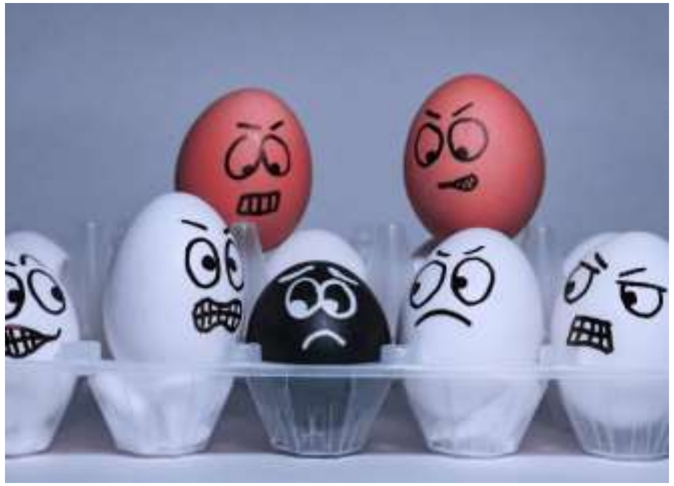
- Example uses: pick and place, assembly, packaging, AGV
- AI learns optimal paths vs following a set route
- AI/CV can identify objects for robot to interact with
- AI coordinates robot interaction with people (collaborative robots)

Issues with AI

Not provably correct

Sometimes fatally wrong

Biases possibly trained in



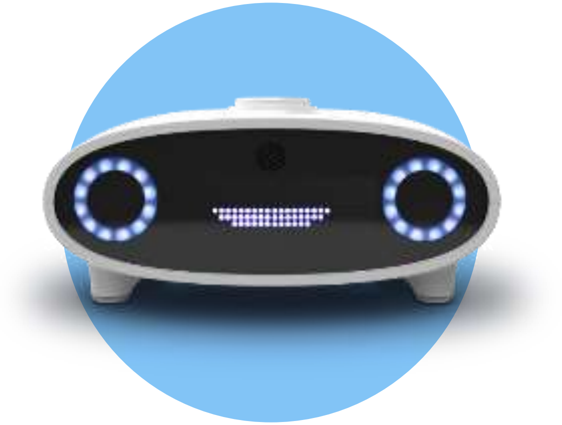
5 steps to AI/ML on Layerscape

- Application
  - Learning vs. inferencing, model creation.
- Breakdown
  - Mapping I/O processing and CNN
- Optimization
  - Accelerators or cores ?
- System
  - Peripherals, communication, security
- Deployment
  - Deploying AI applications and models to Edge nodes

AI @ the Edge vs. Cloud

Smart Endpoints

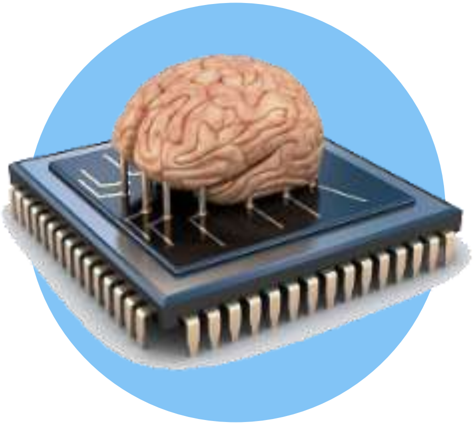
Integrated ML Optimization



- Optimizing ML operations running locally
- Dedicated AI/ML accelerator is optional
- Inferencing only

Edge Gateway

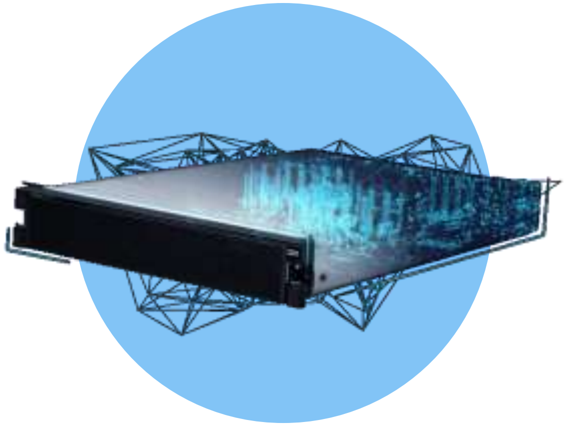
Integrated ML Optimization & Acceleration



- Enabling ML operations for connected devices
- Dedicated AI/ML accelerator is required
- Training may be turned off

Cloud & Data Center

ML App Acceleration & Offloading



- Leveraging GPU and TPU
- iNIC or smart offloading line cards
- Support both training and inferencing

ARM VSPA TPU GPU

AI @ the Edge vs. Cloud – Performance vs. Practicality

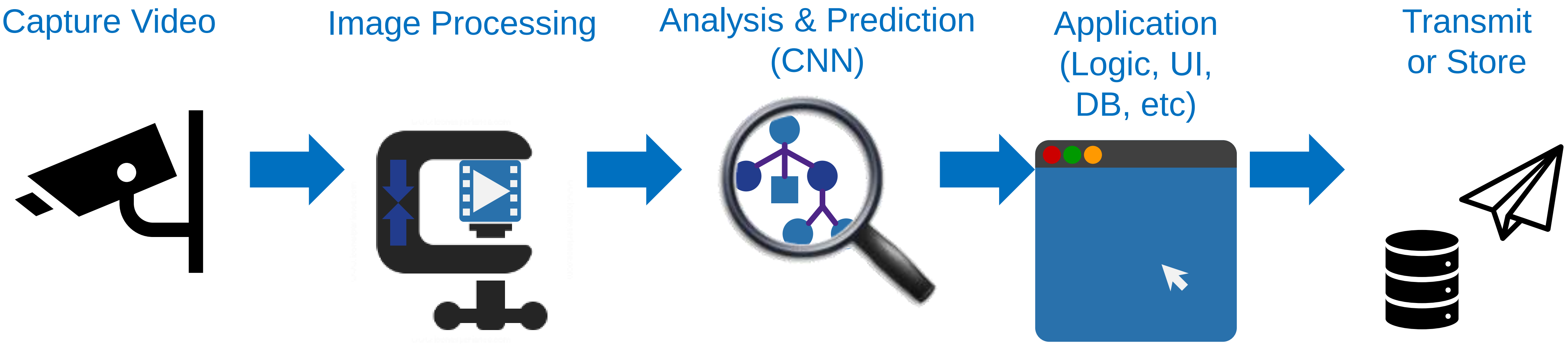
	Cloud server + GPU	Edge appliance
Compute	AMD Ryzen 2600 + nVidia GTX 1080Ti (12 TOps)	Layerscape LS1046
Power	~250W	~10W
Input video	MI 6 trailer – 1080p	MI 6 trailer – 720p
Algorithm	YOLOv3	YOLOv3
Object Detection – fps	25 fps	3 fps
CPU Utilization	100% 2 cores @ 3.4 Ghz + 85% GPU	100% 4 cores @ 1.8 Ghz

Great for Formula 1 close finishes.

