OCP TAIWAN DAY

Road to 5G · AI · Edge Computing

OPEN Compute Project

TAIWAN Regional Community

OPEN Compute Project

TAIWAN Regional Community

OCP TAIWAN DAY

Road to 5G · AI · Edge Computing

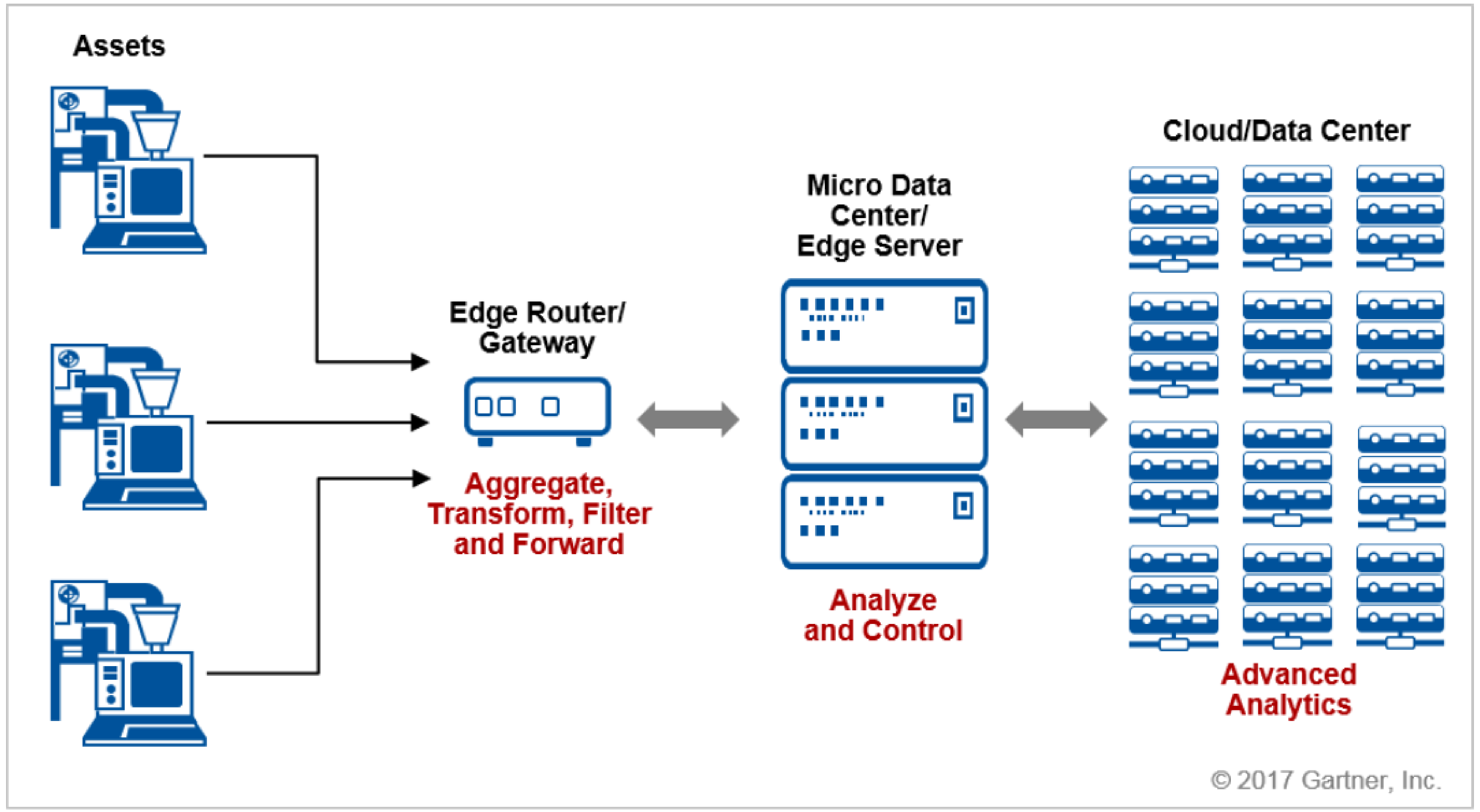
AI Edge Computing

Chilung Wang,
Division Director of Data Center Architecture and Cloud Applications Division,
Information and Communications Laboratories, ITRI

工業技術研究院
Industrial Technology Research Institute

What Is Edge Computing?

- Gartner's definition is:
  - Edge computing is a part of a distributed computing topology where information processing is located close to the edge, where things and people produce or consume that information.
- Cloud computing and edge computing are complementary, rather than competitive or mutually exclusive.



Category Based on Location of Computing Capabilities	Computing Capabilities	Type of Analytics	Preferred Response Time to Event/Activity	System Capabilities	Location
Embedded Computing	Low to high	Simple event processing	Microseconds to seconds	Analyze and control	Static or dynamic
Gateway	Low	Simple event processing; event stream processing (low velocity only)	N/A	Analyze	Static or dynamic
Edge Server	Medium	Event stream processing; complex-event processing; application hosting	Seconds to minutes	Analyze and control	Static
Edge Cluster (hosted in a micro data center)	High	Complex-event processing; batch processing; application hosting	Seconds to minutes to hours	Analyze and control	Static or dynamic

Source : Gartner

Edge Computing vs. Cloud Computing

Items	Cloud Computing	Edge Computing
Time-sensitivity	Low	High (Low latency)
Responsive	Slow	Quick
Delay Jitter	High	Low
Real-time analysis request	Low	High
Distance to data	Multiple hops	Most of them are one hop
Data-intensive	Deal with huge collected data	Dedicate to specific data
Network topology	Multiple hops connectivity	Mesh network based
Security requirement	High	Very High
Architecture	Centralized	Decentralized (dedicated processing for a single, specific task)

What Does a $12 Camera Entail?

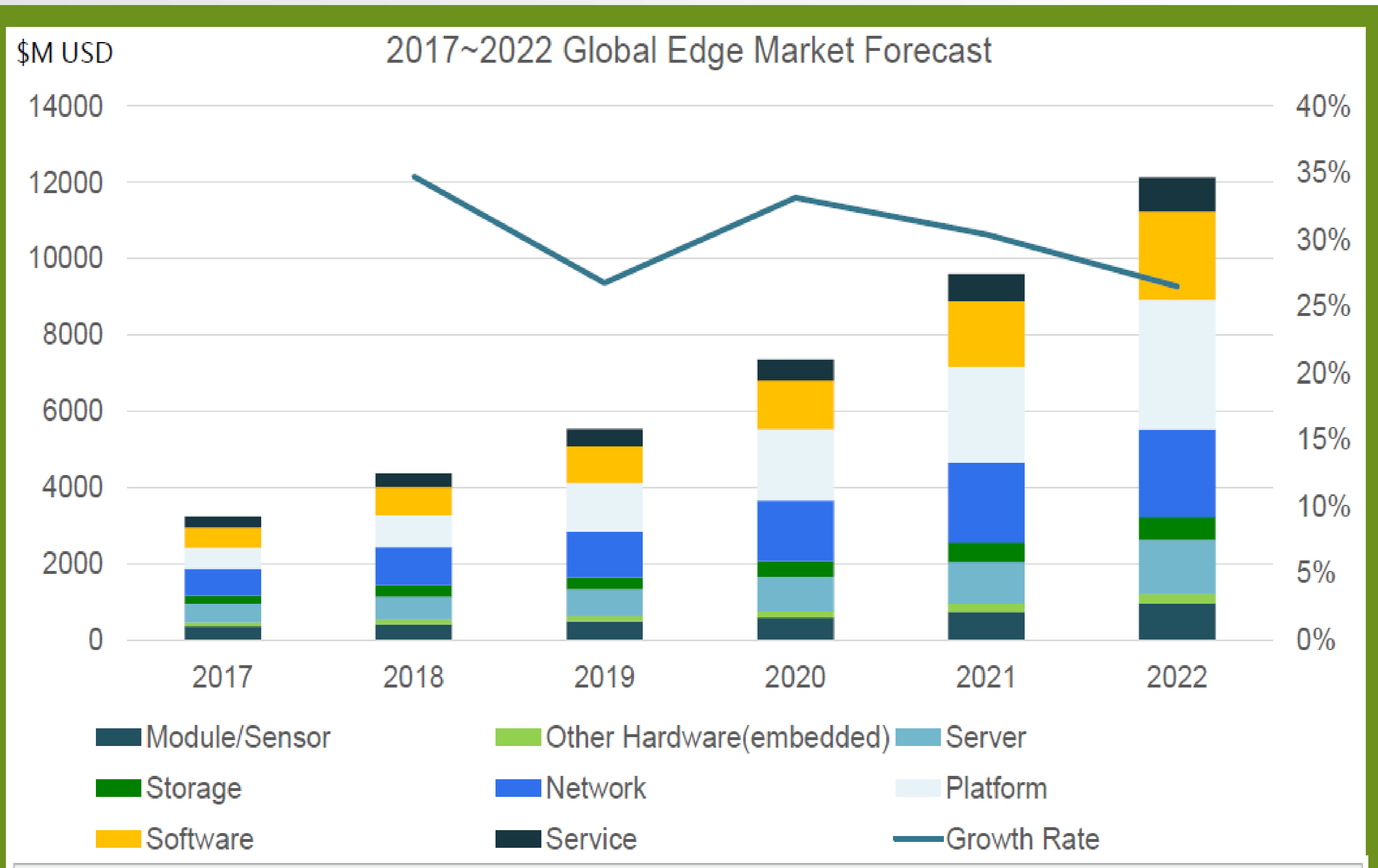
- The camera consumes power.
  - The $12 ZOSI camera takes about 6W. As a rule of thumb, electricity in the US costs about $1 per watt per year, so we get a three-year power cost of about $9.
- The camera needs a network connection.
  - Local cabling or Wi-Fi may just have capital costs, but almost all Internet access technologies cost real money.
  - FTTH (Fiber To The Home) : $0.10 per TB (Terabyte), Cable/DSL: $8-20 per TB, T1: $100 per TB.
- The camera needs storage in the cloud.
  - $12.50/TB-month for standard storage and $2.5/TB-month for glacier storage.
- The camera needs computing in the cloud.
  - YOLO inference on AWS GPUs (Graphics Processing Units) might cost about $0.58 per million frames.

