2020 embedded vision summit

Edge Inferencing-Scaling with Intel® Vision Accelerator Design Cards

Rama Karamsetty
Edge.AI-IOTG
Intel Corporation
September 2020

intel

Notice and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

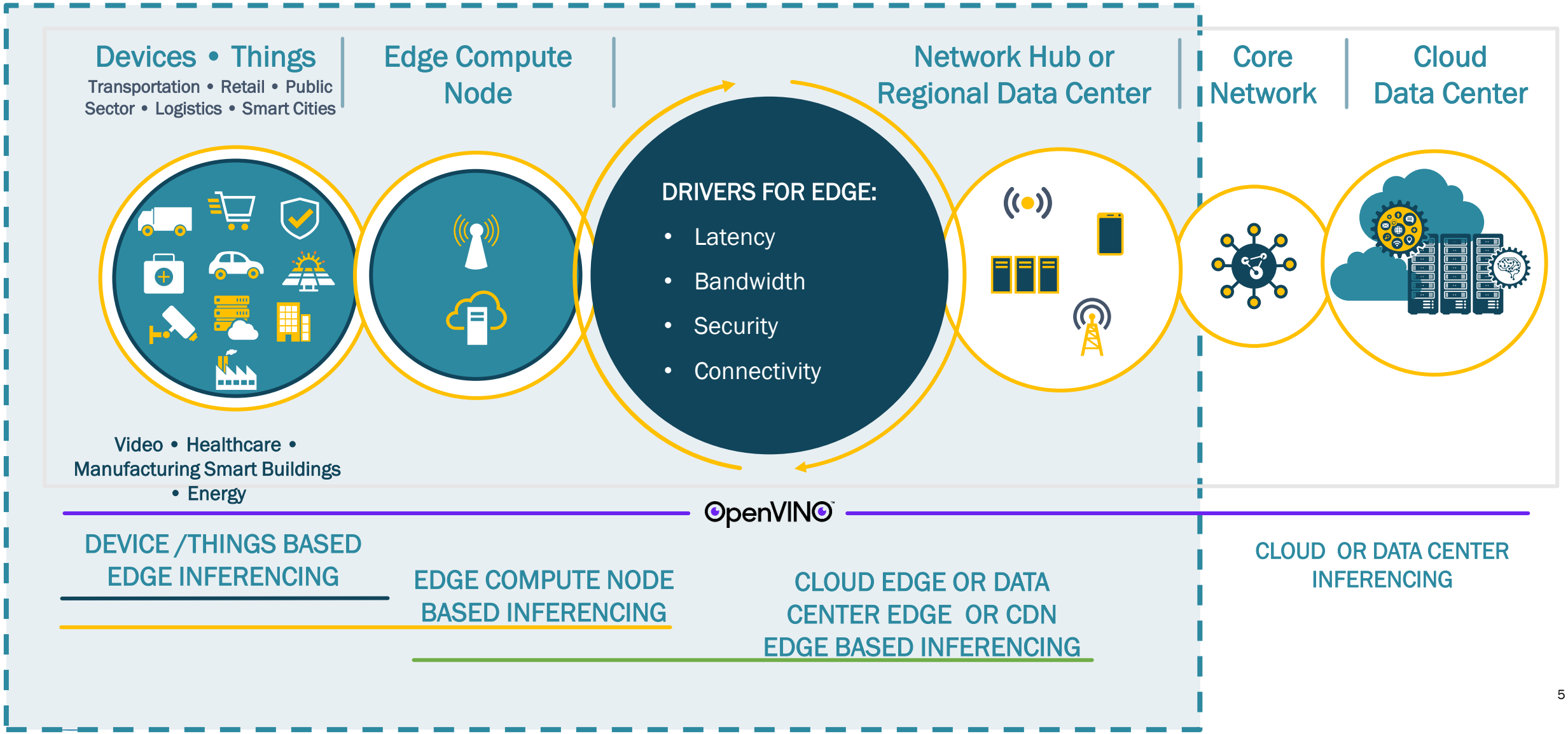
This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Agenda

- Intel's Edge AI journey
- Applications & Use cases
- Intel® Vision Accelerator Design Cards
- Intel Ecosystem
- Partner Journey, Challenges, Solutions
- Conclusion

One Size fits all- A Myth

Reach of Accelerator Cards – Does one size Fit all?