Efficient at counting cars and people in a parking lot.

Watch LS1046 object detection sample @ https://youtu.be/EEc5-oiccuM

Breakdown of an Edge Application Using AI



Capture Video

Image Processing

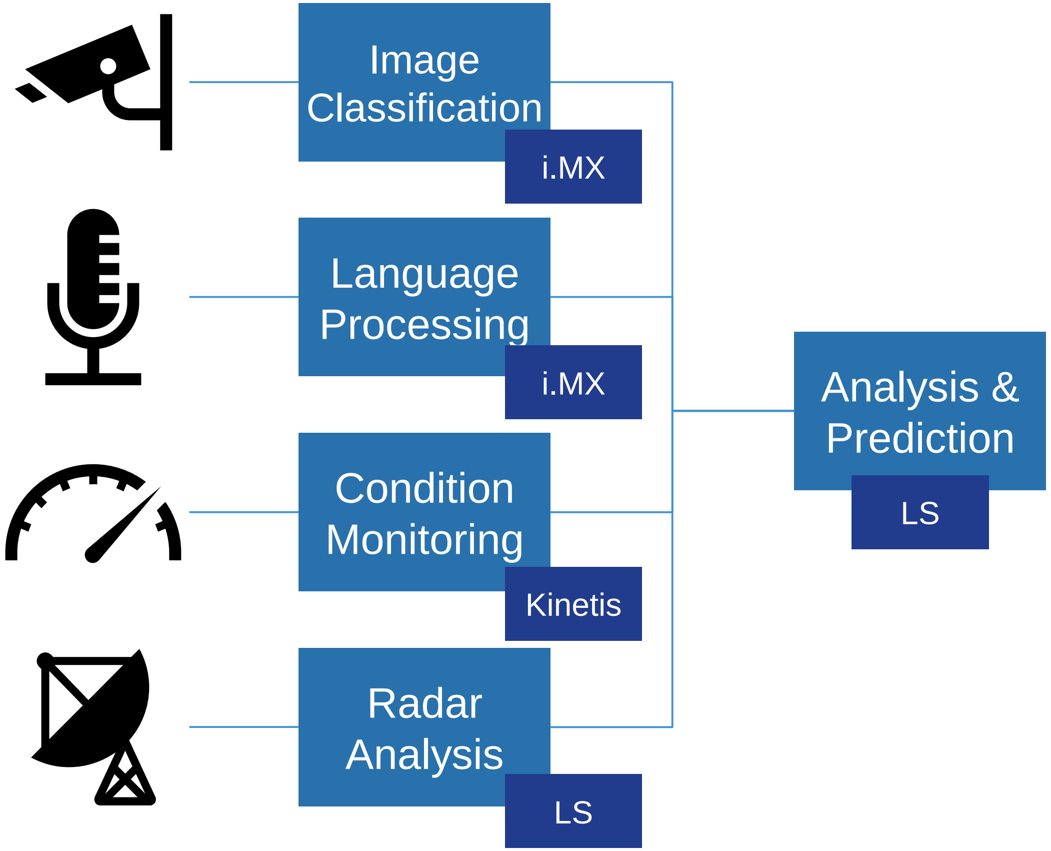
Analysis & Prediction (CNN)

Application (Logic, UI, DB, etc)

Transmit or Store

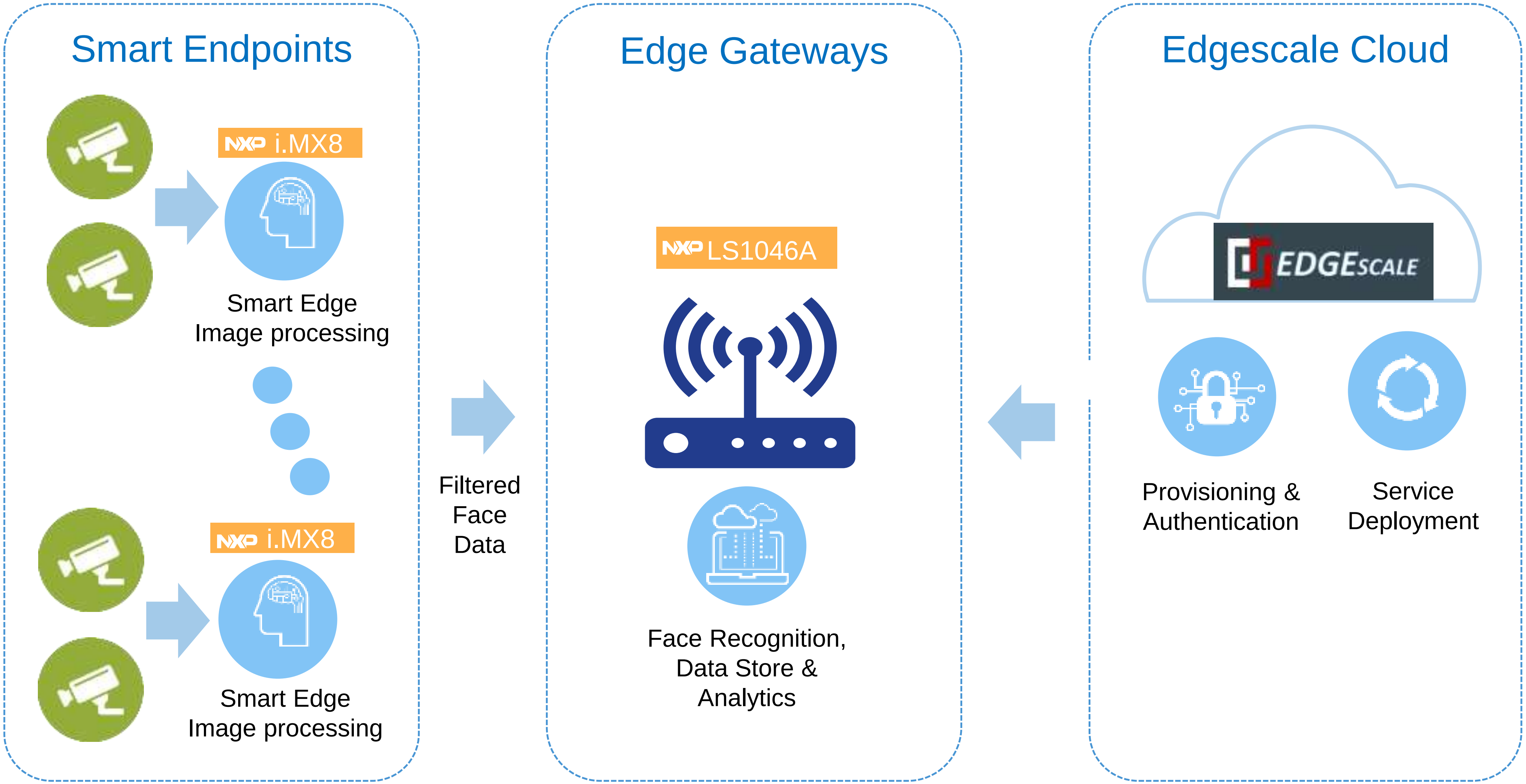
Hardware	USB Ethernet/IP MIPI to ISP (pref)	CPU, GPU (preferred)	CPU, NN Accel (GPU, TPU, VSPA)	CPU GPIO, USB, PCIE	USB Ethernet/IP SATA NVMe
Software	Drivers, V4L	V4L, G-streamer, video codecs	Frameworks (e.g., TensorFlow) Turnkey Models Training	Custom	Linux Network stack, File-system