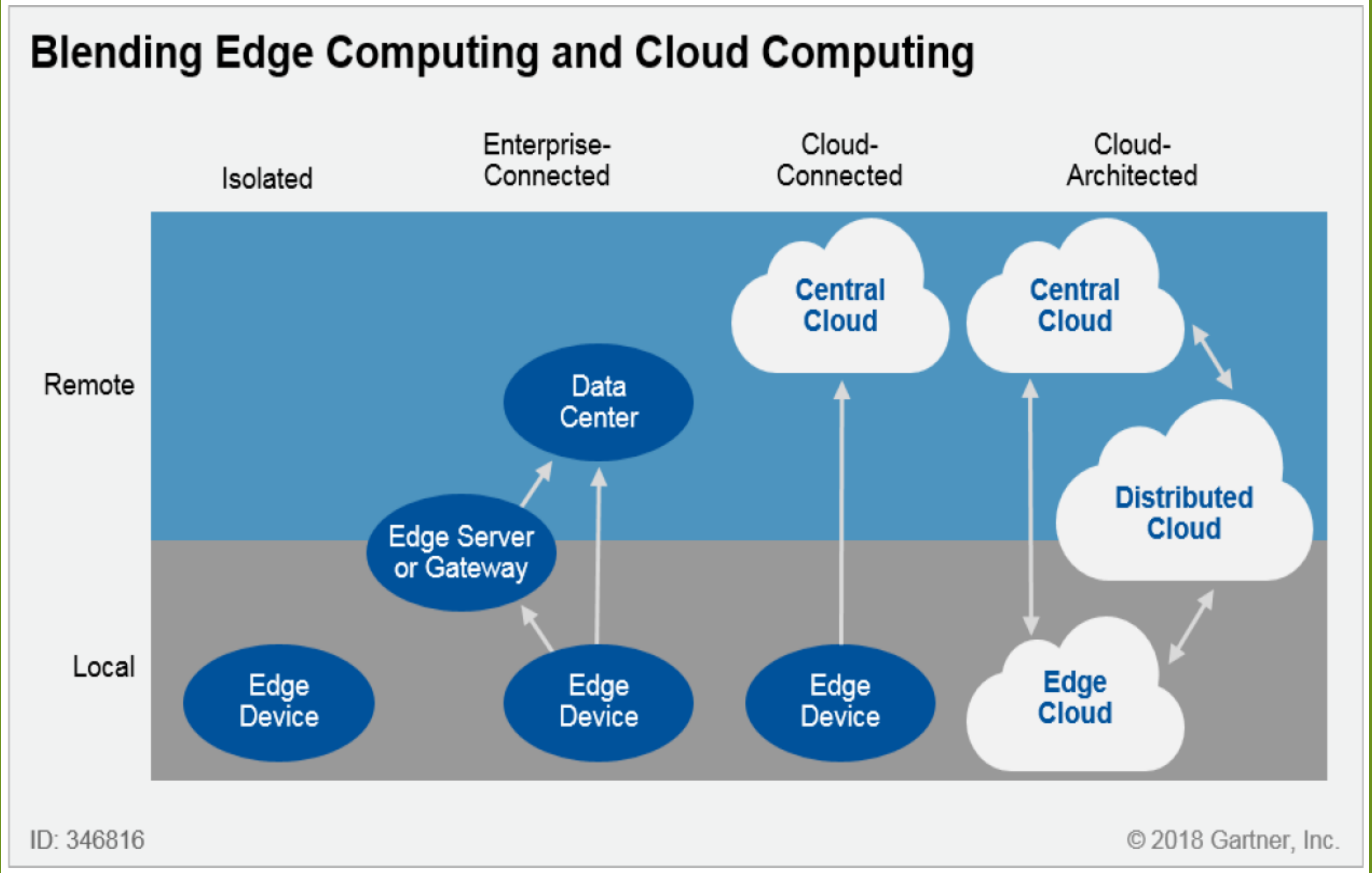
Data stream	Network Cost		Storage Cost		Compute Cost	Total Cost
	Fiber	Conv	1 day	1 month		
8Mp raw	$9,500	$950,000	$79,000	$476,000	$3,300	$92,000-$1,430,000
H.264 p60	$47	$4,700	$400	$2,400	$3,300	$3,700-$10,000
H.264 @ 1 fps	$1	$79	$7	$39	$55	$62-$173
H.264 @ 1 fpm	$0	$1.30	$0.11	$0.66	$0.91	$1-$3

Lessons:
- On-device or edge computing for video frame filtering is a must.
- Privacy concern for video cameras is rising.

Market Opportunities

- **IDC:** The global edge computing market size in 2018 is estimated to be USD **4.36B** in 2018, and will grow to USD **12.1B** in 2022, at a CAGR of **26%**.
- **MarketsandMarkets**
  - The AI in computer vision market is estimated at USD **3.62B** in 2018 and is projected to grow to USD **25.43B** by 2023, at a CAGR of **45.74%** during the forecast period.
  - The surveillance video analytics market size is expected to grow from USD **2.77B** in 2017 to USD **8.55** Billion by 2023, at a CAGR of **21.5%** during the forecast period.
- Opportunity: **AI + Edge computing**



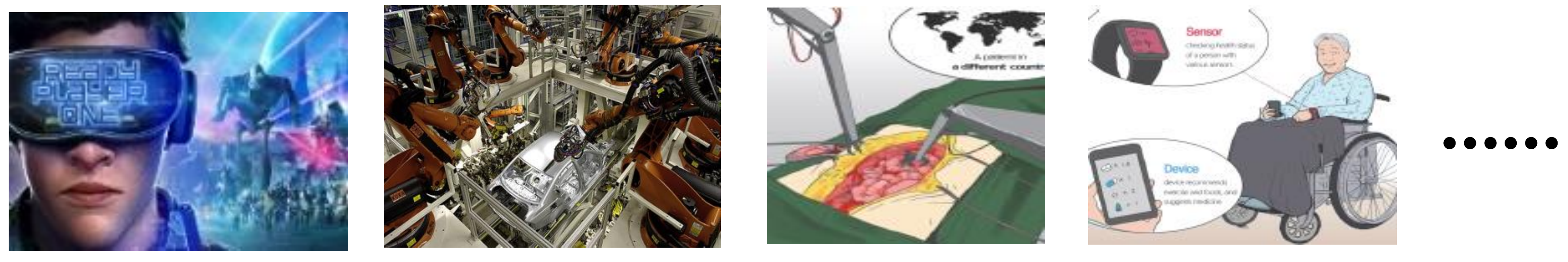


Source: Gartner (March 2018)

What is AI Edge Computing?

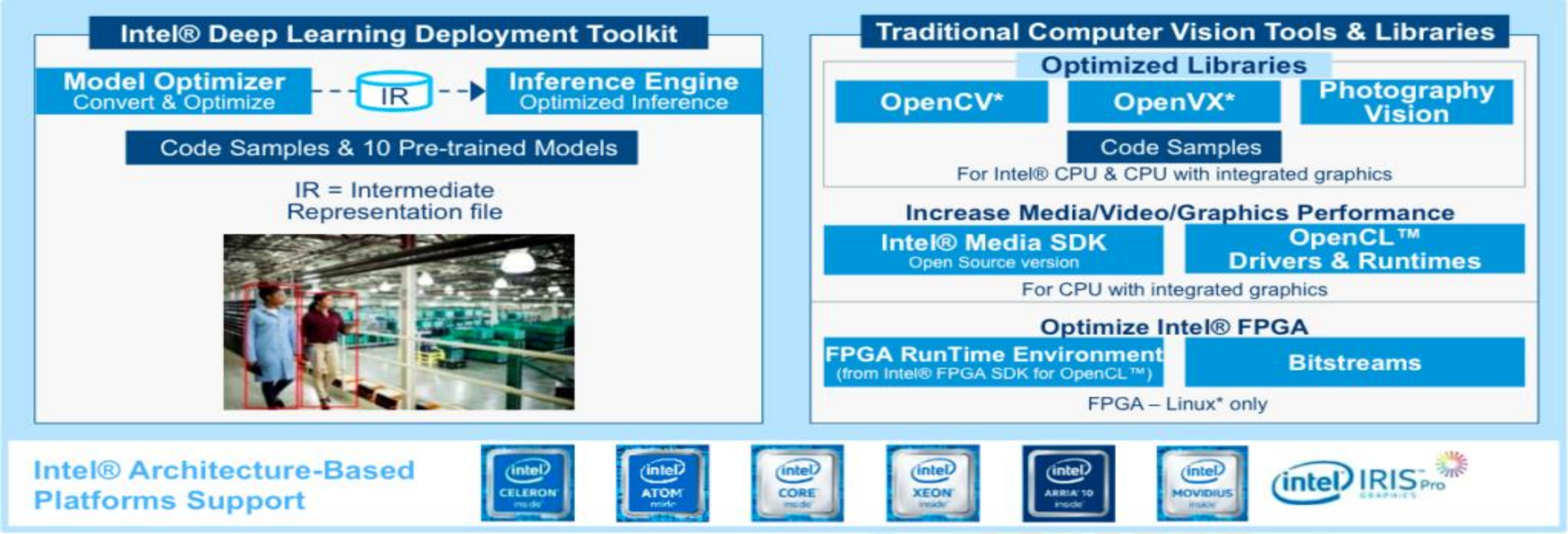
Key building blocks:

- Network: 5G or high-speed wired/wireless network
- Computing: edge data center
- Target applications: Deep Neural Network-based video processing



