Innovating at the Edge Meetups

Opportunities & Challenges for AI-based Services at the Edge

- Sunku Ranganath, Global Solutions Architect, Intel
- Hassnaa Moustafa, Principal Engineer, Intel

Notices and Disclaimers

- Intel provides these materials as-is, with no express or implied warranties.
- All products, dates, and figures specified are preliminary, based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- 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 product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://intel.com.
- Some results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.
- Intel and the Intel logo are trademarks of Intel Corporation in the United States and other countries.
- Other names and brands may be claimed as the property of others.
- © Intel Corporation

Opportunity at the Edge

By 2024,
the edge silicon
opportunity
will reach
$65 billion.[1]

By 2025,
the edge computing
(multi-access edge compute
+ private wireless)
opportunity across hardware,
software, and services
will reach
$29 billion.[2]

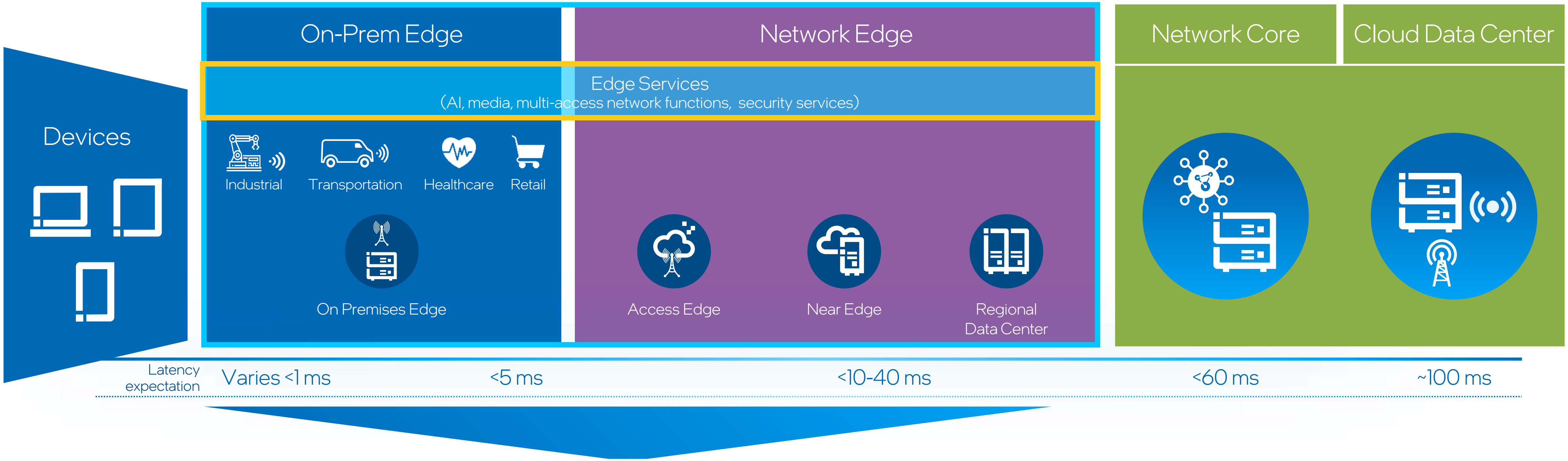
By 2025,
75 percent
of data will be
created outside
of central
data centers.[3]

[1] Intel Fuels the Edge Today With Expanded Tech, Customer Deployments, Businesswire, Sept 23, 2020.

[2] MEC definition here refers to MEC2.0 hyperconverged edge. Source: IDC, Omdia, Intel Judgment.

[3] What Edge Computing Means for Infrastructure and Operations Leaders, Gartner, Oct 3, 2018.