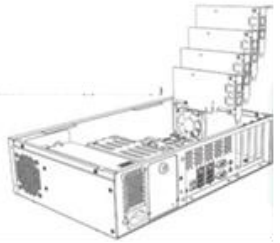
Industry/Applications for Edge Inference Cards

Across Many Industries…

And Many Devices/Applications…

Edge devices
(NVR, Robotics, AI Boxes…)

Edge servers
Video, VA/AI servers, CDN



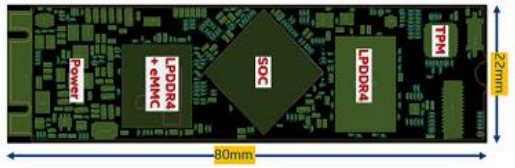
Common DL Imaging/Vision Use Cases in Edge Segments

- INTELLIGENT TRAFFIC MONITORING
- ANOMALY DETECTION
- ANONYMOUS ANALYTICS
- LICENSE PLATE RECOGNITION
- PERIMETER PROTECTION
- OBJECT DETECTION , TRACKING
- REAL-TIME ANALYTICS
- PEOPLE MOVEMENT
- THERMAL MAPPING

Intel® Vision Accelerator Design Cards

Intel® Vision Accelerator Design cards–Description

- IVAD cards-Specialized cards designed with one or more Vision Processing Units(VPUs) to deliver high-performance machine vision at ultra-low power.
- Small Form Factors-PCIe, M.2, mPCIe connectivity based pre-validated edge inference engines for size constrained systems
- Plug into any existing Intel Architecture based Host ecosystem solution
- Help offload encode, detection, recognition on to the accelerator card



Intel® Vision Accelerator Design cards value

Flexible and Scalable options
- Low, Med and High-performance options
- Scalable price points

Relative lower power consumption

Long Life Support w/ 24/7 usage

OpenVINO™ supported
- Application portability and forward/backward compatibility

Lower System level TCO

Glimpse of Ecosystem products

Example card based on Vision Accelerator Designs	 x1 VPU	 x2 VPUs	 x4 VPUs	 x8 VPUs
Interface	mPCIe, M.2	mPCIe**, M.2	PCIe x4, PCIe x2, M.2	PCIe x4
Currently manufactured by*	AAEON an ASUS assoc. co. ; ADLINK Leading EDGE COMPUTING ; ADVANTECH Enabling an Intelligent Planet ; Hewlett Packard Enterprise ; iEi ; NEXCOM ; tinyGO ; uzel			
Software tools	**Intel® Distribution of OpenVINO™ Toolkit** Develop NN Model; Deploy across Intel® CPU, GPU, VPU, FPGA; Leverage common algorithms			

**Other names and brands may be claimed as the property of others

Application Case Studies

Story of an ISV - AXXONSOFT

50/100/150 CAMERA STREAMS W/ DELL POWEREDGE SYSTEM

OBJECT TRACKING

PEOPLE MOVEMENT

Intel® Components

- Intel® Xeon™ Processor
- x2, x3, x4 'Gen 1 Intel® Movidius™ Myriad™ X VPU Based x8 IVAD'
- Intel® Distribution of OpenVINO™ toolkit

Story of an ISV -- AXXONSOFT

COMPONENT	SETTINGS
Input video stream	640x360 @ 25 fps
Number of input channels (potential camera feeds)	50 100 150
Neural NW processing framerate	5 fps
Number of active neural NW channels	50 100 150
Video archive	Raid 5, 65 TB
Number of archiving streams	50 100 150

- Key challenge was to apply neural network detection/classification to 50, 100 and 150 camera streams' images with the *same scalable system design*, and no rework in partner application and still be within overall power envelope
- 640x360 pixel color images captured @ 25 fps and encoded in H.264 format from 50 cameras, 100 cameras and 150 cameras
- Realtime inferencing need: Application 'tags and boxes' areas of interest and displays meta-data onscreen for live viewing, for remote viewing and for local storage

Story of an ISV --AXXONSOFT

NUMBER OF VIDEO CHANNELS AND PROCESSING FRAME RATE (5 fps inference per channel)	FRAME RATE (FPS) (PROCESSED)			TEMPERATURE HDDL NODES (DEGREES C)		
	Min	Max	Avg	Min	Max	Avg
150 channels	748	750	749	59	74	67
100 channels	499	499	499	55	71	63
50 channels	248	250	249	53	68	61