Intel's OpenVINO

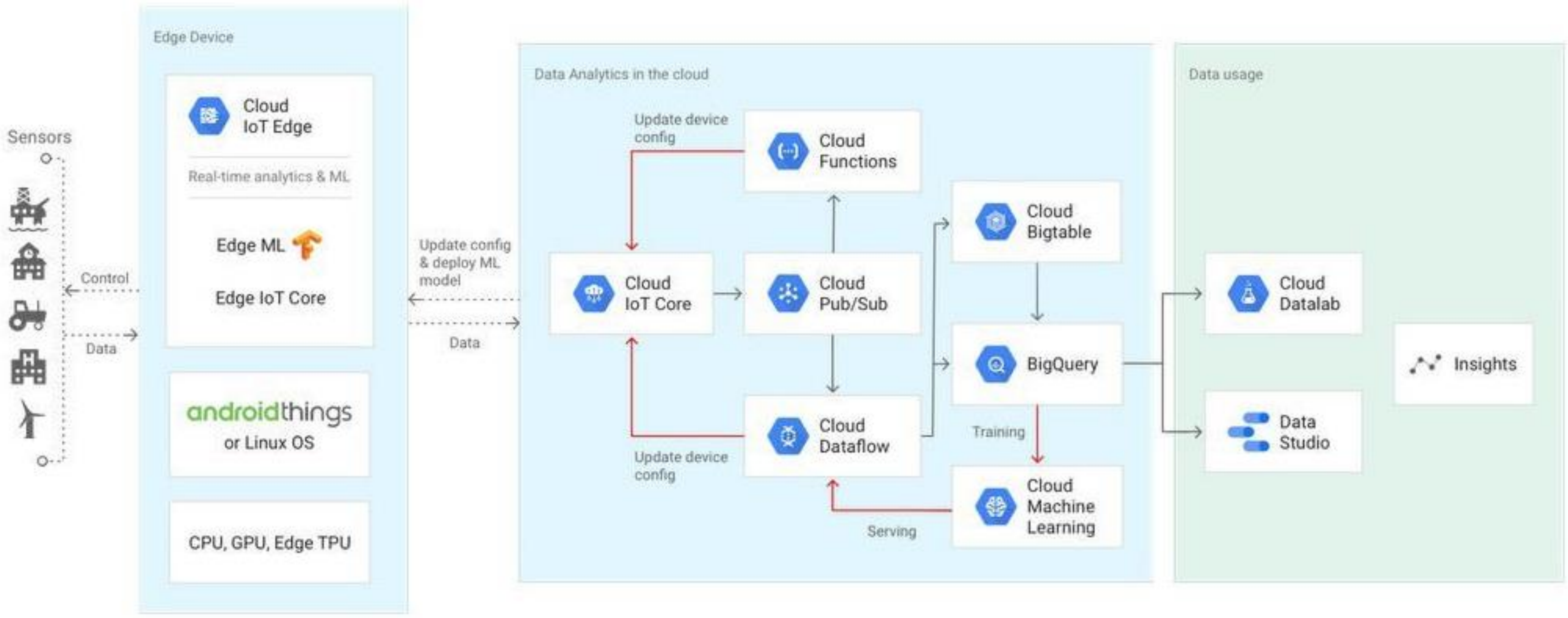
- By combining the technologies from four acquired companies, Nervana, Movidius, MobileEye and Altera, Intel creates the OpenVINO development kit.
  - OpenVINO integrate open source software such as OpenCV, OpenVX, and OpenCL and support hardware acceleration chips such as CPU, GPU, FPGA, ASIC (IPU, VPU) as well as deep learning frameworks such as Caffe, TensorFlow, Mxnet, and ONNX.
- OpenVINO is mainly used for inference. In addition to providing hardware acceleration, it also provides Model Optimizer to speed up inference performance by 10 to 100 times.



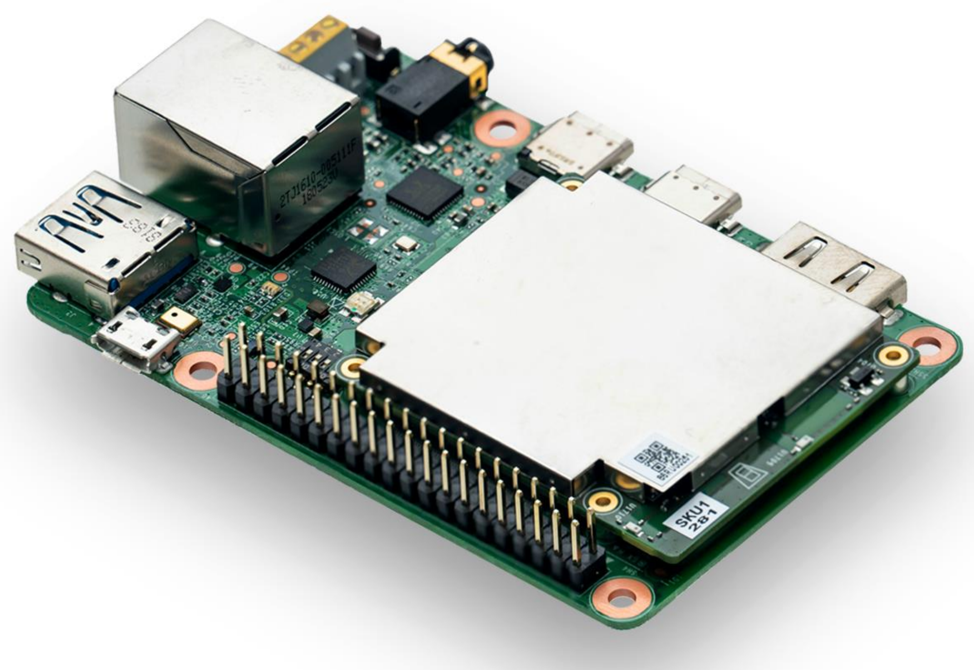
Google's Edge TPU (Tensor Processing Unit)

- End-to-end AI infrastructure: Edge TPU complements Cloud TPU and Google Cloud services
- High performance in a small physical and power footprint
- Co-design of AI hardware, software and algorithms
- A broad range of applications: predictive maintenance, anomaly detection, machine vision, robotics, voice recognition, and many more