Intelligent, Secure, & Multi-Access Services
Cloud Native Edge Solutions



Key challenges to overcome

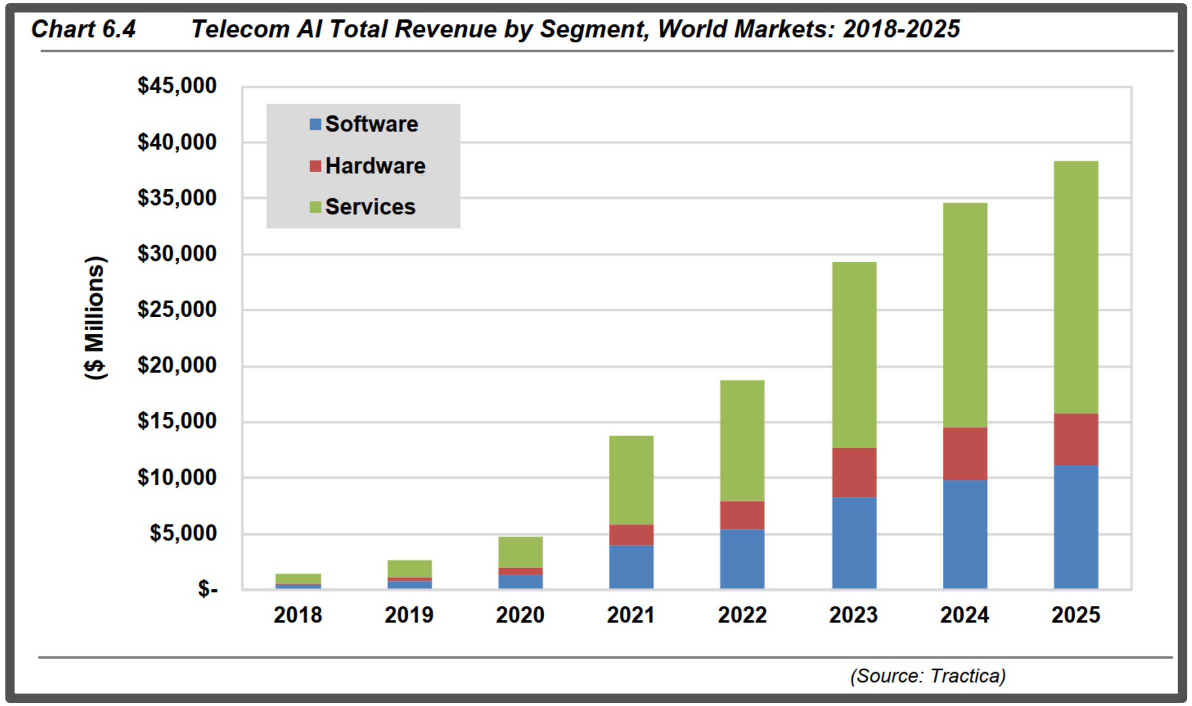
Edge SW solutions consistency and scalability cross diverse edge platforms and location requirements

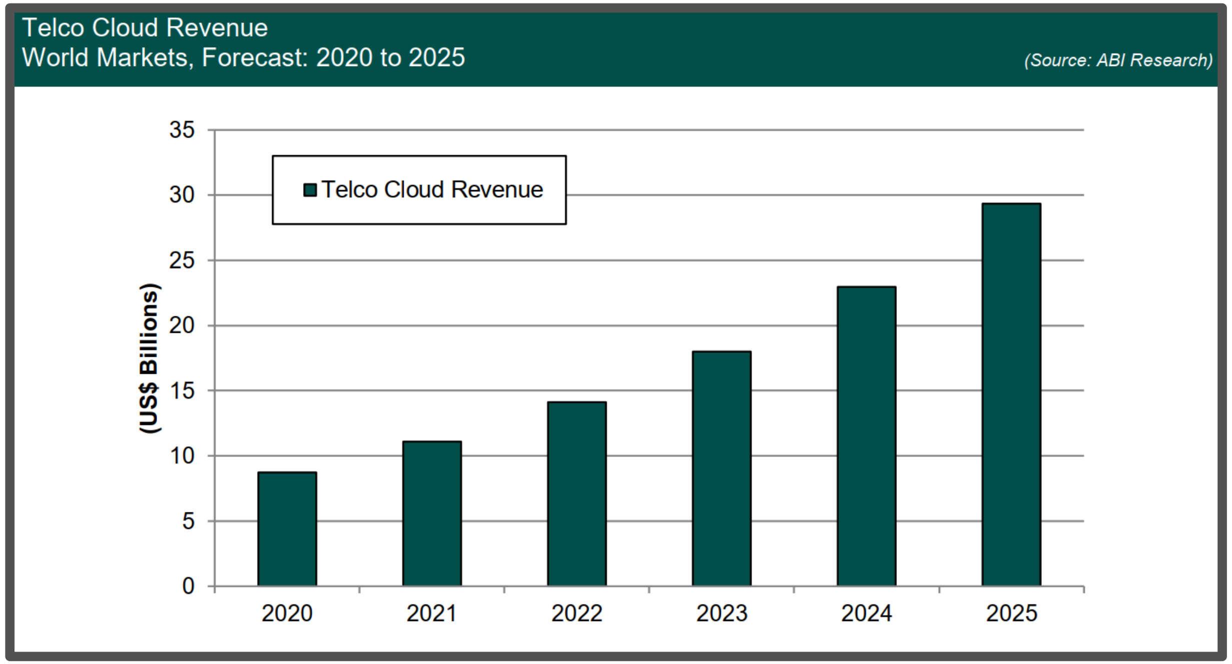
SW Defined Multi-access Network Functions as network services across any edge location and w/o vendor lock in

