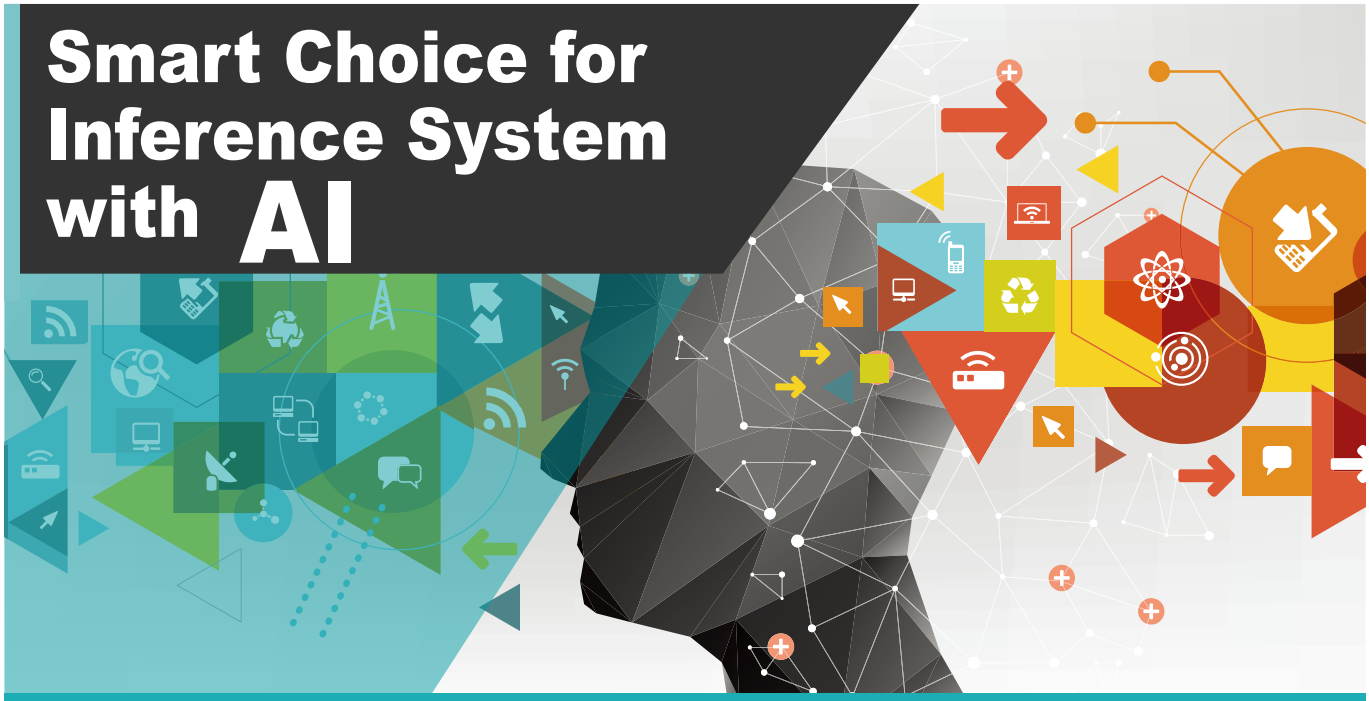
Smart Choice for Inference System with AI

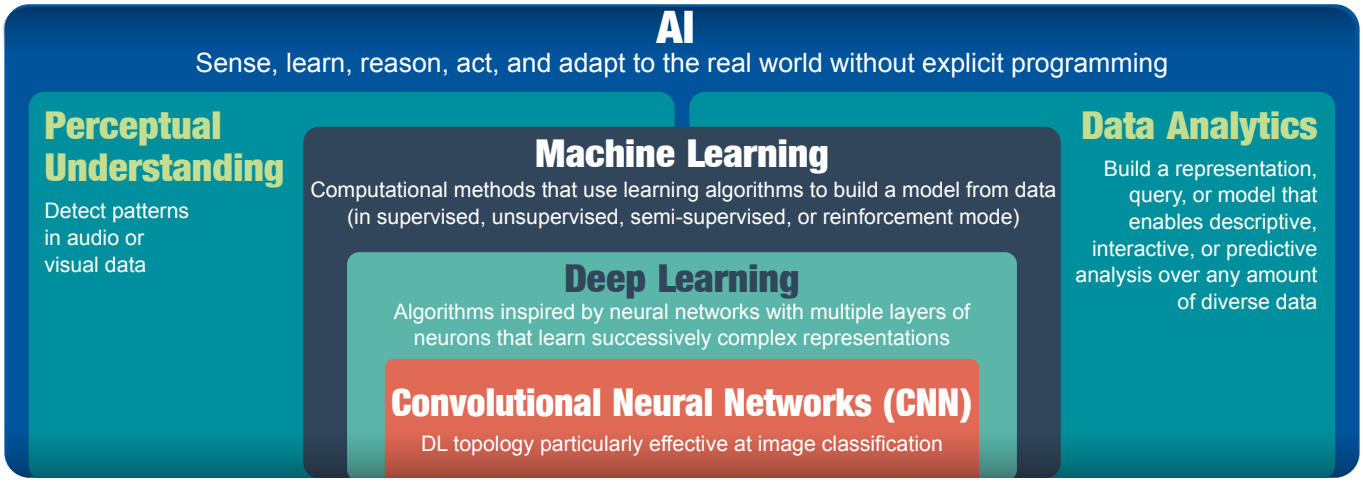


Artificial Intelligence, AI, is changing our lives from the past to the future. It enables machine learning by using a variety of training models to simulate and infer the status or appearance of objects. For example, the inference system with the video analysis model can perform face and vehicle license plate analysis for safety and security purposes.