Google Releases Edge TPU and Cloud IoT Edge

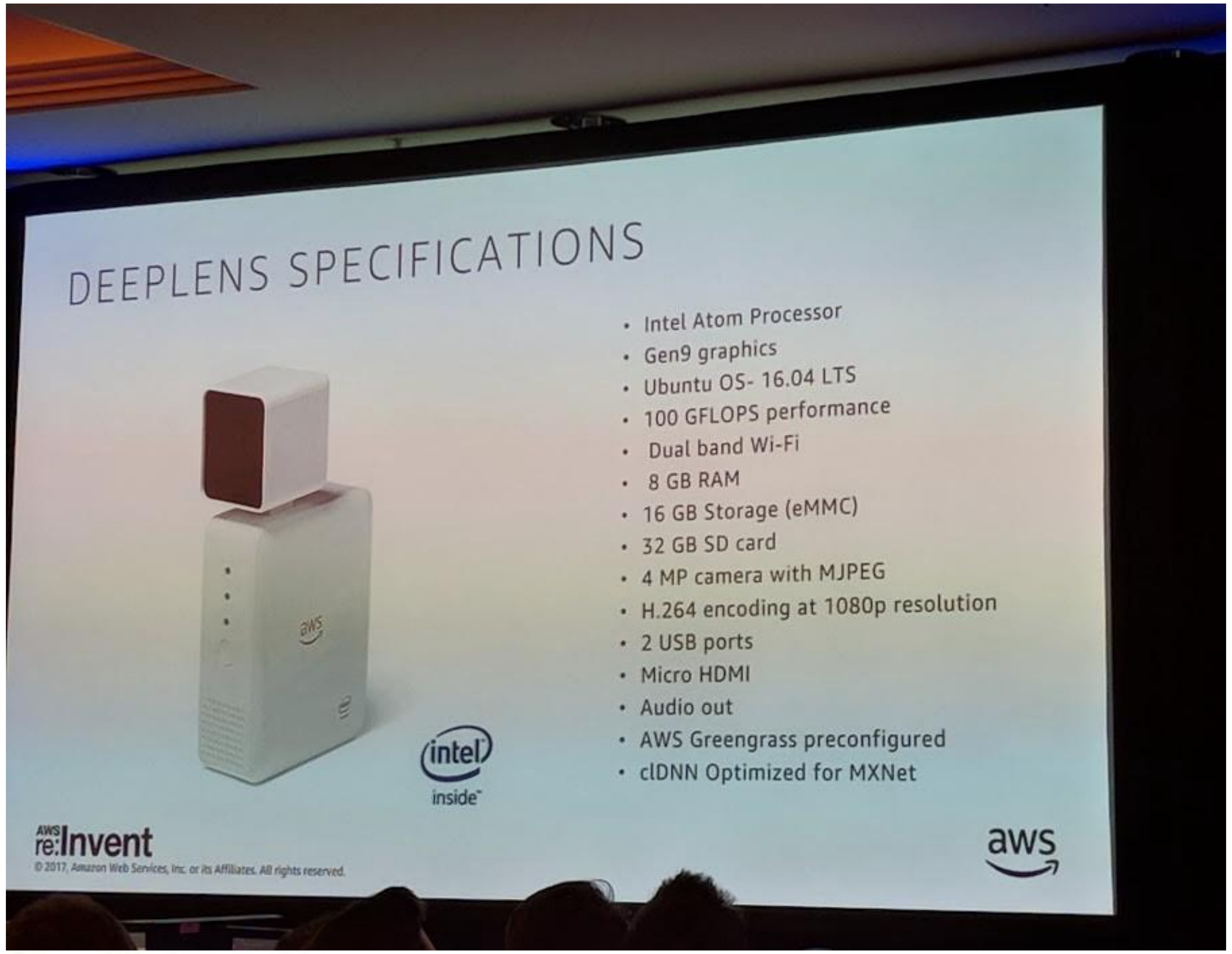


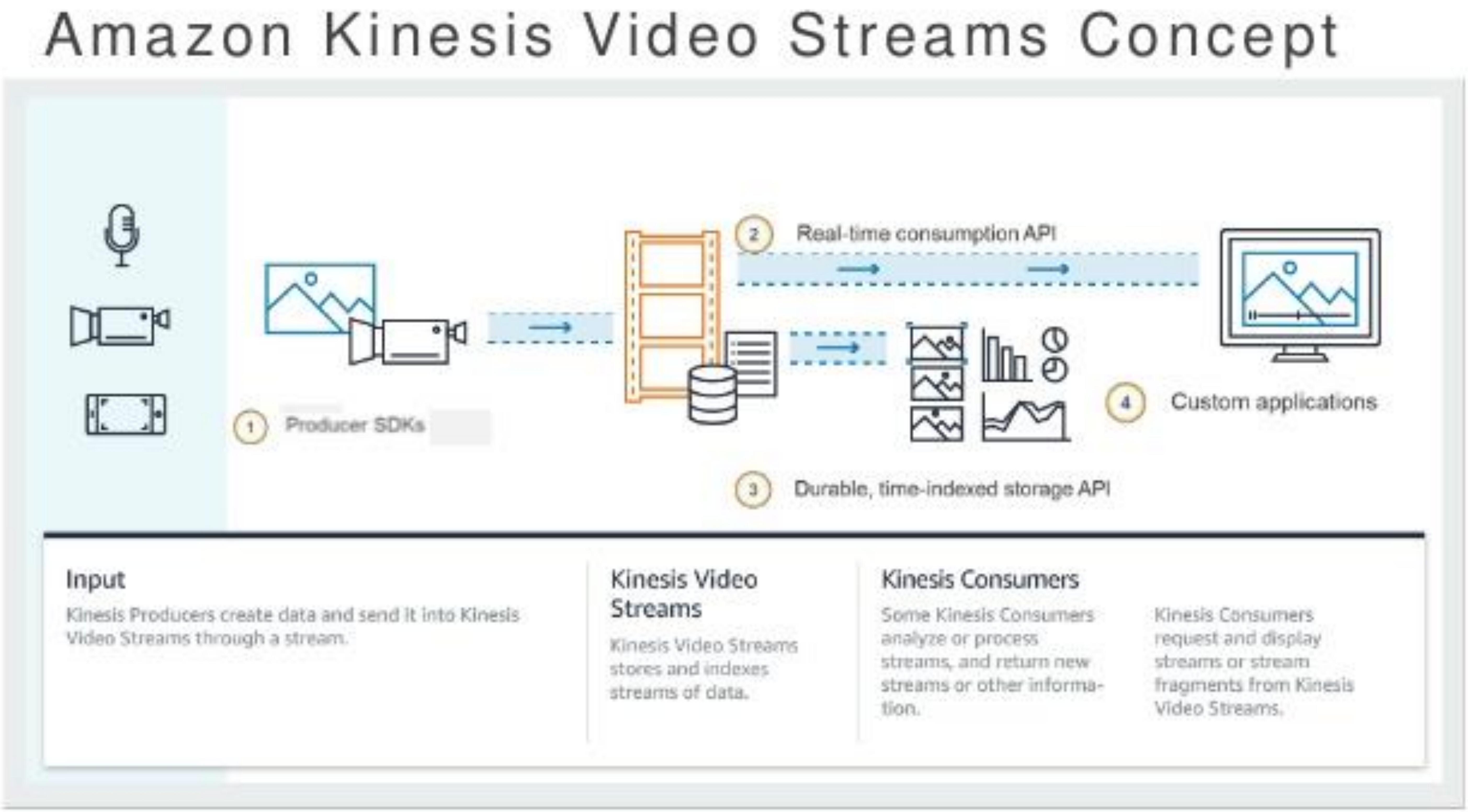
AIY Edge TPU Dev Board



AIY Edge TPU Accelerator

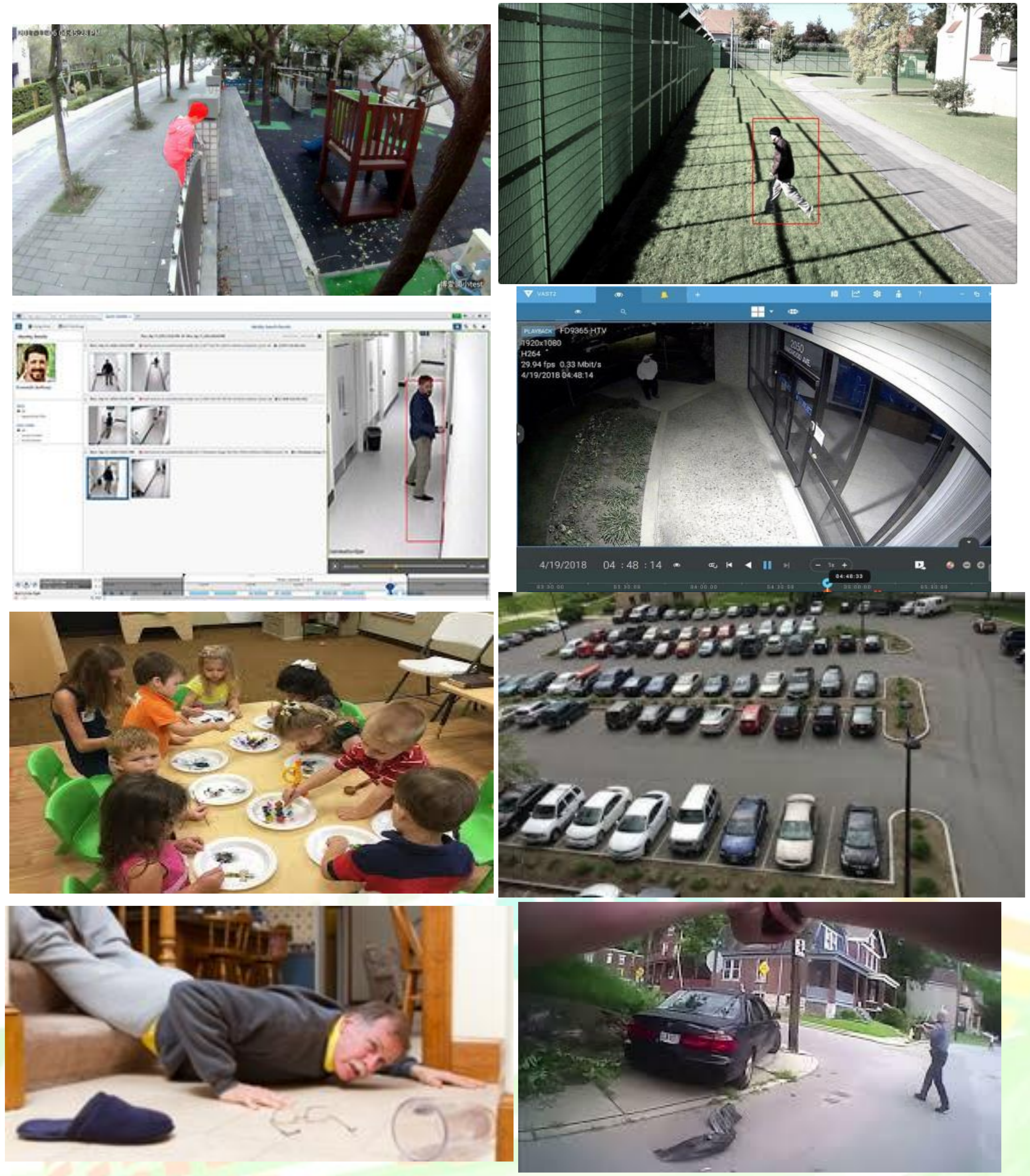
Amazon's DeepLens and Kinesis Cloud





Application 1: Intelligent Video Surveillance

- Intelligent Video Analytic Applications
  - Virtual border control and policing
  - Community/parking space surveillance
  - Building patrol
  - Children/elder monitoring, e.g. fall detection
- Target Receivers
  - 研華科技、國興資訊、凌群、仁寶、大猩猩科技
  - 商湯科技 (sensetime)、盾心科技 (UmboCV)...
  - 新光保全 (SKS)、中興保全 (SECOM)
  - Vivotek (晶睿)、Zoips (零壹科技)、Synology (群暉)
  - 移民署：國境安全管理需求



Application 2 : Smart Hospital

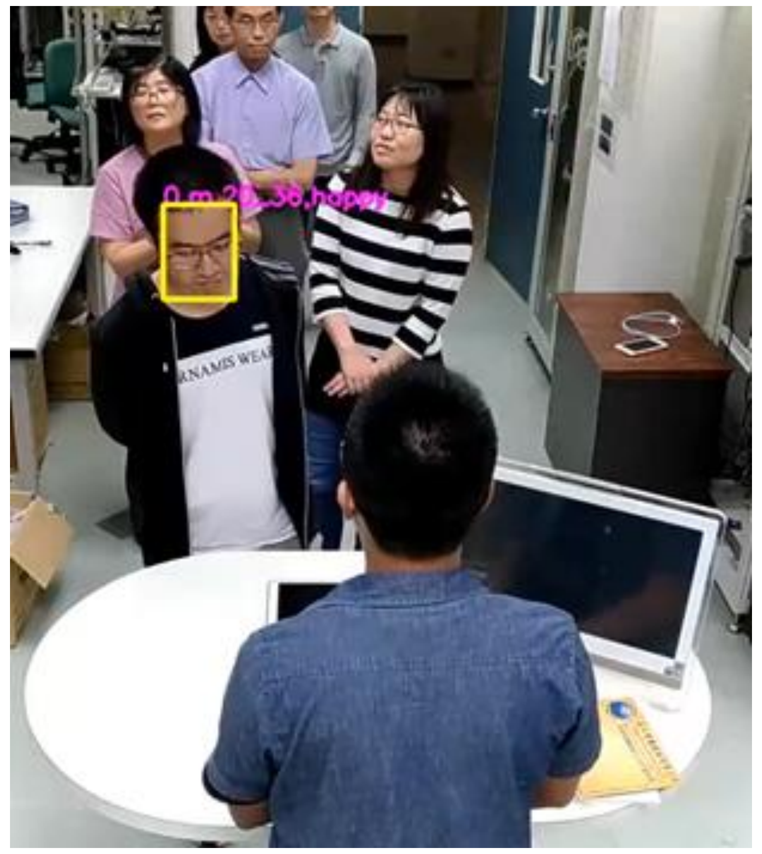
- **Unmanned video-based patient monitoring**
  - **Real time face recognition for patient/personnel tracking, especially with masked faces**
  - **Video-based dementia patient monitoring and care-giver behavior monitoring**
  - **Real time pain and emergency detection for ICU patients**
- **AI-based medical image analysis**
  - **aetherAI develops a platform to enable users to upload their DICOM (digital imaging and communication in medicine) files and analyze them automatically. The detection accuracy of specific types of cancers is more than 90%.**
  - **CS-eHealth for intelligent patient monitoring**

Application 3: Next-Generation Smart Retail

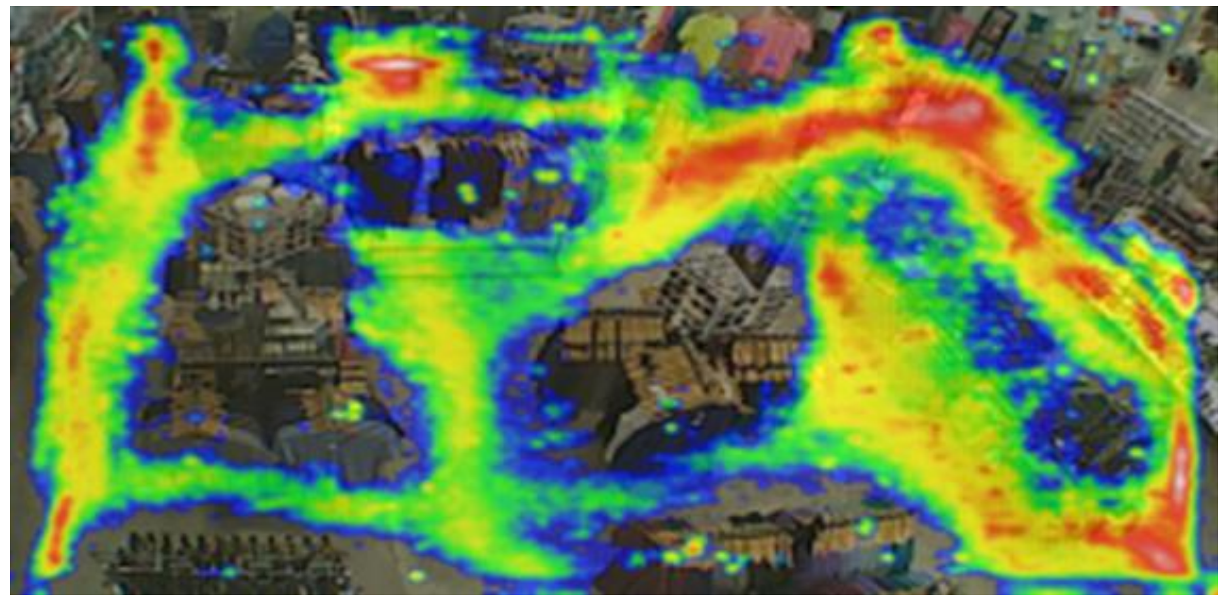
- Benefits:
  - Optimize shoppers’ offline shopping/consuming experiences
  - Capture offline shopper-merchandise interactions so as to combine them with on-line shopping behaviors
  - Unmanned retail store
- Capture whatever can be captured in an on-line store in a physical retail store and integrate them across stores
  - How many shoppers?
  - Each shopper’s trajectory?
  - What merchandise is touched?
  - Does she like/dislike it?