Cascade Layerscape and i.MX Processors for Complex Designs



- First-level functions (i.MX, LS, MCU)
  - Classify/perceive
  - Recognize/model
  - Preprocess
- Second-level functions (Layerscape)
  - Fuse first-level inputs
  - Interpret data and model behavior
  - Predict and plan responses
  - Log data
  - Communicate

Scalable Video Analytics Solution



Smart Endpoints

NXP i.MX8

Smart Edge Image processing

NXP i.MX8

Smart Edge Image processing

Filtered Face Data

Edge Gateways

NXP LS1046A

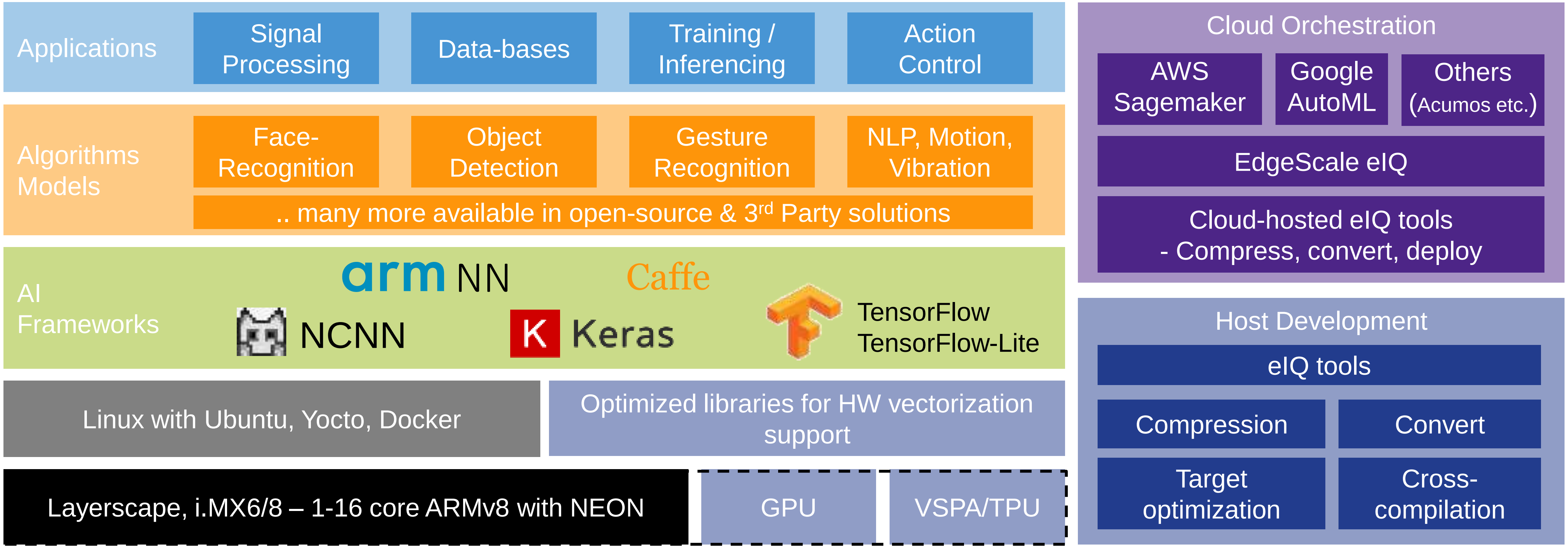
Face Recognition, Data Store & Analytics

Edgescale Cloud

Provisioning & Authentication

Service Deployment

Edgescale and eIQ for AI on Layerscape & i.MX



Applications

Signal Processing

Data-bases

Training / Inferencing

Action Control

Algorithms Models

Face-Recognition

Object Detection

Gesture Recognition

NLP, Motion, Vibration

.. many more available in open-source & 3rd Party solutions

AI Frameworks

arm NN

Caffe

NCNN

Keras

TensorFlow
TensorFlow-Lite

Linux with Ubuntu, Yocto, Docker

Optimized libraries for HW vectorization support

Layerscape, i.MX6/8 – 1-16 core ARMv8 with NEON

GPU

VSPA/TPU

Cloud Orchestration

AWS Sagemaker

Google AutoML

Others (Acumos etc.)

EdgeScale eIQ

Cloud-hosted eIQ tools
- Compress, convert, deploy

Host Development

eIQ tools

Compression

Convert

Target optimization

Cross-compilation

- NXP provides the right enablement for cloud-connected AI/ML applications @ Edge.
- Host-based eIQ tools for model conversion, optimization and target optimization.
- Edgescale leverages eIQ tools for cloud-based orchestration and integration with Sagemaker, AutoML etc.
- Helps customer leverage open-source frameworks, models and communities.

AI Frameworks Running on Layerscape

- Layerscape SDK supports popular AI/ML frameworks
  - Documentation available
  - Customer support available
- Other supported software
  - Video codecs
  - Camera drivers