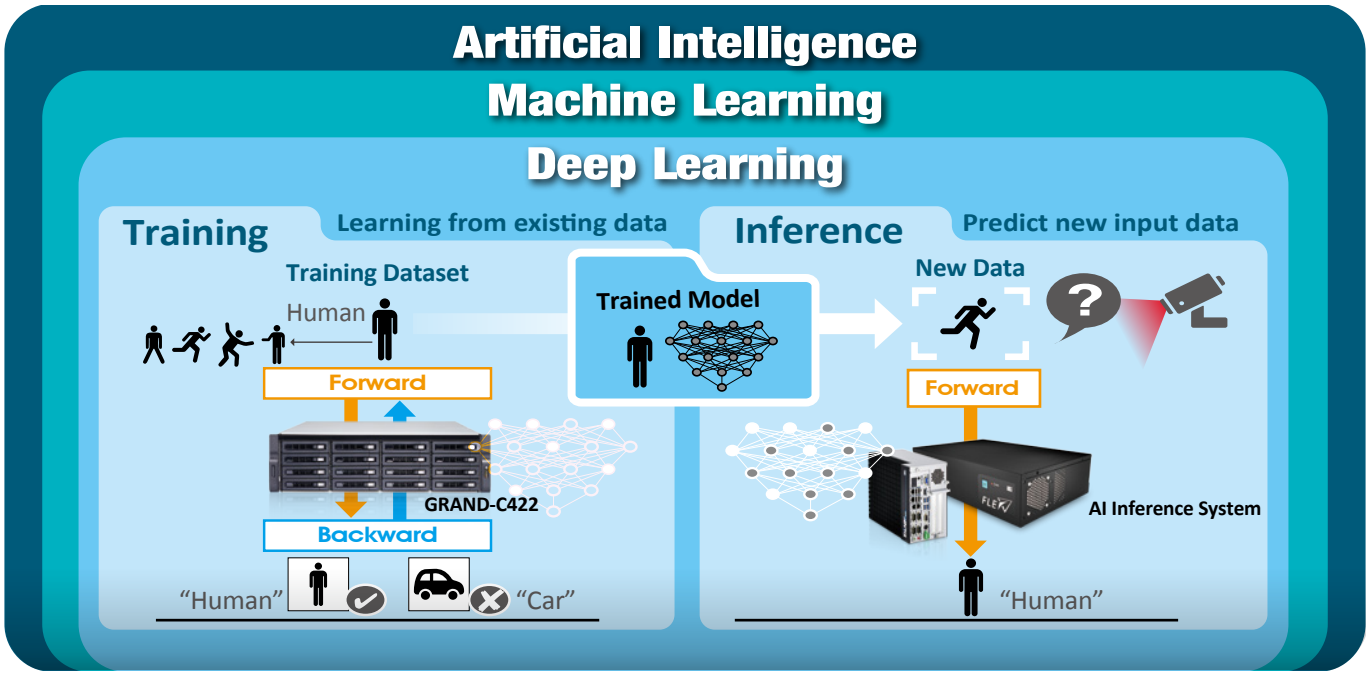
Today, most of AI technology still rely on the data center to execute the inference, which will increase the risk of real-time application for applications such as traffic monitoring, security CCTV, etc. Therefore, it's crucial to implement a low-latency. real-time edge computing platform.

Deep learning and inference

Deep learning is part of the machine learning method. It allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.Deep neural network and recurrent neural network architectures have been used in applications such as object recognition, object detection, feature segmentation, text-to-speech, speech-to-text, translation, etc.In some cases the performance of deep learning algorithms can be even more accurate than human judgement.



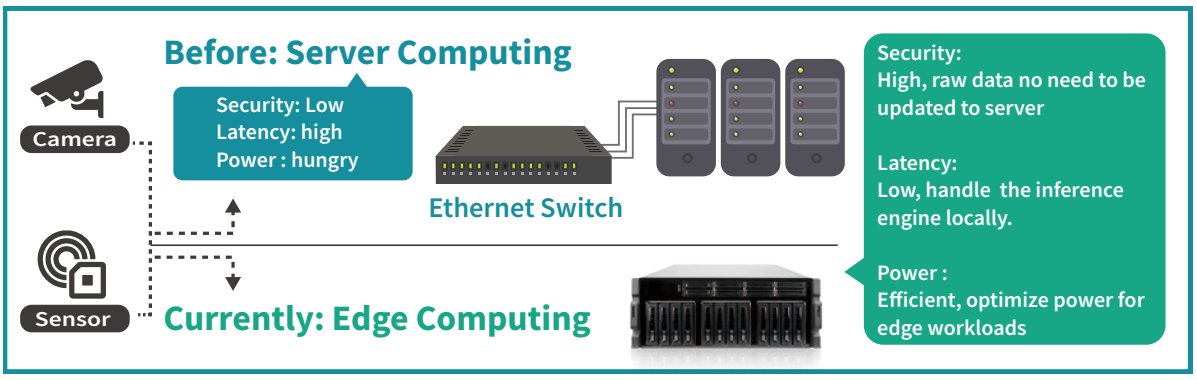
In the past, machine learning required researchers and domain experts knowledge to design filters that extracted the raw data into feature vectors. However, with the contributions of deep learning accelerators and algorithms, trained models can be applied to the raw data, which could be utilized to recognize new input data in inference.