Unmanned Store

Profile/ID Registration

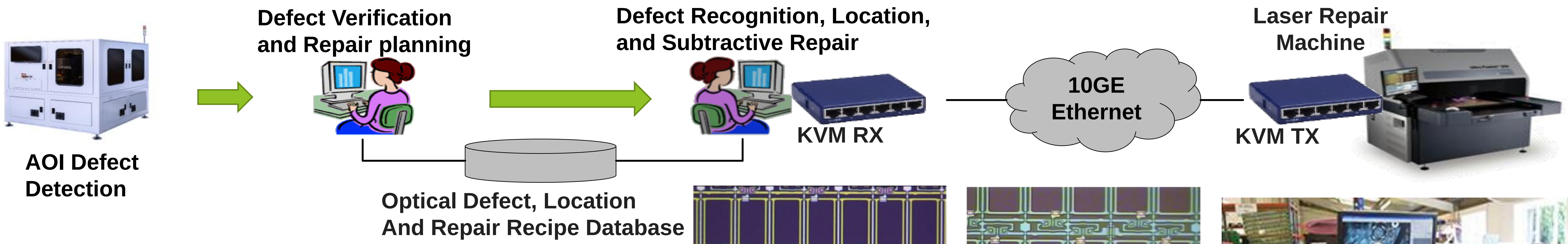


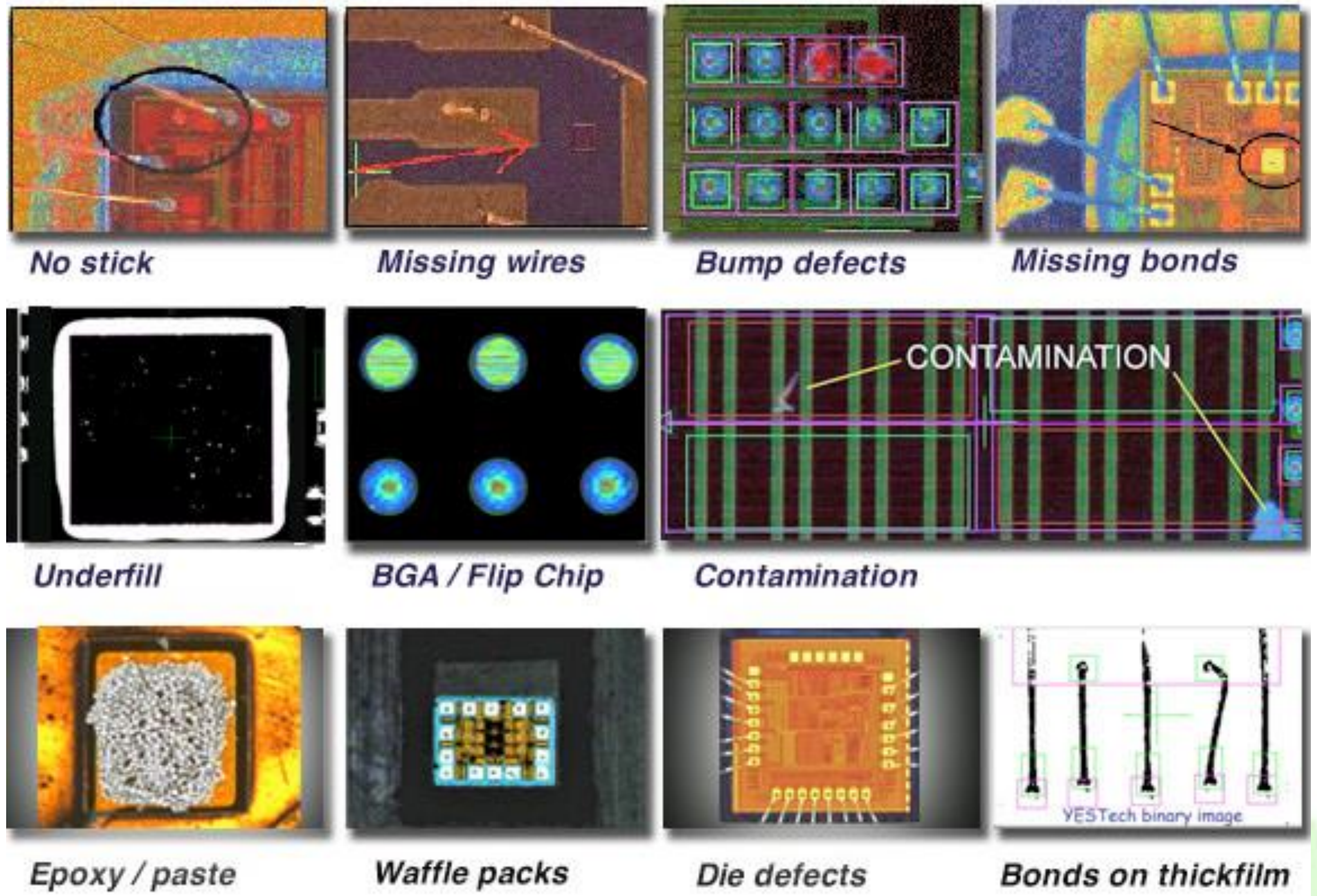
Physical Store Shopping Analytics

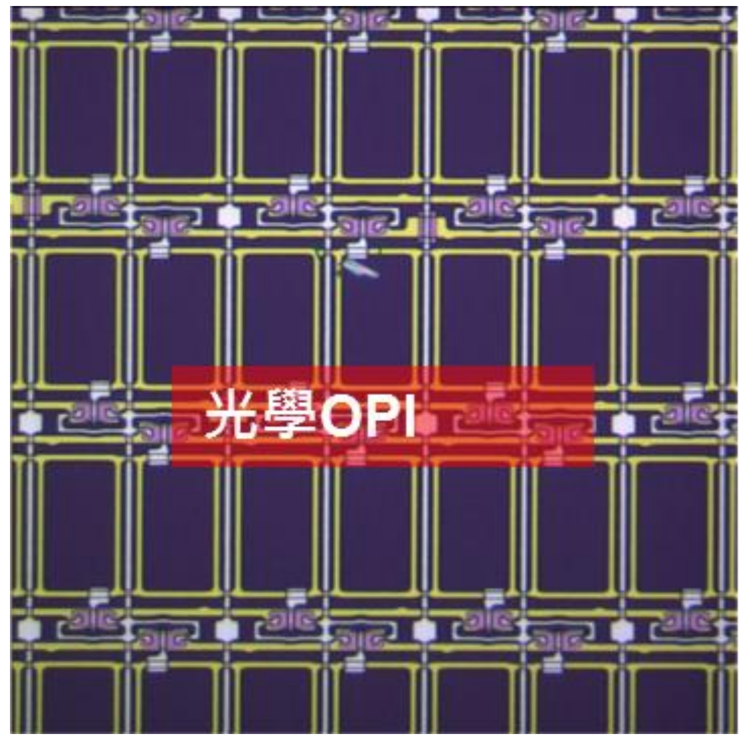


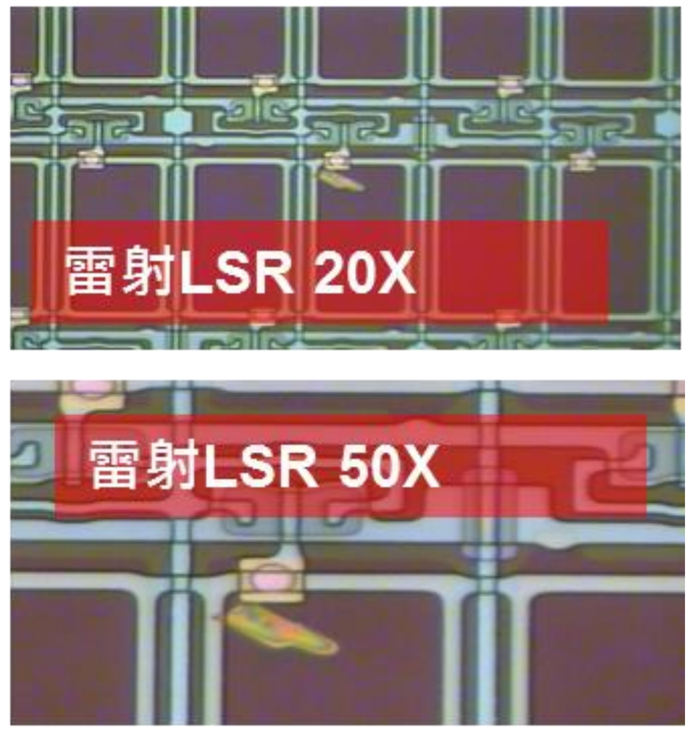
Wave for Service

Application 4 : Smart Manufacturing









Case A	Manual Inspection	AOI Throughput：300000 images/day * 4 human inspectors= **1,200,000 images/day** False negative rate : **5%**
	DNN Inspection	False negative rate: <0.01%, manual inspection load: 5% Throughput：**1.2 M images/day** ➔ **14.4 M images/day**
Case B	Manual Inspection	AOI Throughput：300000 images/day * 10 human inspectors= **3,000,000 images/day** False negative rate : **12.9%**
	DNN Inspection	False negative rate: <1%, manual inspection load: 10% Throughput：**3 M images/day** ➔ **8.6 M images/day**