NCNN

armNN

Caffe

Keras

TensorFlow
TensorFlow-Lite

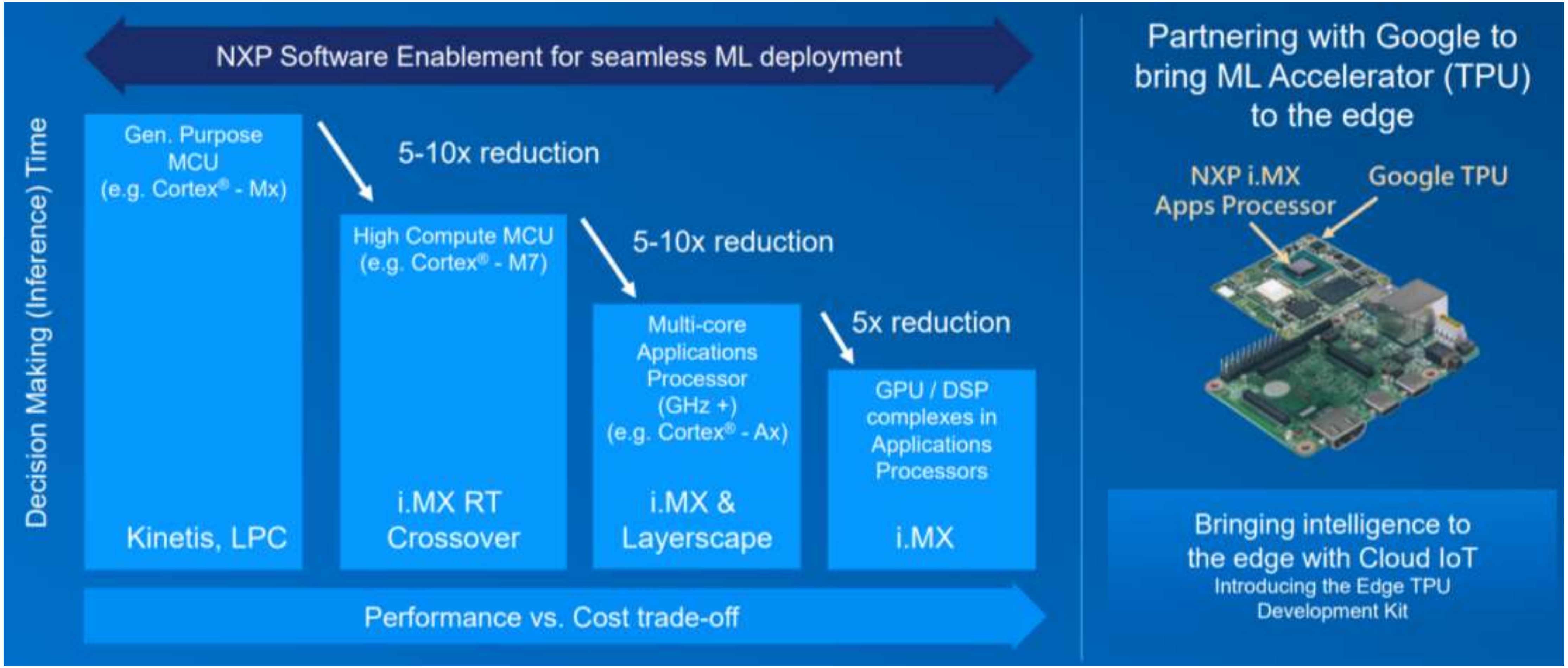
Choosing the Right Algorithm Matters

	Option 1	Option 2
Algorithm	FaceNet	MobileFaceNet
Inference Framework	Tensorflow	NCNN
Implementation	Tensorflow (Python)	C++ (no lib dependency)
Performance (LS1046 – 4x A72@1.8GHz)	**4 core: ~200 msec**	**4 core: ~10ms** **1 core: ~50ms**
Accuracy (improvable with training)	**99.6%**	**99.5%**
Model Complexity (#weights)	**19.5M**	**1M**
Model File Size (MB @Float32)	**93**	**4**
OS	Linux	Linux, Android, Portable to RTOS

- AI Algorithms and Frameworks are rapidly evolving.
- What works well on servers may not be optimized for the embedded Edge.
- General purpose cores may perform as well as accelerators for certain workloads.

Edge Compute Enabler – Scalable Inference

Balancing Cost vs. End-user Experience



NXP Software Enablement for seamless ML deployment

Decision Making (Inference) Time

Gen. Purpose MCU (e.g. Cortex® - Mx)

Kinetis, LPC

5-10x reduction

High Compute MCU (e.g. Cortex® - M7)

i.MX RT Crossover

5-10x reduction

Multi-core Applications Processor (GHz +) (e.g. Cortex® - Ax)

i.MX & Layerscape

5x reduction

GPU / DSP complexes in Applications Processors

i.MX

Performance vs. Cost trade-off

Partnering with Google to bring ML Accelerator (TPU) to the edge

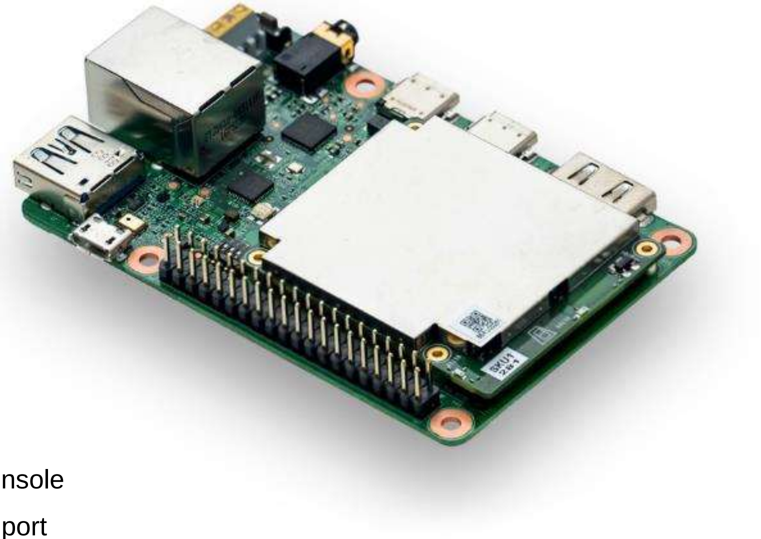
Bringing intelligence to the edge with Cloud IoT

Introducing the Edge TPU Development Kit

Google Edge TPU SOM w/ NXP SoC

Edge TPU Module (SOM) Specifications

CPU	NXP i.MX 8M SOC (quad Cortex-A53, Cortex-M4F)
GPU	Integrated GC7000 Lite Graphics
ML accelerator	Google Edge TPU coprocessor
RAM	1 GB LPDDR4
Flash memory	8 GB eMMC
Wireless	Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5GHz)
	Bluetooth 4.1
Dimensions	40 mm x 48 mm

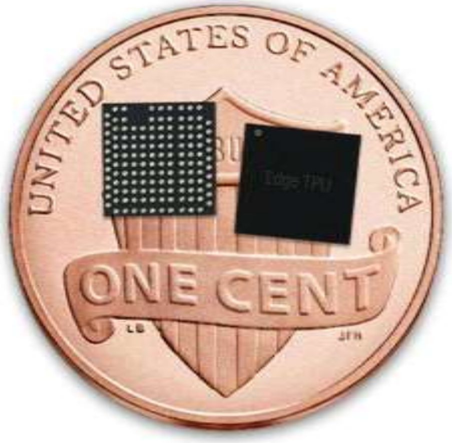


Board Features

Flash memory	MicroSD slot
USB	Type-C OTG
	Type-C power
	Type-A 3.0 host
	Micro-B serial console
LAN	Gigabit Ethernet port
Audio	3.5mm audio jack (CTIA compliant)
	Digital PDM microphone (x2)
	2.54mm 4-pin terminal for stereo speakers
Video	HDMI 2.0a (full size)
	39-pin FFC connector for MIPI-DSI display (4-lane)
	24-pin FFC connector for MIPI-CSI2 camera (4-lane)
GPIO	40-pin expansion header
Power	5V DC (USB Type-C)
Dimensions	85 mm x 56 mm