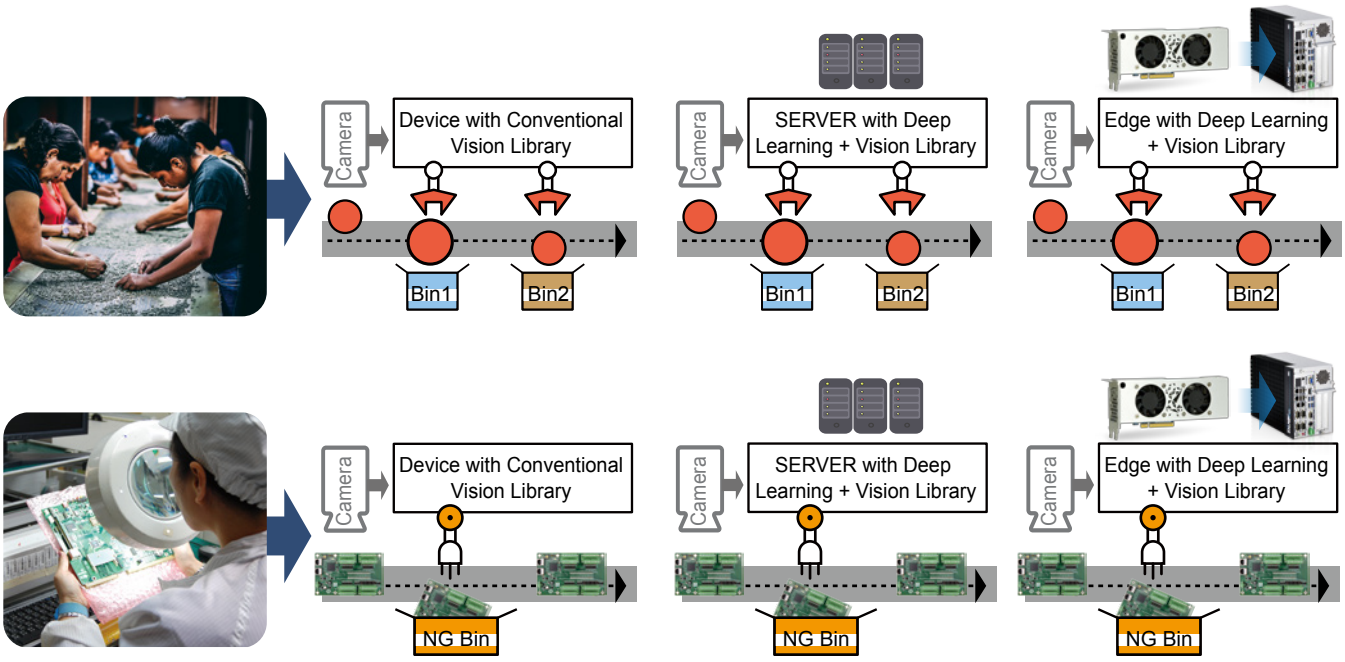
Edge Computing

The advantages of edge computing:

- Reduce data center loading, transmit less data, reduce network traffic bottlenecks.
- Real-time applications, the data is analyzed locally, no need long distant data center.
- Lower costs, no need to implement sever grade machine to achieve non complex applications.