AI Edge Computing for Video Analysis

- **Target product**: A highly cost-effective edge computing system that leverages 5G low-latency high-bandwidth communication capabilities and caters specifically to real-time AI-based video analytics ➔ Video DNN inference
- Supported general-purpose CPUs
  - Intel/AMD's X86 CPU
- Supported accelerators
  - Nvidia's Tesla P100/V100 and GeForce GTX 1080Ti
  - AMD's Radeon RX Vega 64 and Radeon Instinct MI25
  - FPGA
  - Taiwan's own AI processor (from MTK, RealTek and ITRI)
- Target scale: up to 10 X86 servers and 100 GPUs
- Tight integration with the iMEC software stack developed in 5G systems
- Support for multi-tenancy

Build-up of an AI Edge Computing System

Operating System for Video DNN Model Execution — **How Frequently, When, Which DNN computation**

Deep Learning Framework — **Training Computation to evolve a DNN Model (TensorFlow, Caffe, and PyTorch)**

DNN Model Compiler — **DNN Model ➔ Executable Code (TVM)**

Baseline Operating System — **Linux container (LXC) and orchestration (Docker)**

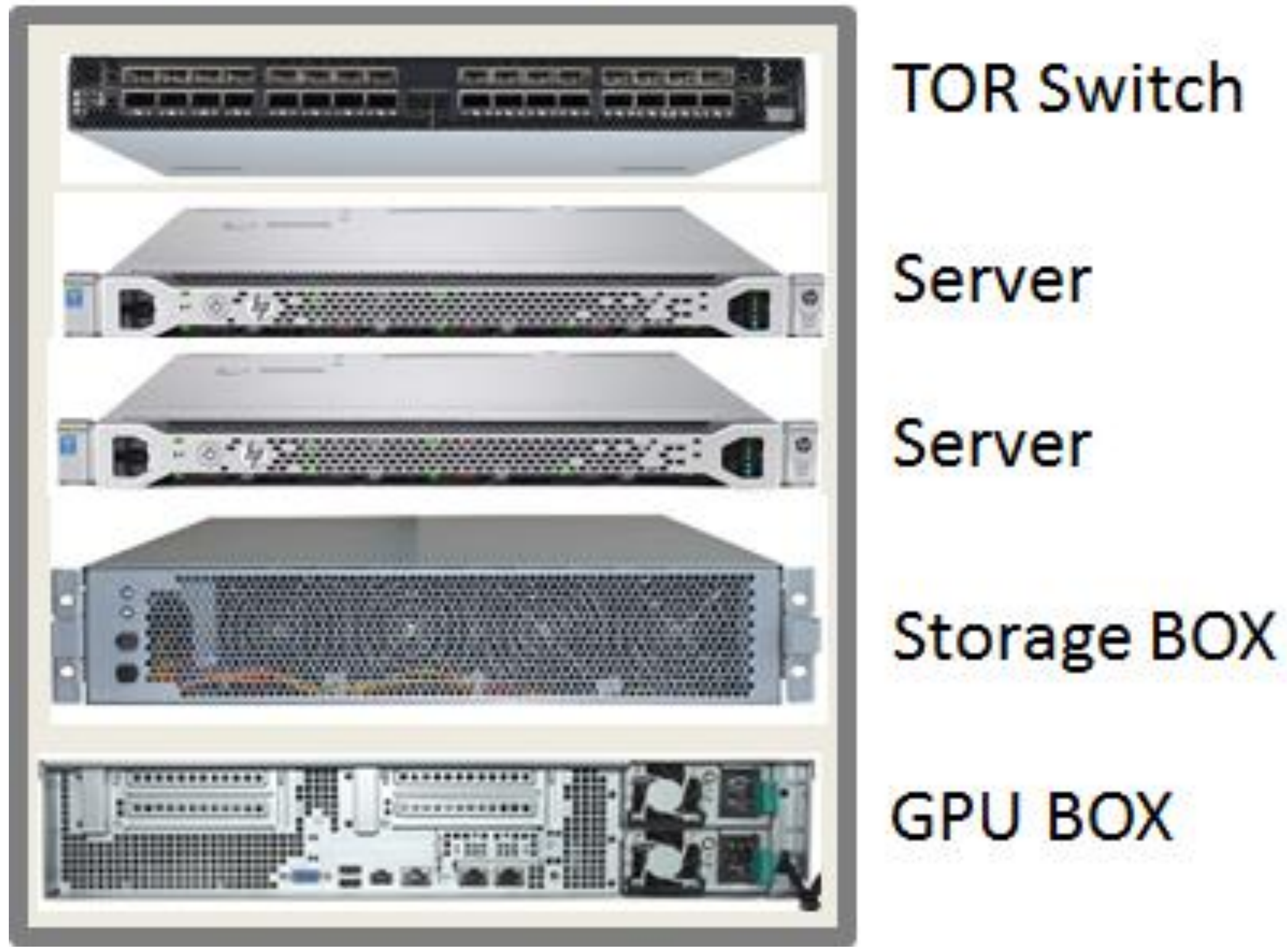
Nvidia/AMD/Intel GPU, FPGA

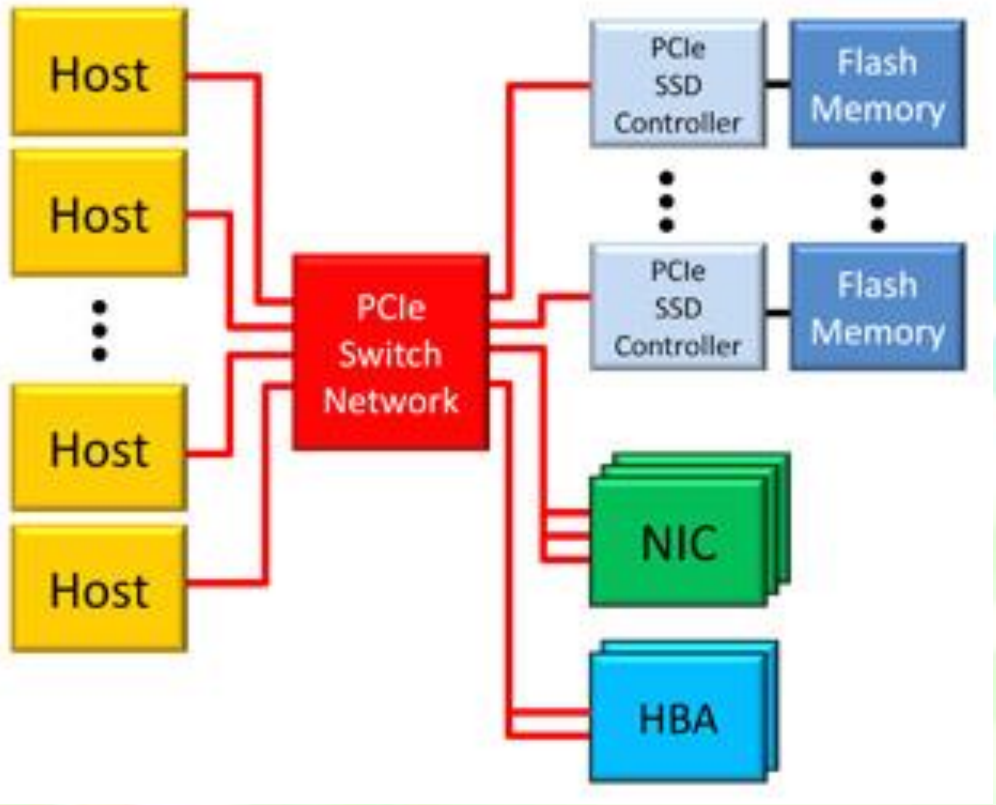
X86/ARM server with multiple PCIe slots and effective cooling

PCIe-based Disaggregated Rack Architecture

- PCIe as an I/O interconnect and inter-server communication backbone
- Resources within a rack form a global pool that could be dynamically assigned to individual servers
  - GPU, NVMe SSD, SAS SSD, and 100GE NIC
  - Direct access from a server's CPU to its assigned HW resources
    - Zero protocol SW/FW overhead
  - Software-defined server that is tailored to CPU-intensive, GPU-intensive, memory-intensive, and network-intensive workloads
- High-bandwidth low-latency intra-rack or east-west communications over PCIe
- Inter-rack or north-wouth communication traffic still goes through Ethernet
- Examples: Adlink's NexTCA and Liqid