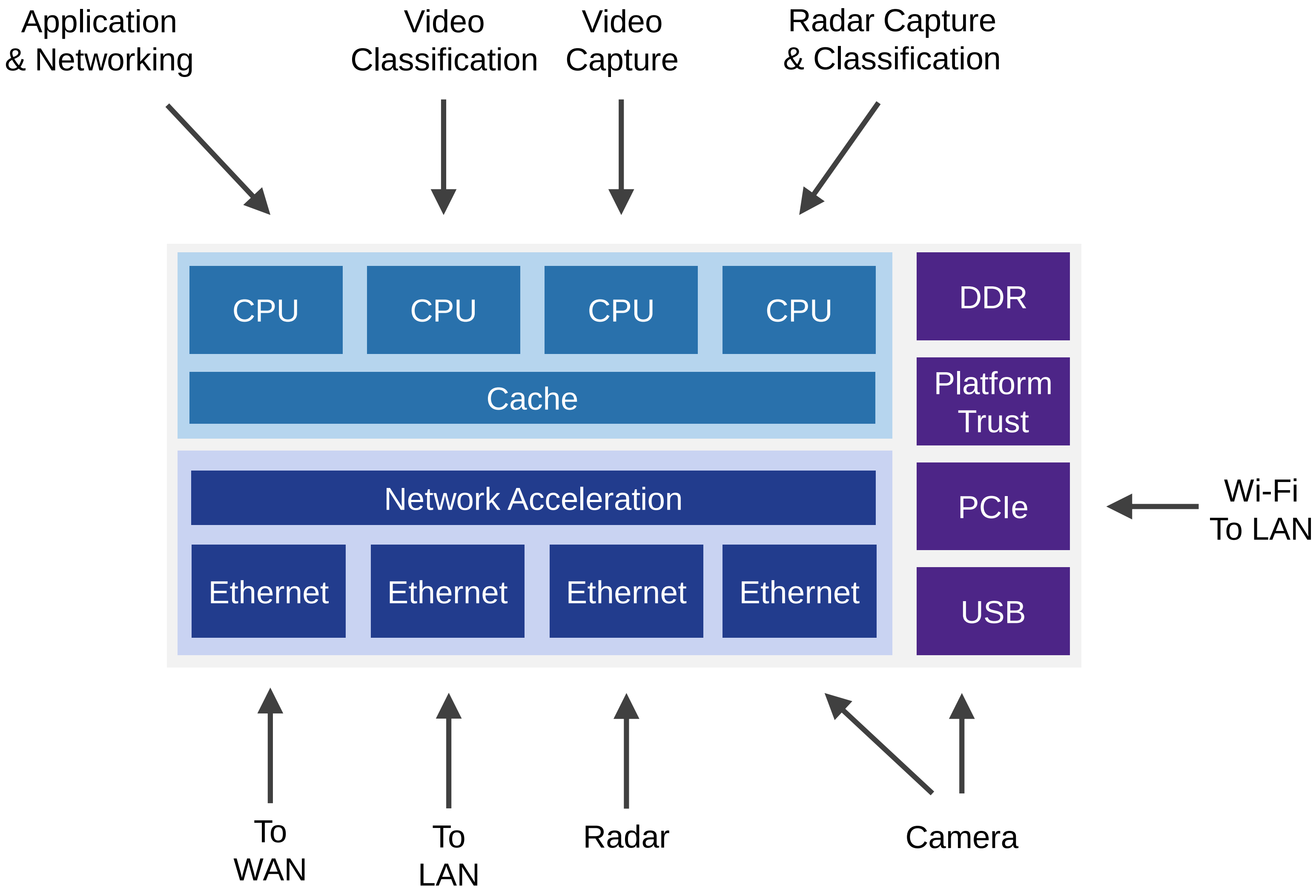
Google Edge TPU Performance on Common Vision Models

Model	Performance (connected through USB or PCIe)
GoogleNet:	600 fps
Inception v2:	400 fps
MobileNet:	700 fps



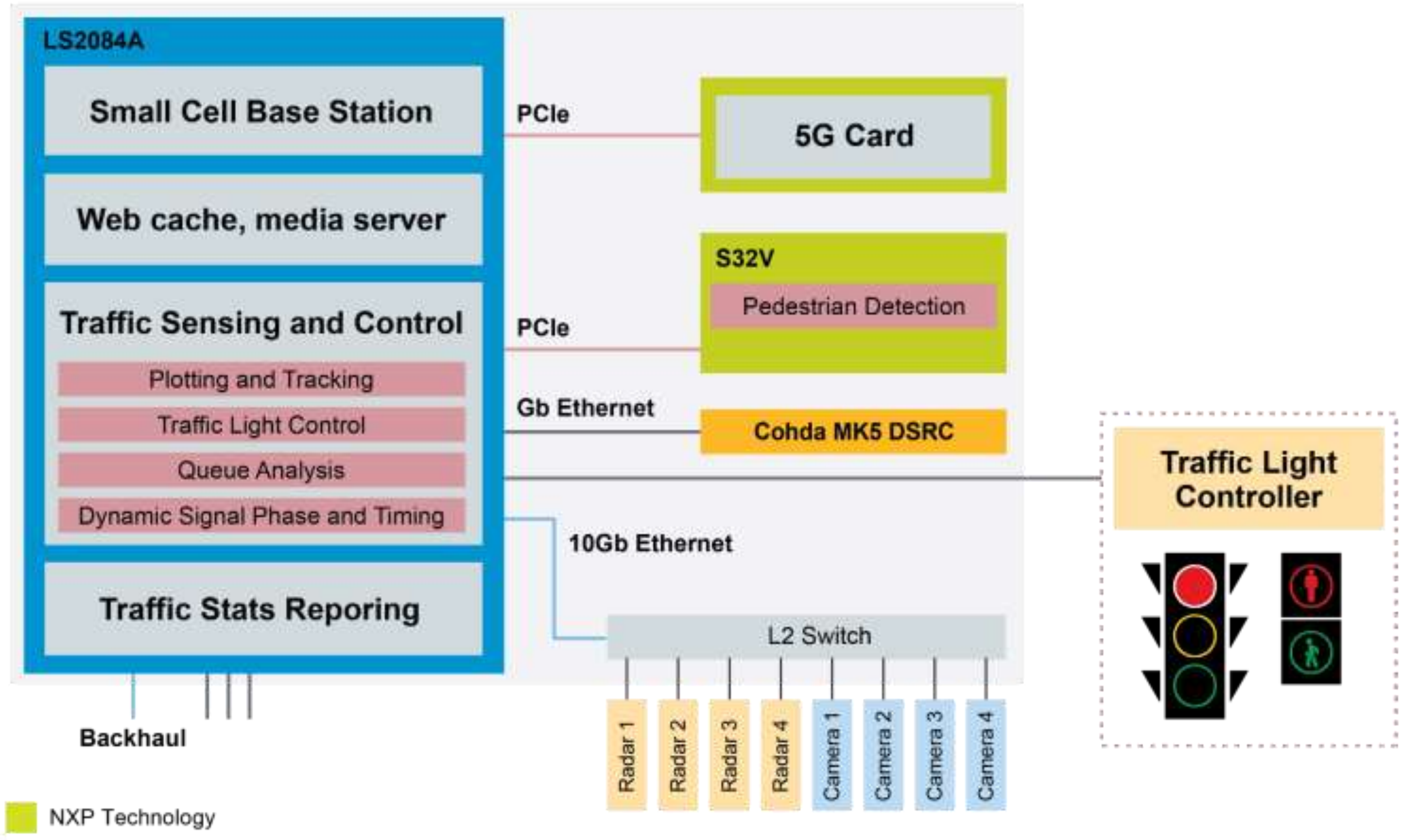
NXP is working with Google to explore Edge TPU usage in professional/industrial markets.