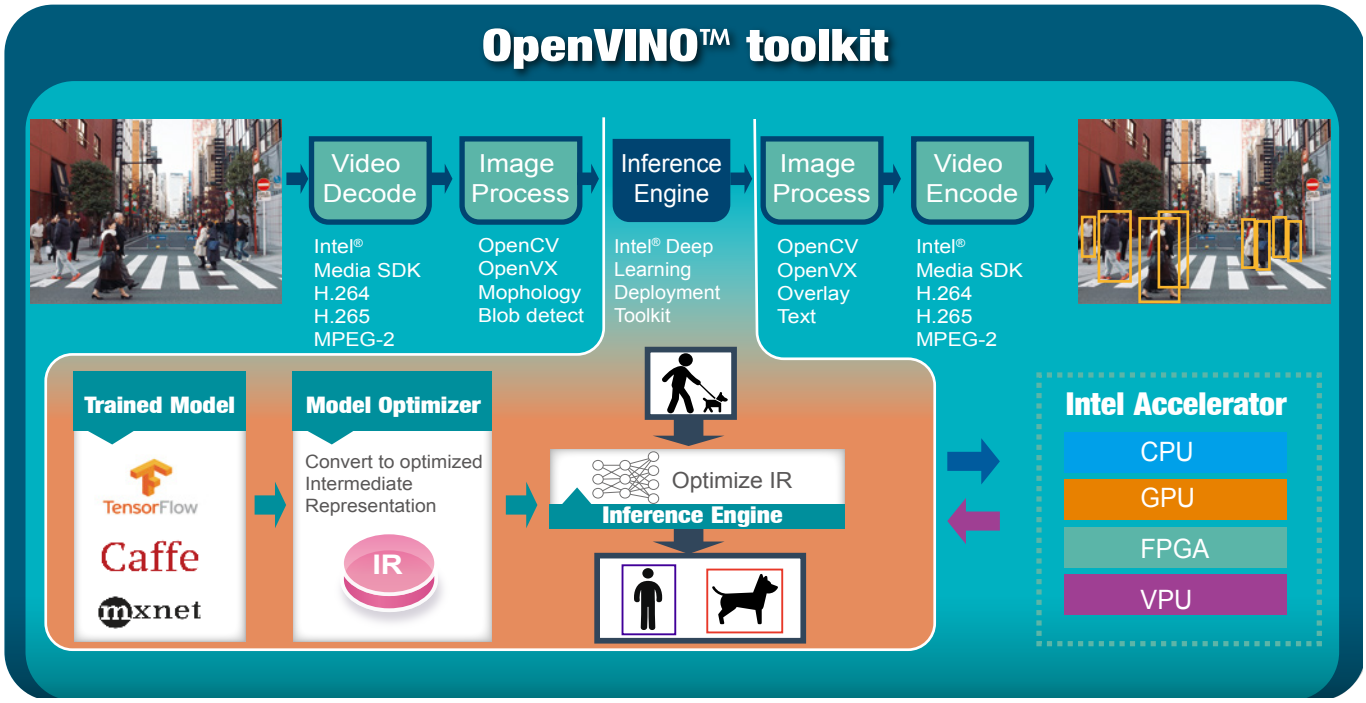
Human Sorting	Machine Vision	Deep Learning Server	Deep learning Edge
Sorting by human bare-eyes, the standard can not fix due to different person. Low accuracy, low productivity	Training and sorting by conventional machine vision method,need well-trained engineer to teach and modify the inspection criteria. Low latency, but low accuracy if product has too much variations.	Implement deep learning method for training and sorting, reduce human bias and machine vision training process. High CTO, high latency, power hungry, high power consumption	Implement deep learning method for training and sorting, reduce human bias and machine vision training process. Low latency, low CTO, power efficiency, low power consumption



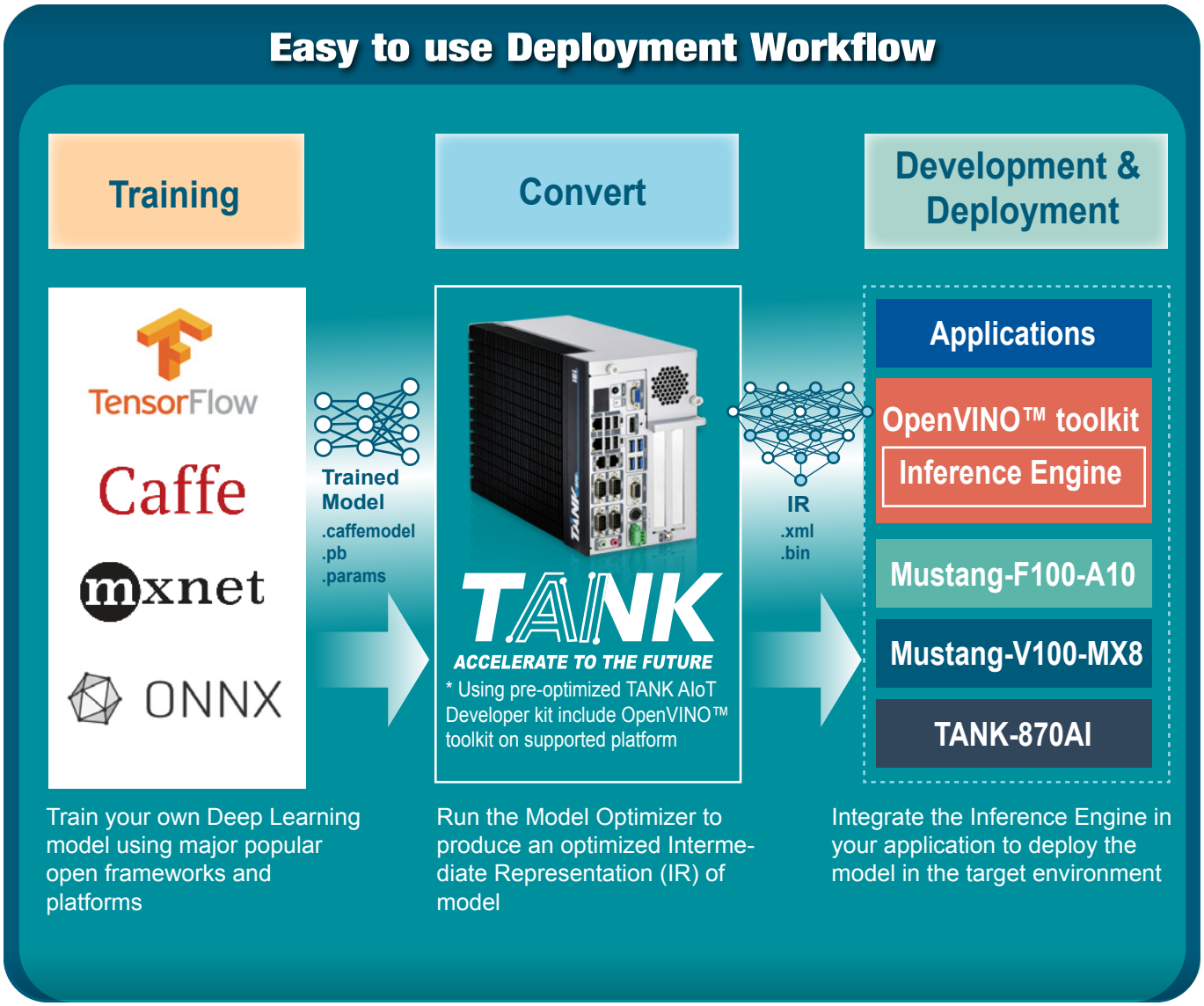
Intel® Distribution of OpenVINO™ toolkit