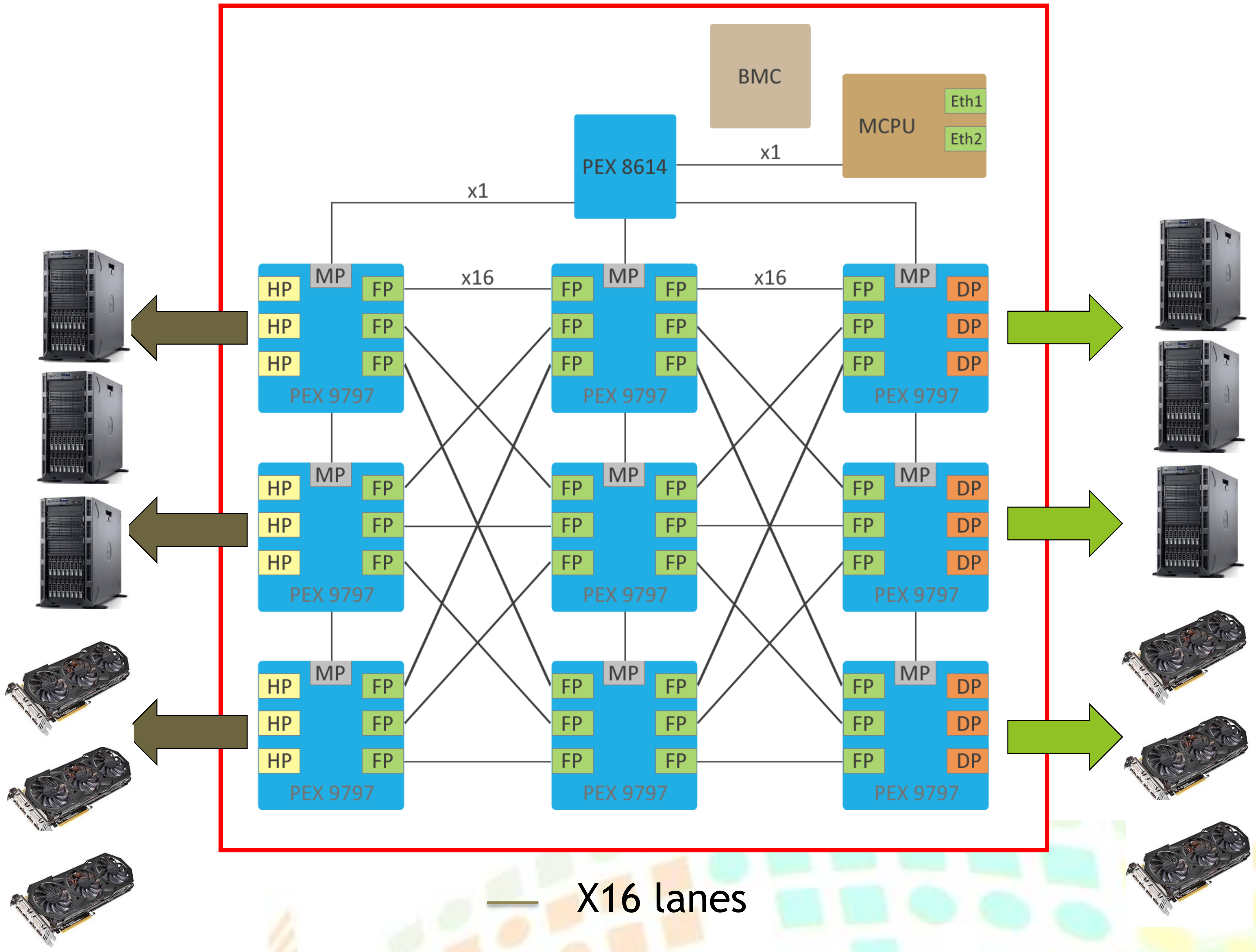
Edge Computing Rack





Meshed PCIe Network

- Main motivation: Support for scalable inter-CPU, inter-GPU and CPU-GPU communication
- 3x3 PEX88098s used to form a mesh-like topology
- 18 end points each with a 16-lane link or 36 end points each with a 8-lane link to the switch
- Gen4-based: 16Gbps per lane
  16 lanes = 256 Gbps



DNN Model Compiler

- **Objective**: Compile a DNN model's training/inference computation into efficient code that runs on a wide variety of AI acceleration processors
- **Compiler toolchain**: simulator, accelerator driver, DNN-specific library, code generator, and optimizer
- **Consolidation of Multiple DNN Models:** Given M DNN models that operate on a video stream, how to apply feature extraction on each video frame exactly once?
  1. Apply the M models on each video frame a breadth-first rather than depth-first manner
     – Apply all the models on an input image before it is evicted out of the CPU/GPU cache
  2. Retrain the M models so that they share the same feature extraction layers
     – Cross-model optimization by consolidating their feature extraction layers into a common representation

Resource Management for DNN Model Execution

- Target applications of DNN-based video computing
  - Perception subsystem for ADAS and autonomous driving
  - Face recognition for patient/personnel tracking in smart hospital
  - Limb motion and body language recognition for smart retail
  - Automated optical inspection (AOI) for real-time manufacturing process
- **Objective**: Maximize the resource utilization efficiency of an AI edge computing system customized for scalable real-time video object detection, location, recognition, tracking, and analysis
- Key ideas
  - Selective: There is no need to apply FULL video analysis to EVERY video frame.
  - Incremental: Take advantage of N-1th frame's result when processing Nth frame
  - DNN model-aware: Exploit a DNN model's structure for data prefetching and reuse

OS for Video DNN Model Execution

- Building blocks
  - DNN models for video object location, recognition, tracking, and feature analysis
  - Image frames in a video sequence
- Proposed approach
  - **Baseline**: Apply **M** DNN models to each frame in a video sequence
  - **Optimized**: Apply a subset of the possibly reduced versions of the **M** DNN models to each frame in a video sequence while achieving the same video analysis accuracy as Baseline

Model 1: 1.1, 1.2, 1.3
Model 2: 2.1, 2.2, 2.3
Model 3: 3.1, 3.2, 3.3

1.1	1.1	1.1	1.1	1.1	1.1	1.1
2.2	2.1	2.2	2.1	2.1	2.1	2.2
3.3	3.1	3.1	3.1	3.2	3.1	3.3

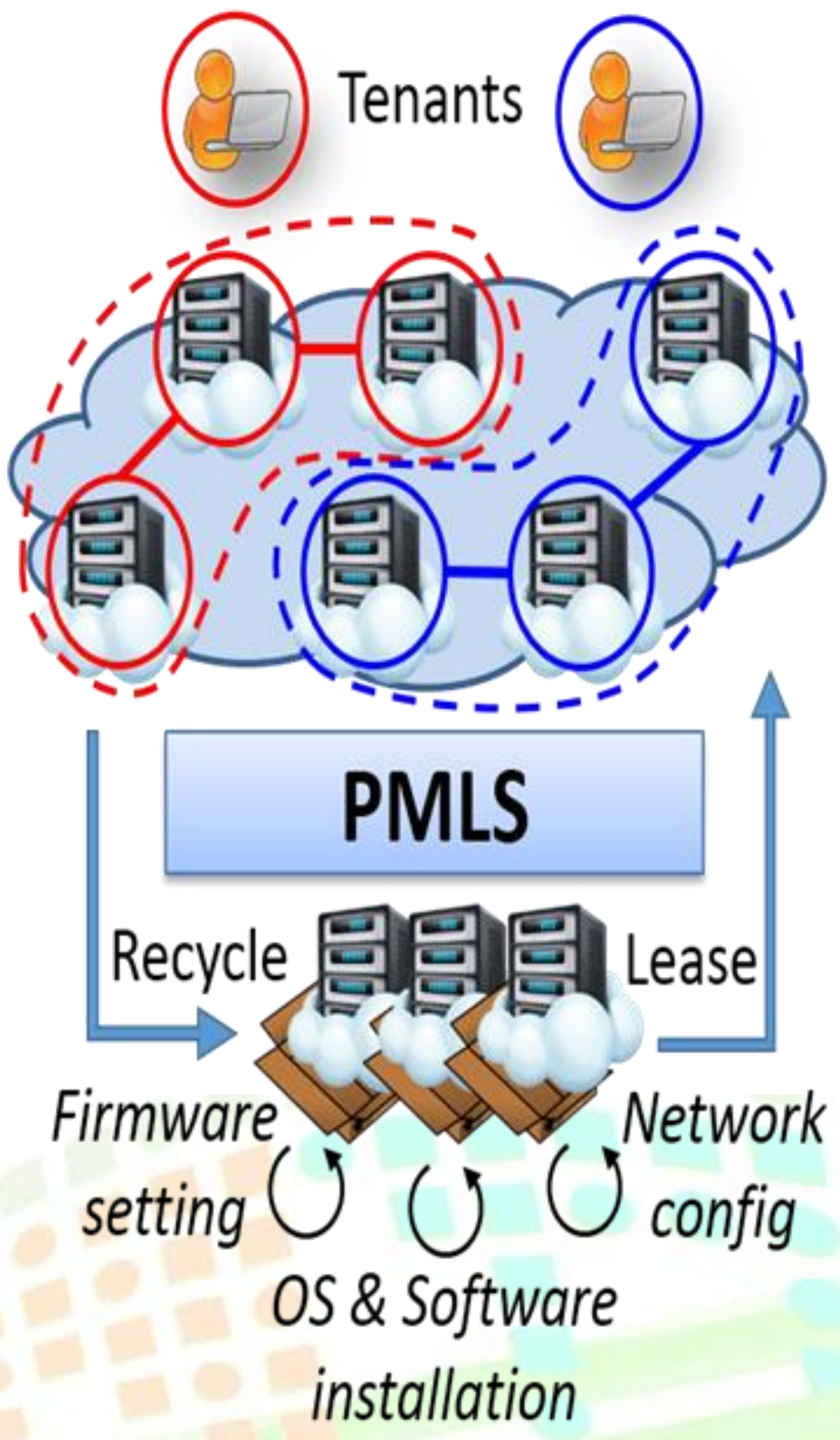
Multi-Tenancy Edge Computing

- Each tenant wants to rent a couple of edge computing servers from a geographically distributed set of edge computing data centers
  - Each tenant gets a physical data center instance (PDCI), which consists of a set of physical servers, a physical network connecting them, and a set of local/remote storage volumes accessible to the servers.
- Why Hardware as a Service (HaaS)?
  - Preferred hypervisor
  - Big data/DNN training/HPC: efficient utilization of HW resources is critical
  - Container-based virtualization is sufficient.
- Comparison among service models:

Model	Rental Unit	IT HW Ownership	HW Management
IaaS	Virtual machine	Service provider	Service provider
HaaS	Physical machine	Service provider	User
Colocation	Rack space	User	User