Mapping Home Automation & Safety to Layerscape LS1046



Mapping Layerscape LS2084 to Roadside Unit

INTELLIGENT TRAFFIC CONTROLLER BLOCK DIAGRAM



LS2084A

Small Cell Base Station

Web cache, media server

Traffic Sensing and Control

Plotting and Tracking

Traffic Light Control

Queue Analysis

Dynamic Signal Phase and Timing

Traffic Stats Reporing

PCIe

5G Card

S32V

Pedestrian Detection

PCIe

Gb Ethernet

Cohda MK5 DSRC

Traffic Light Controller

10Gb Ethernet

L2 Switch

Radar 1

Radar 2

Radar 3

Radar 4

Camera 1

Camera 2

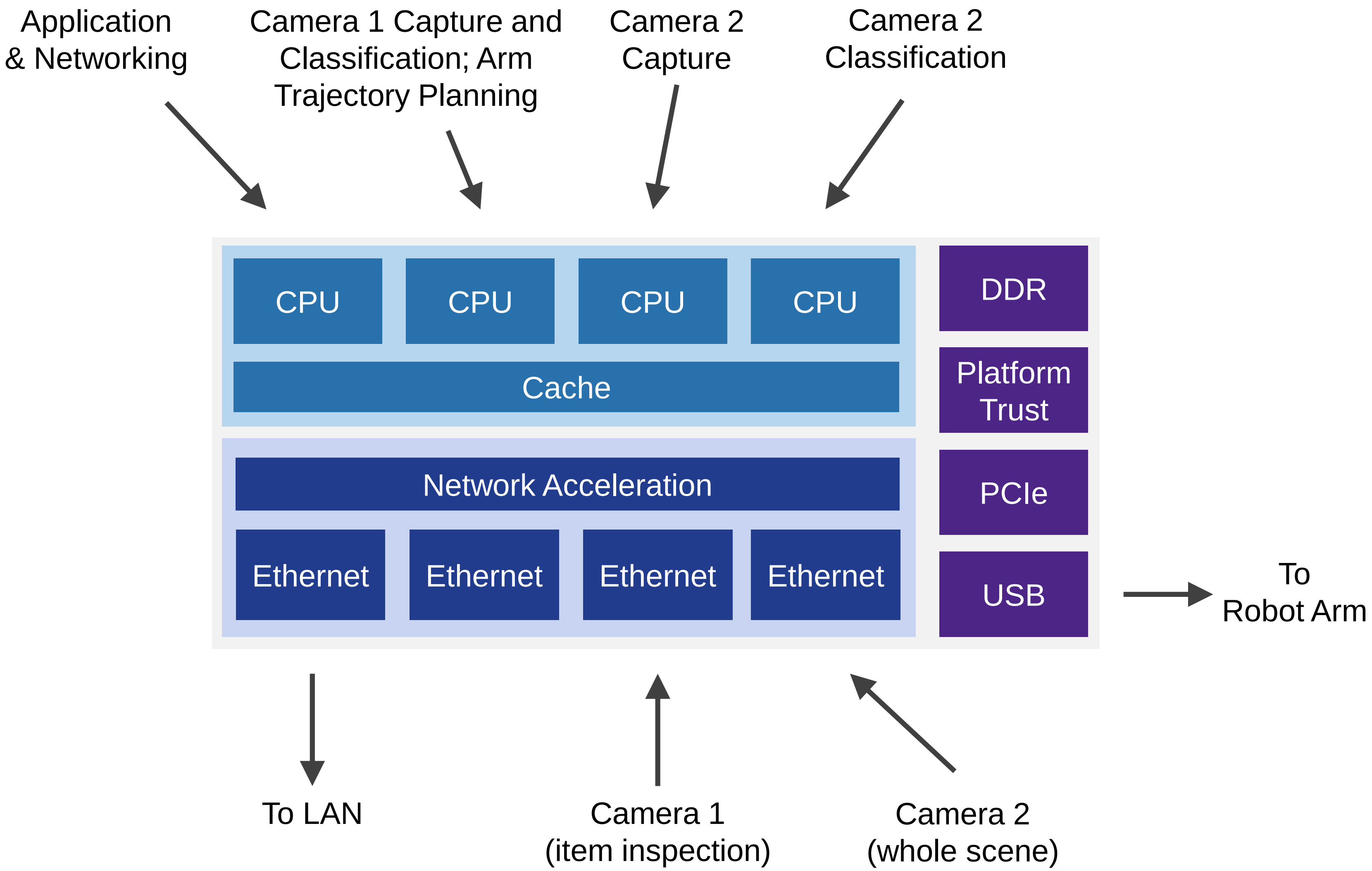
Camera 3

Camera 4

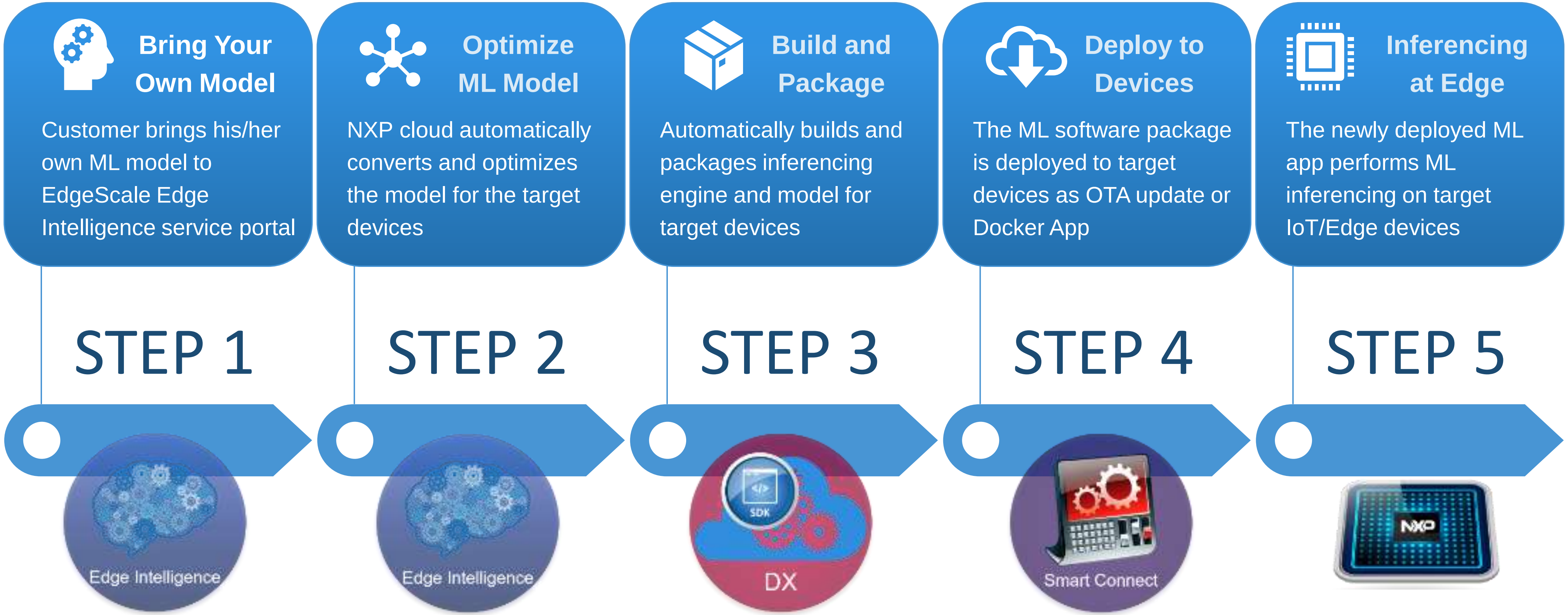
Backhaul

🟩 NXP Technology

Mapping Robot Arm Picker to Layerscape LS1046



AI/ML DX Example – Bring Your Own Model



Bring Your Own Model

Customer brings his/her own ML model to EdgeScale Edge Intelligence service portal