Intel® Distribution of OpenVINO™ toolkit is based on convolutional neural networks (CNN), the toolkit extends workloads across multiple types of Intel® platforms and maximizes performance.

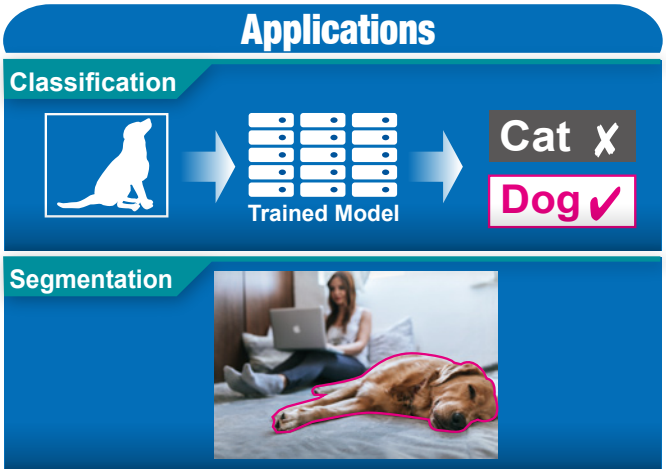
It can optimize pre-trained deep learning models such as Caffe, MXNET, and ONNX Tensorflow. The tool suite includes more than 20 pre-trained models, and supports 100+ public and custom models (includes Caffe*, MXNet, TensorFlow*, ONNX*, Kaldi*) for easier deployments across Intel® silicon products (CPU, GPU/Intel®Processor Graphics, FPGA, VPU).

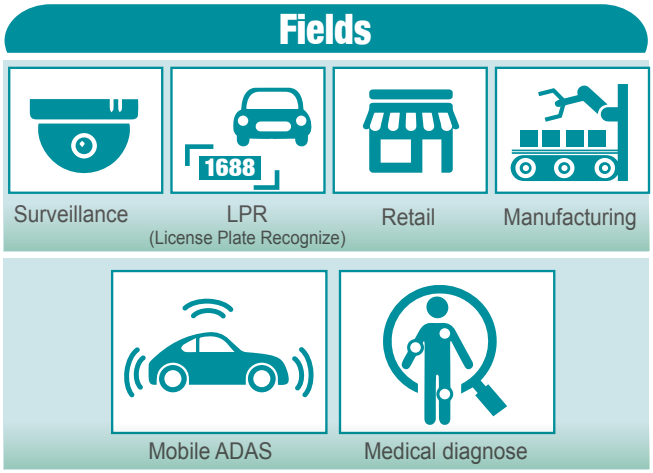


Software



- **Operating Systems**
  Ubuntu 16.04.3 LTS 64bit, CentOS 7.4 64bit, Windows 10 64bit
- **OpenVINO™ toolkit**
  – Intel® Deep Learning Deployment Toolkit
    • Model Optimizer
    • Inference Engine
  – Optimized computer vision libraries
  – Intel® Media SDK
  – Current Supported Topologies: AlexNet, GoogleNet V1/V2/V4, Yolo Tiny V1/V2, Yolo V2/V3, SSD300,SSD512, ResNet-18/50/101/152, DenseNet121/161/169/201, SqueezeNet 1.0/1.1, VGG16/19, MobileNet-SSD, Inception-ResNet-v2,Inception-V1/V2/V3/V4,SSD-MobileNet-V2-coco, MobileNet-V1-0.25-128, MobileNet-V1-0.50-160, MobileNet-V1-1.0-224, MobileNet-V1/V2, Faster-RCNN
  (more variants are coming soon)
- **High flexibility, develop on OpenVINO™ toolkit structure which allows trained data such as Caffe, TensorFlow, and MXNet to execute on it after convert to optimized IR.**