HaaS Service Model

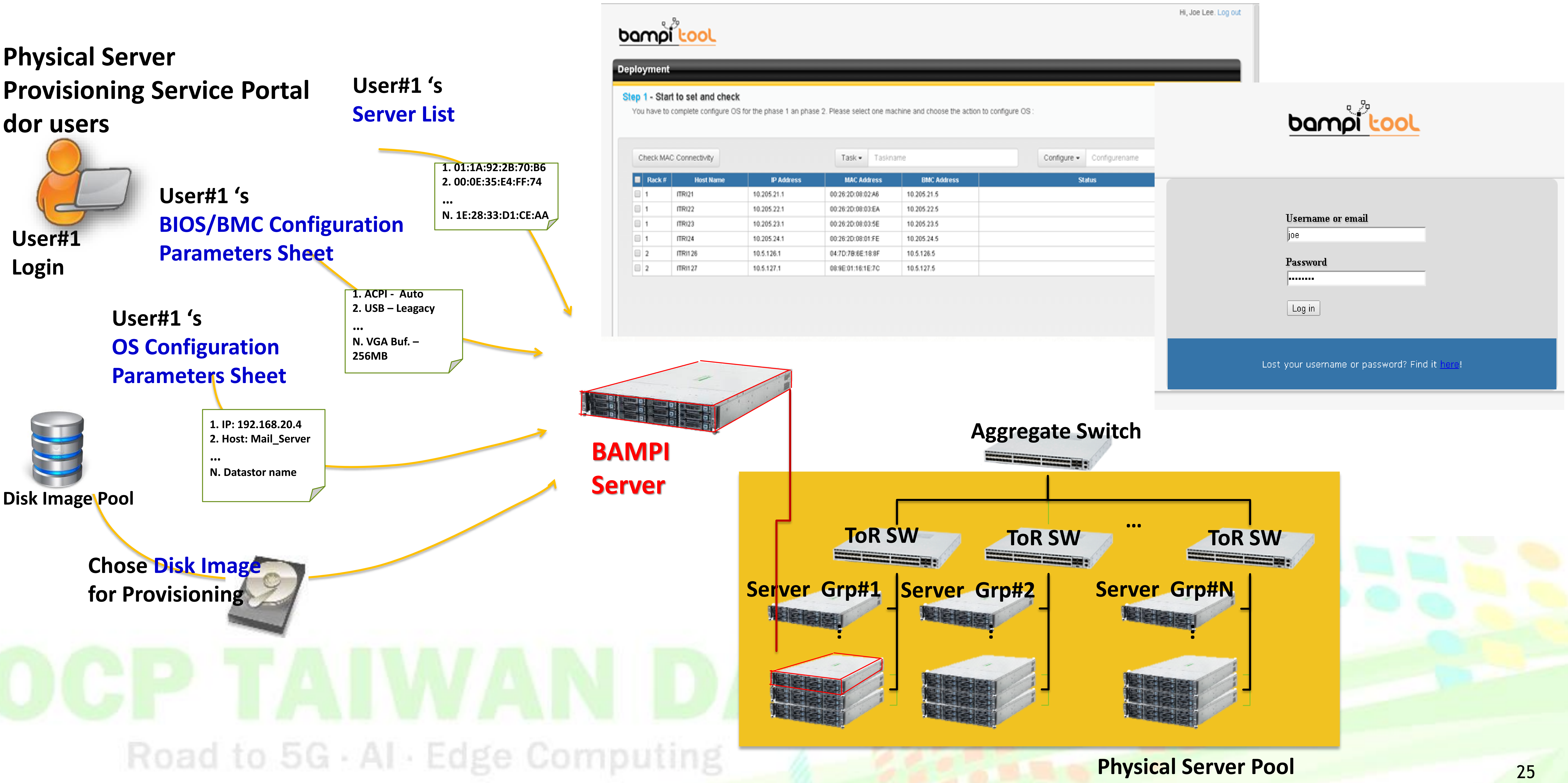
- An HaaS reservation consists of the following:
  - **A set of servers**, each with its hardware specification and configurations on BIOS, BMC, PCI devices, and OS
  - **A set of storage volumes** that exist in local or shared storage, and are attached to the servers
  - **A set of IP subnets** that connect the servers and how they are connected
  - **A set of public IP addresses** to be bound to some of the servers, and their **firewall/NAT policies**
- User could remotely configure and install OS and applications on servers in its PDCI, and manage their operations at run time.
- Players: **HaaS operator** and **HaaS tenant**



Building Blocks for ITRI HaaS

- **Bare-metal provisioning (HaaS operator)**
  - Hardware asset discovery, inventory and configuration
- **Service request (HaaS tenant): BAMPI**
  - Server provisioning: Server hardware/firmware configuration and verification
  - Storage provisioning: local storage vs. shared storage
  - Network provisioning:
    - Agentless and scalable multi-tenancy network isolation: One HaaS tenant's virtual network is isolated from other HaaS tenants
      - Support up to hundreds of thousands of IP subnets across all tenants
    - Switch hardware/firmware configuration and check
- **Run-time administration (HaaS operator and tenant)**
  - Hardware usage and maintenance tracking: HaaS provider
  - System monitoring and administration: HaaS provider and HaaS tenant

HaaS Operator’s View of BAMPI



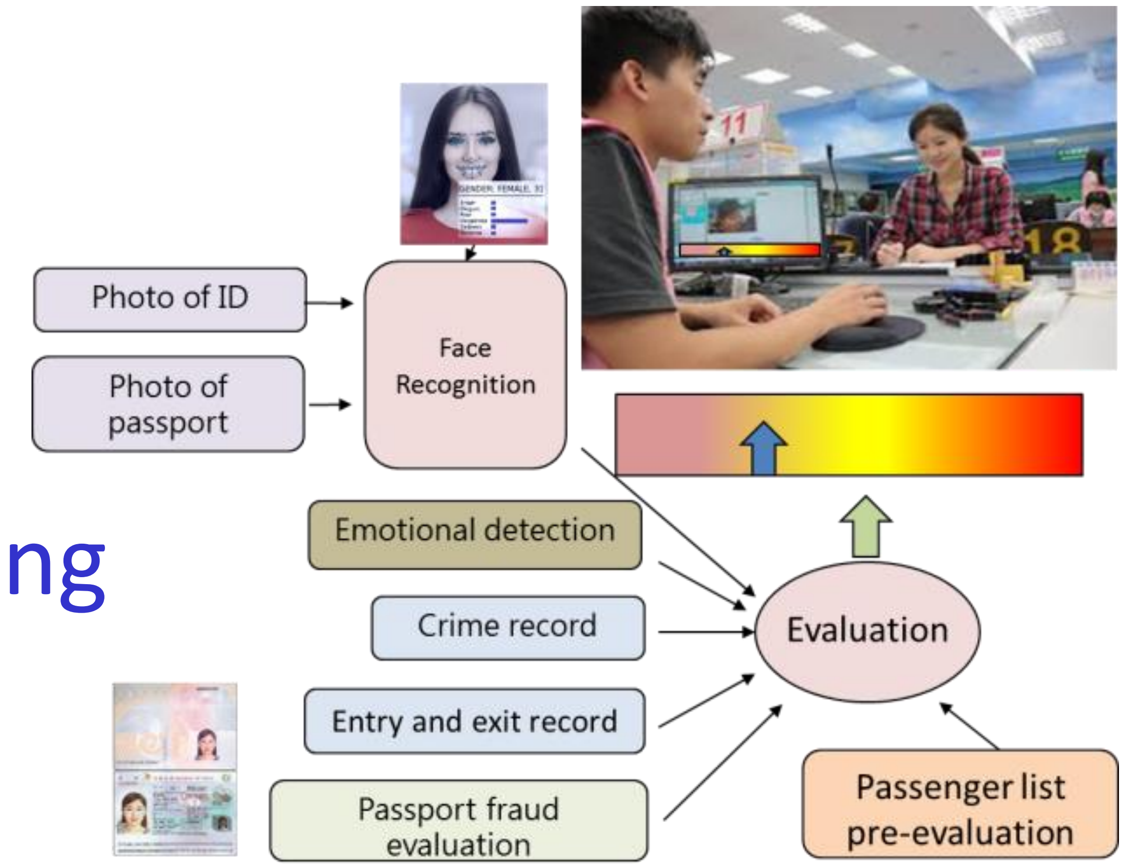
Effectiveness of Bare Metal Server Provisioning

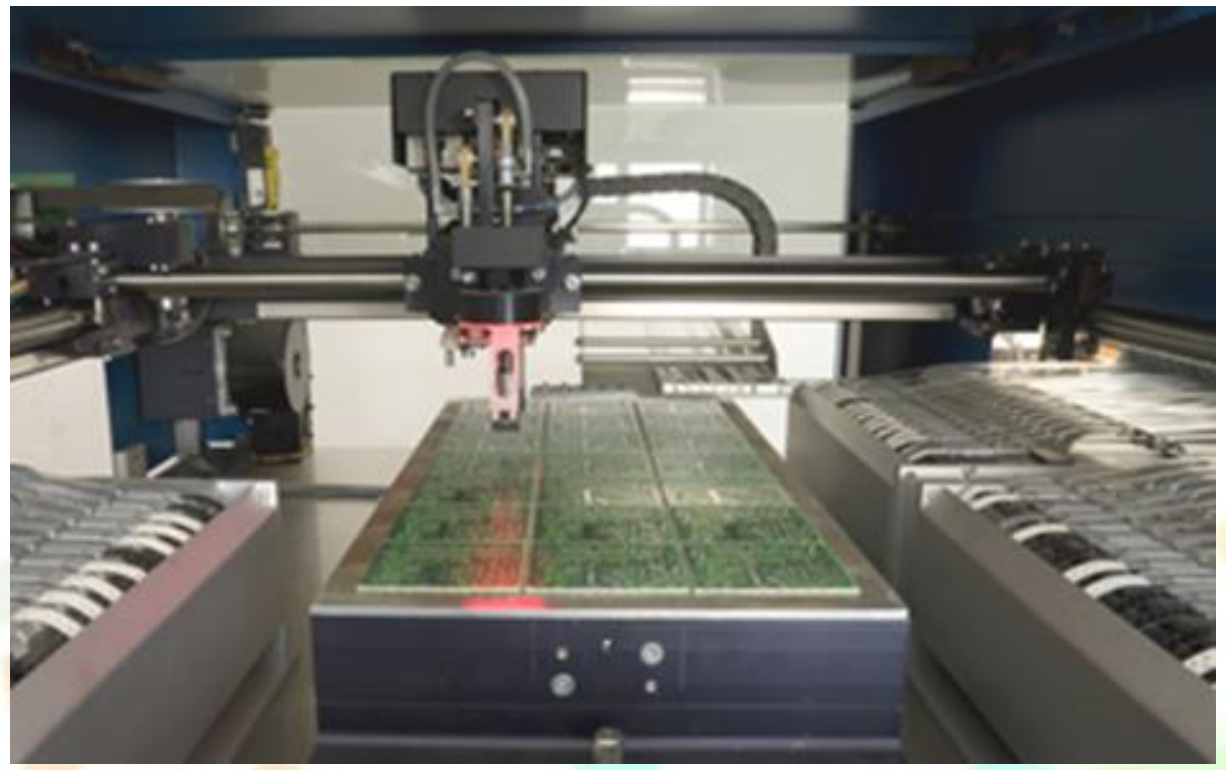
KDDI experience: About 200 times faster than manual operation

	Manually Done	BAMPI
Initialize BMC Network	※ Time of Completion for 80 servers: 288 man-hours	※ Time of Completion for 80 servers: 1.5 man-hours
Find the MAC Address of Server		
Upgrade BIOS / BMC / RAID Firmware		
Configure BIOS / BMC / RAID / OS		
Check BIOS / BMC / RAID / OS		
Restore OS		
Check Network Connectivity		
Configure BMC Network		
Delete Kitting VMkernel		

Summary

- **Target products**
  - AI edge computing system
  - DNN training appliance: TASA (Taiwan AI Systems Alliance)
  - Application-specific DNN inference models: hospital, retail, manufacturing
- **Application solutions**
  - Video-based 3rd-gen airport border control system
  - ETC based on real-time vehicle license plate recognition engine
  - Perception subsystem for ADAS and autonomous driving
  - Scalable automated optical inspection system
- **Technical building blocks**
  - Flexible allocation and sharing of GPU/FPGAs
  - Efficient inter-GPU communication
  - DNN model-oriented compiler and OS
  - Multi-tenancy hardware as a service (HaaS)





Thank You!

Questions and Comments?

chilung@itri.org.tw