Story of an OEM

Shopper Engagement occupancy, shelf inventory

Covid-19 occupancy, social distancing solution

Curbside Wait time

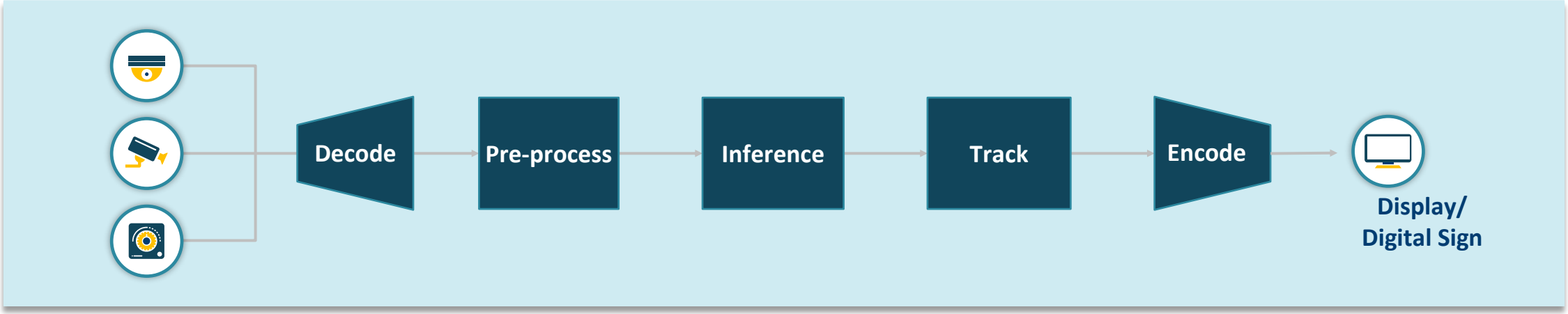
Intel® Components

- Intel® Atom™, Core ™, Xeon ™ Processors
- x1 M.2, x2 PCIe ‘Gen 1 Intel® Movidius™ Myriad™ X VPU Based IVAD Card’
- Intel® Distribution of OpenVINO™ toolkit

Sensormatic by Johnson Controls

Story of an OEM

- Technical challenges
  - Scalable compute w/o software rework to move up/down performance stack.
  - Heterogenous compute needed with different workloads for normal operations vs inferencing pipeline.
  - Decode, Tracking and Encode are better done on CPU.
  - Pre-processing and Inferencing are better handled by VPUs.
- Intel IVADs allow for workloads to be split amongst CPU, VPU compute resources without any application level re-programming.
  - Application level re-programming for workload optimization between compute engines is not scalable.
  - IVADs w/ IA hosts allow for scalability of different use cases just by adding more IVAD cards with no application level reprogramming .