STEP 1

Optimize ML Model

NXP cloud automatically converts and optimizes the model for the target devices

STEP 2

Build and Package

Automatically builds and packages inferencing engine and model for target devices

STEP 3

Deploy to Devices

The ML software package is deployed to target devices as OTA update or Docker App

STEP 4

Inferencing at Edge

The newly deployed ML app performs ML inferencing on target IoT/Edge devices

STEP 5

AI/ML DX Example – Pick Your Own Engine

Pick Your Own Engine

EdgeScale Edge Intelligence service offers options for customer to pick his/her own inference engine framework

Build and Package

EdgeScale DX service automatically builds and packages inferencing engine and model for target devices

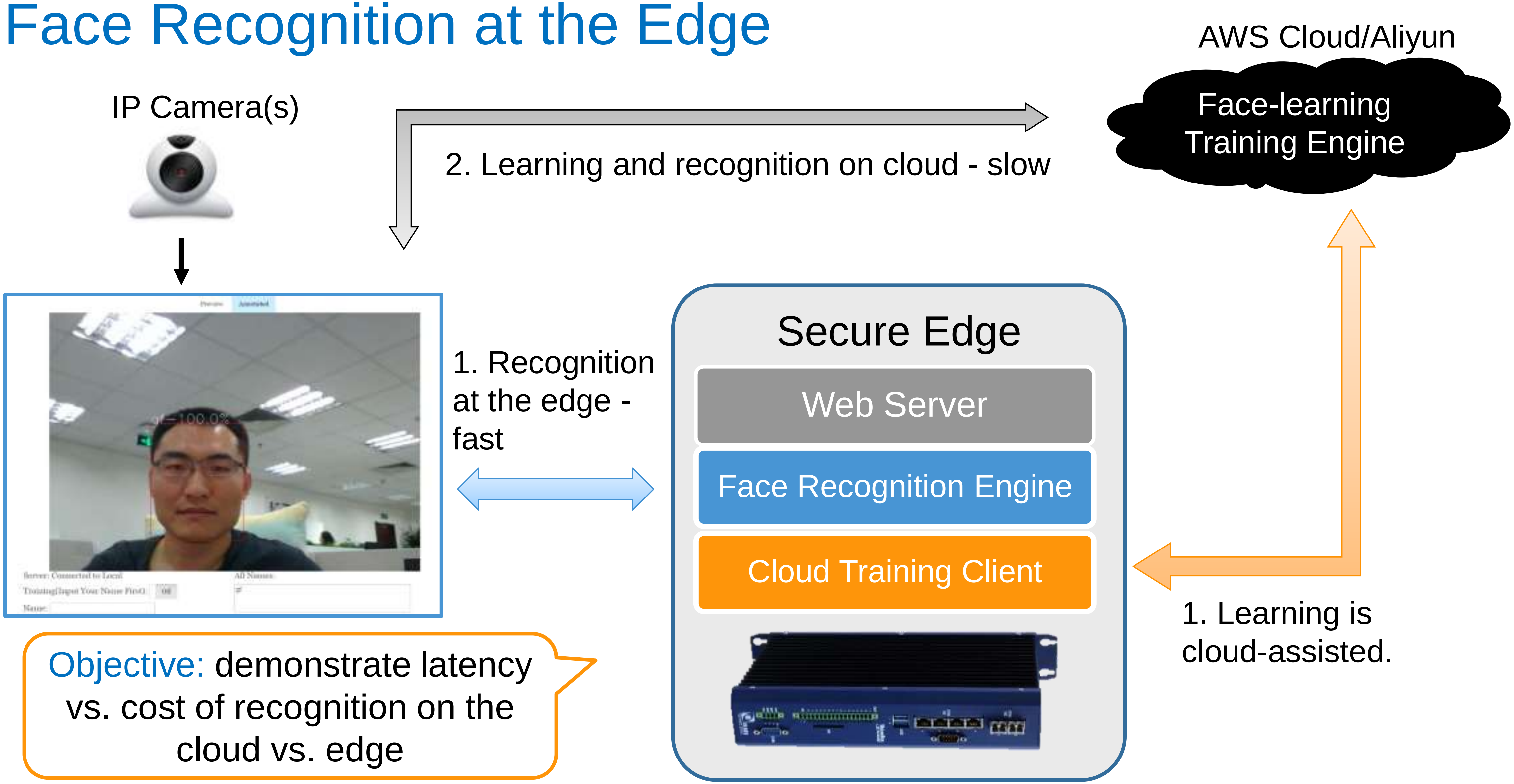
Deploy to Devices

EdgeScale Smart Connect service deploys the ML software package to target devices as OTA update or Docker App