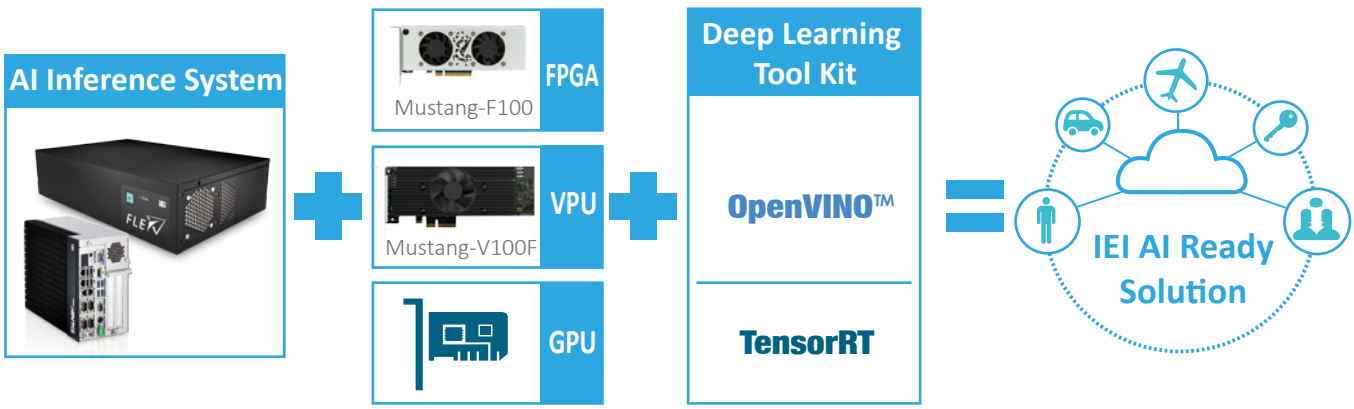




		TS-X77 with GPU	GRANG-C422 with GPU	TANK-870AI with Mustang-F100-A10	TANK-870AI with Mustang-V100-MX8	FLEX-BX200-Q370 with Mustang-F100-A10	FLEX-BX200-Q370 with Mustang-V100-MX8
Applications	Inference Training	O	O				
	Inference Engine	O	O	O	O	O	O
	Image Classification	O	O	O	O	O	O
	Image Localization	O	O	O	O	O	O
Features	Energy Efficent			O	O	O	O
	Low-latency.			O	O	O	O
	Compact Size			O	O	O	O

IEI AI Ready Solution Accelerates Your AI Initiative

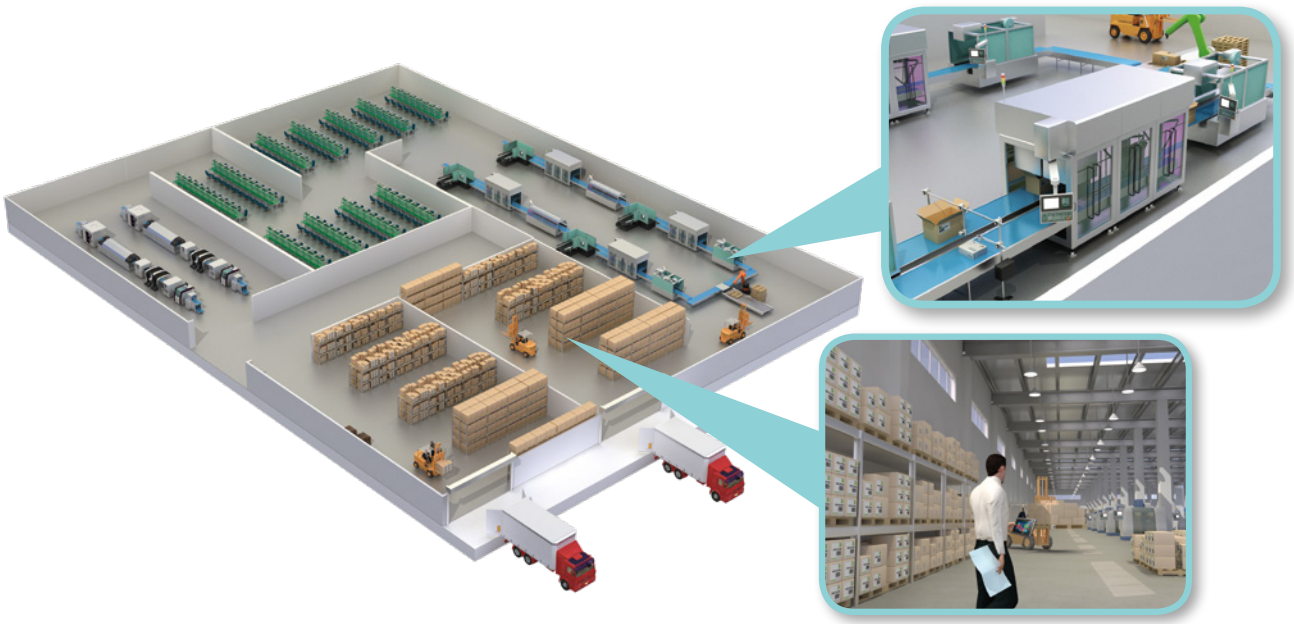
The FLEX-BX200 and TANK-870AI dev. kit are AI hardware ready system ideal for deep learning inference computing to help you get faster, deeper insights into your customers and your business. IEI's FLEX-BX200 and TANK-870AI dev. support graphics cards, Intel® FPGA acceleration cards, and Intel® VPU acceleration cards, and provides additional computational power plus end-to-end solution to run your tasks more efficiently. With the Intel® OpenVINO toolkit and NVIDIA TensorRT, it can help you deploy your solutions faster than ever.