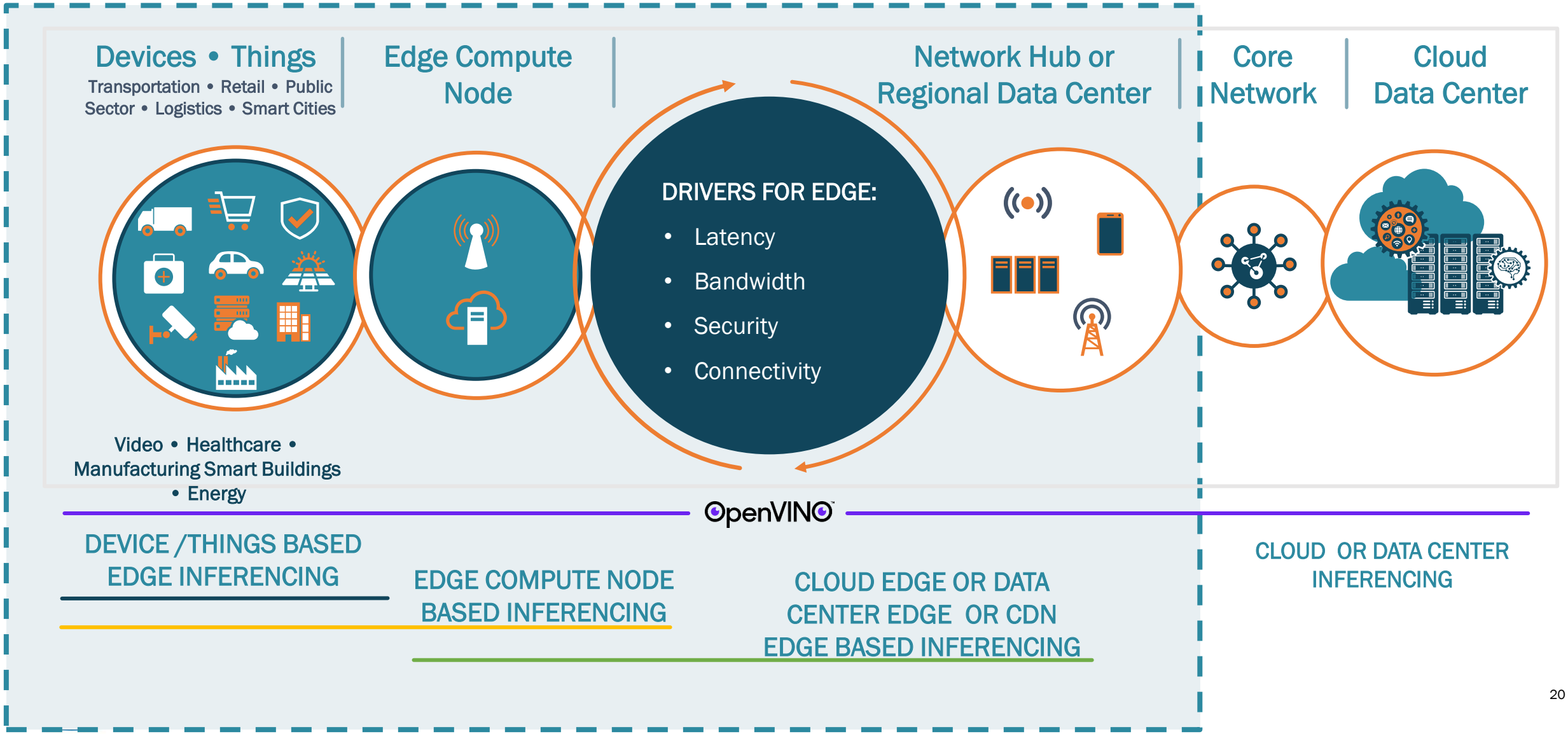
Story of an OEM – Partner Quote

"The retail industry continues to evolve, and our collaboration with Intel will help us take on the industry's biggest challenges," **said Subramanian Kunchithapatham, vice president, Engineering, Sensormatic Solutions.** ***"The collaboration will allow us to deliver smart, <u>connected and scalable</u> solutions that allow retailers to gain real-time insights into inventory, shoppers, associates and the retail environment throughout the entire customer journey."***

Source: https://www.bloomberg.com/press-releases/2020-01-13/sensormatic-solutions-and-intel-announce-technology-collaboration

Conclusion

Reach of Accelerator Cards – Does one size Fit all?



Contact Information

Additional information

Intel® Vision Accelerator Design Cards

Additional information:
Email Rama Karamsetty

Intel Workshops and Demos to Visit

General Session Speaker:

- Bill Pearson, VP IOTG, GM Developer Enabling, Tuesday, September 15, 10:00 a.m. to 10:30 a.m. PDT: **Streamline, Simplify and Solve for the Edge of the Future**

In-depth technical workshops

- Friday, September 18, 9:00 a.m. to 1:30 p.m. PDT: **Using the Intel® Distribution of the OpenVINO™ Toolkit for Deploying Accelerated Deep Learning Applications**
- Monday, September 21, 9:00 a.m. to 1:30 p.m. PDT: **Intel's Edge AI for Retail**
- Wednesday, September 23, 9:00 a.m. to 1:30 p.m. PDT: **Intel's Edge AI for Industrial**

Technical presentations

- Alexander Kozlov, Deep Learning R&D engineer, Intel: **Recent Advances in Post-Training Quantization**
- Dr. Manas Pathak, Global AI lead for oil and gas, Intel: **Acceleration of Deep Learning for 3D Seismic**
- Tara K. Thimmanaik, solutions architect, Intel: **Smarter Manufacturing Achieved with Intel's Deep Learning-Based Machine Vision**
- Gary Brown, Director of AI Marketing, Intel: **Getting Efficient DL Inference Performance: Is It Really All About The TOPS?**
- Rama Karamsetty, Edge AI Marketing Manager, Intel: **Edge Inferencing-- Scaling w/ Vision Accelerator Cards**
- Vaidyanathan Krishnamoorthy, edge inference solutions architect, Intel: **Federated Edge Inferencing**

Dedicated demos and networking space