Inferencing at Edge

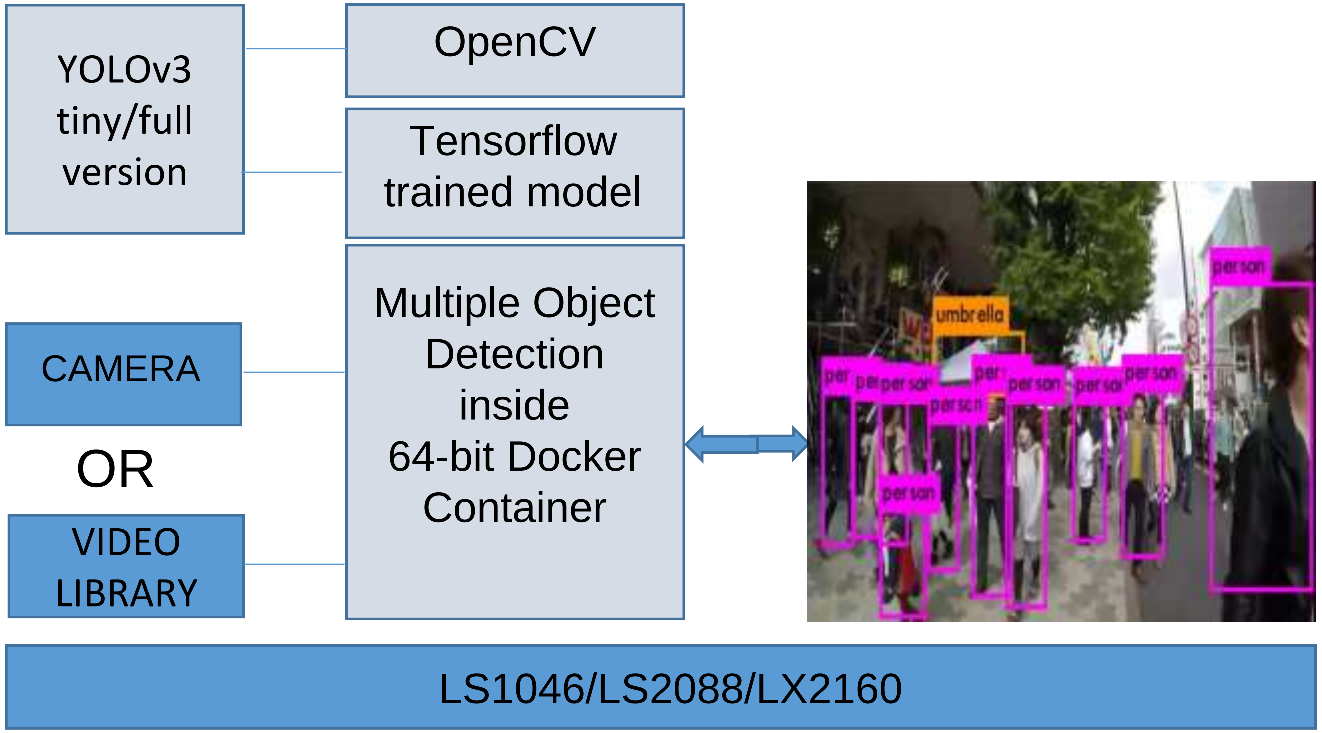
The newly deployed ML app performs ML inferencing on target IoT/Edge devices

Face Recognition at the Edge


AWS Cloud/Aliyun
Face-learning Training Engine
IP Camera(s)
2. Learning and recognition on cloud - slow
Preview
Annotated
100.0%
Server: Connected to Local
Training(Input Your Name First):
OK
All Names:
Name:
1. Recognition at the edge - fast
Secure Edge
Web Server
Face Recognition Engine
Cloud Training Client
1. Learning is cloud-assisted.

Objective: demonstrate latency vs. cost of recognition on the cloud vs. edge

People/Object Counting Using Machine Learning on Layerscape



YOLOv3 tiny/full version

OpenCV

Tensorflow trained model

CAMERA

OR

VIDEO LIBRARY

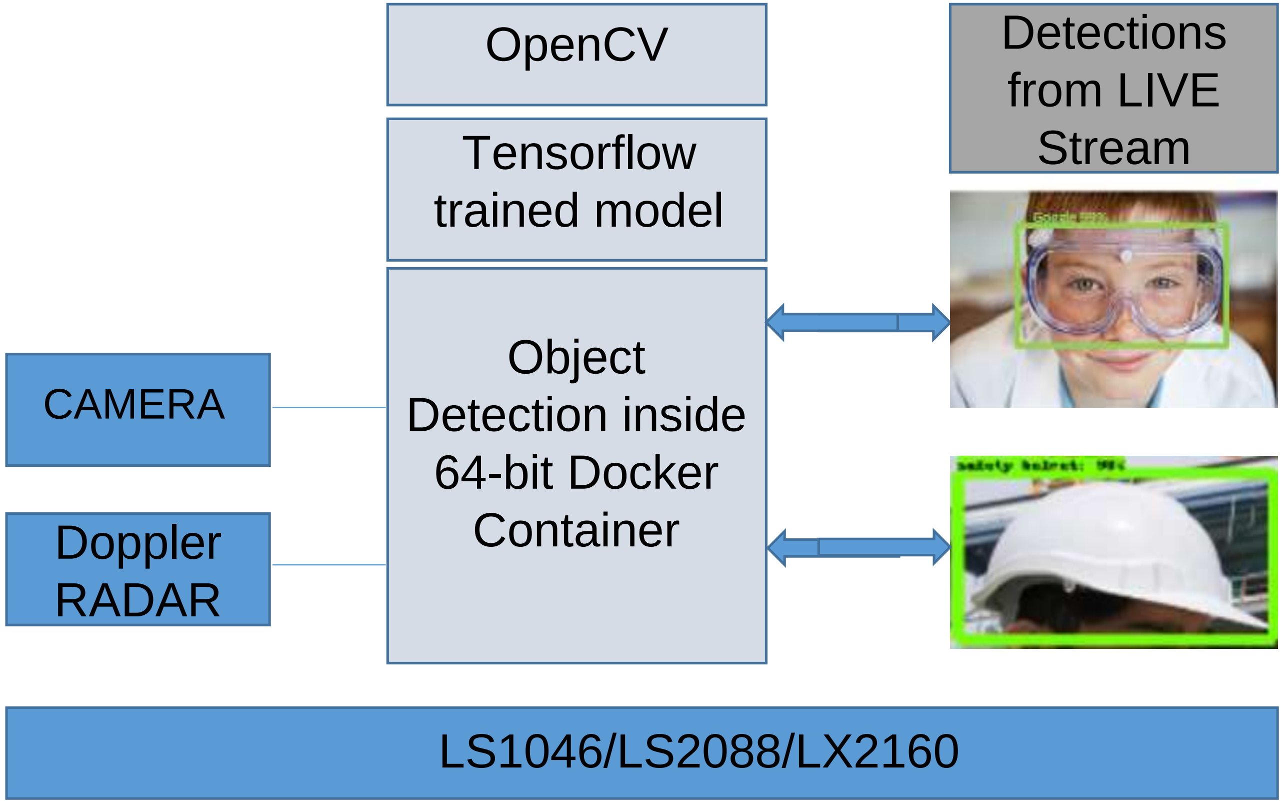
Multiple Object Detection inside 64-bit Docker Container

LS1046/LS2088/LX2160

What does it show/solve?

- Demonstrates Machine learning for People/Object counting in a given area of interest.
- Secure Surveillance: Can be used to count people/objects from Video database or real time Video stream
- Advanced Machine Learning: Detection of multiple persons and objects using tensorflow, OpenCV and YOLOv3 algorithm. Scalable FPS across 4/8/16-core ARM platforms.

Industrial Safety Using Machine Learning on Layerscape



OpenCV

Tensorflow trained model

Object Detection inside 64-bit Docker Container

Detections from LIVE Stream

CAMERA

Doppler RADAR

LS1046/LS2088/LX2160

What does it show/solve?

- Demonstrates Machine learning for object detection of Safety googles and safety helmet with highest accuracy
- Security: Factory Operators flagged at factory entrance without the presence of safety gears.
- Safety: Doppler Radar is used to set digital safety zone flagging operator to wear goggles
- Machine Learning: Detection of googles and helmet using tensorflow, OpenCV and a customized dataset.

Key Takeaways

AI has numerous industrial uses

NXP has the hardware, software, and ecosystem to enable you to get started today

The power of AI will only improve

SECURE CONNECTIONS
FOR A SMARTER WORLD