Industrial Manufacturing

• Industrial automation

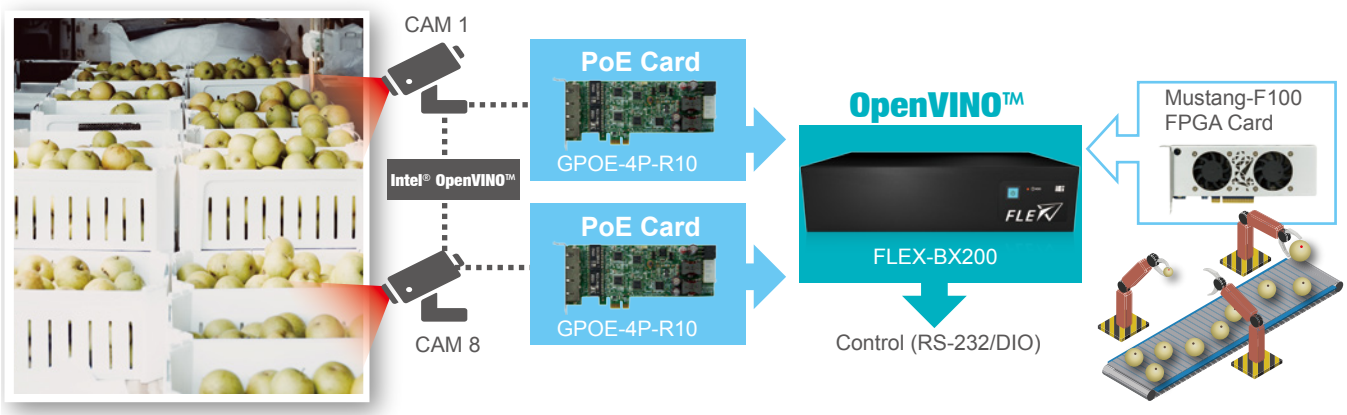
Mustang series solutions help enable intelligent factories to be more efficient on work order schedule arrangements. In today's production line, sticking to manufacturing schedules is becoming more and more important for business efficiency. From raw material storage to fabrication and complete products, all information from factory such as manufacturing equipment process time and warehouse storage status are essential to achieve production goals. Solutions based on AI technology can produce more detailed, accurate, and meaningful digital models of equipment and processes for product management.



• Machine Vision for Sorting and Grading of Agricultural Products

Agricultural products are valued by their appearance. The color indicates parameters like ripeness, defects, etc. The quality decisions vary among the graders and often inconsistent. Machine vision technology offers the solution for all these problems.

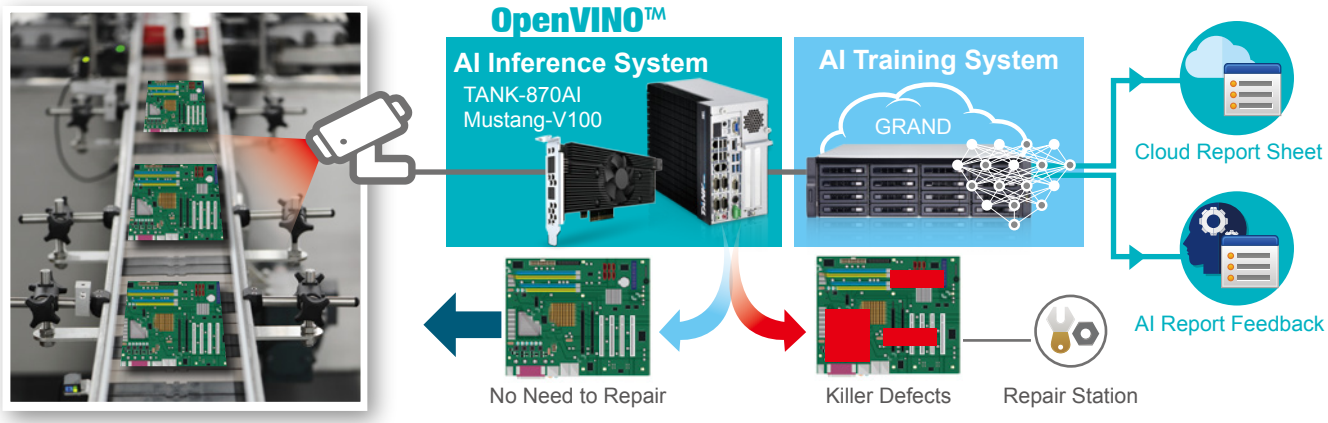
The FLEX series designed for machine vision market has four PCIe 3.0 expansion slots for installing motion controller cards, GP GPU/FPGA/VPU cards and the PoE Ethernet card which is developed by IEI and has four GbE Power over Ethernet (PoE) ports compliant with IEEE 802.3af for direct connection to CCTV cameras without needing separate power.



• AOI Defect Classification

During the manufacturing process, defects could be introduced and harmful to the quality. It is necessary to classify the defects detected by AOI machine appropriately especially killer defects. The higher accuracy to classify defects, the less cost spent on review and repair station.

The TANK AIoT Dev. Kit features rich I/O and dual PCIe slots (x16) to support add-ons like the Acceleration cards (Mustang-F100-A10 & Mustang-V100-MX8) or the PoE to enhance the defects detected performance.



Retail

• Smart Retail

Using the Mustang series for computer vision solutions at the edge of retail sites can quickly recognize the gender and age of the customers and provide relevant product information through digital signage display to improve product sales and inventory control. Self-checkout can reduce human resource cost so that retail owners can spend more resources on promoting products and understanding business patterns.

In addition, it can help to analyze customer's in-store behavior, and provide customer information based on gender and age to facilitate product positioning. Quickly converting the business intelligence gained and help build better business practices and increase profitability.

Self-checkout

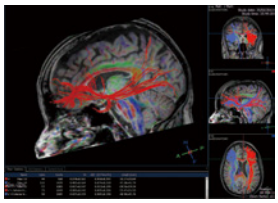
Smart Retail

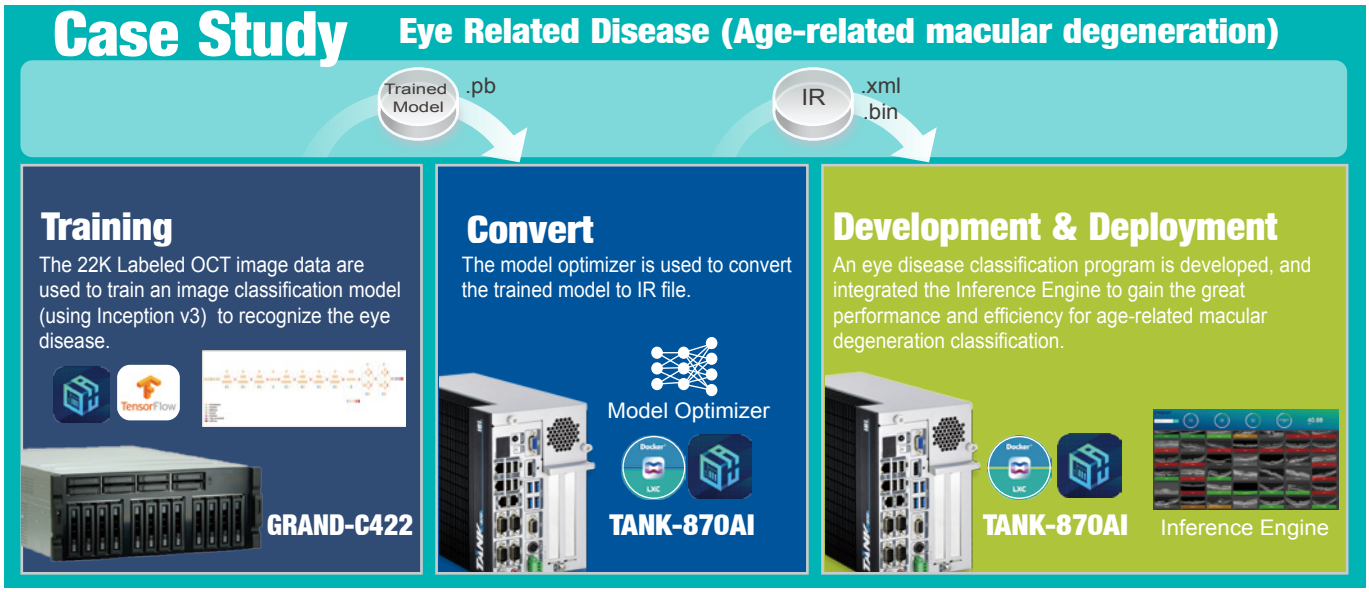

Interactive digital signage

Medical

• Medical Diagnostics

With AI based technology, healthcare and medical centers can diagnose, locate and identify suspicious areas such as tumors and other abnormalities more quickly and accurately. Using segmentation technology and trained models on the Mustang series can be used to locate and identify abnormalities with a high degree of accuracy helping doctors and researchers quickly serve the patient.





Transportation

• Numerous Vehicle License Plate Analysis

Efficient road tolling and parking reduces fraud related to non-payment, makes charging effective, and reduces required manpower to process. Vehicle license plate analysis can be deployed on highways for electronic toll collection, and can be implemented as a method of cataloguing the movement of traffic as well as provide enhanced security by establishing data on suspicious vehicles in a more efficient way.

Traffic management

LPR