Optimize edge solutions for converged services to meet stringent KPIs for each category of service

Broad building blocks (HW & SW) from Intel enabling optimized Computer Vision Services (with AI & Media) to build Cloud Native edge solutions

Your opportunity!





Here and now

HW Capabilities

Enabling Edge Solutions with AI, Security & Multi-Access

AI Workload Acceleration

- Intel® Deep Learning Boost VNNI accelerating AI inference
- Intel® Advanced Matrix Extension (Intel® AMX) accelerating AI and ML
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) FP16 advancing AI workload performance
- Intel® Multi-Purpose GPU (Artic Sound-M) for AI workload processing and inference

Advanced Security Technologies

- Intel® Software Guard Extensions (Intel® SGX)
- Intel ® Total Memory Encryption (Intel® TME)

Optimized Data Streaming & Advanced I/O

- Intel® Data Streaming Accelerator (Intel® DSA) optimizing streaming data movement common with media applications and converged services with Multi-Access
- PCIe Gen 5 boosting link transfer rate to 16 Gt/s

Flexible Configuration to Meet Diverse Edge Services

- Intel® Speed Select Technology (Intel® SST) improve performance and optimize TCO by providing more control over CPU performance
- Intel® Resource Director Technology (Intel® RDT) monitoring and controlling shared resources to enable diverse QoS/SLAs

AI and Analytics Tools - Big Picture

Intel® oneAPI Software Tools for AI and Analytics

Intel® oneAPI Toolkits

Intel® oneAPI AI Analytics Toolkit (AI Kit)

Accelerate machine learning and data science pipelines with optimized deep learning frameworks and high-performing Python libraries

Data Scientists, AI Researchers, DL/ML Developers

Intel® oneAPI Base Toolkit (Base Kit)

Incl. Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), and Intel® oneAPI Data Analytics Library (oneDAL)

Optimize primitives for algorithms and framework development

DL Framework Developers - Optimize algorithms for Machine Learning and Analytics

Toolkit Powered by oneAPI

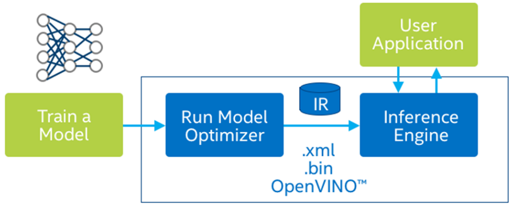
Intel® Distribution of OpenVINO™ Toolkit

Deploy high performance inference and applications from edge to cloud

AI Application, Media, and Vision Developers

OpenVINO™

Intel® Distribution of OpenVINO™ Toolkit



Deep Learning

Intel® Deep Learning Deployment Toolkit

Model Optimizer Convert & Optimize → IR → **Inference Engine** Optimized Inference

IR = Intermediate Representation file

Open Model Zoo

- 40+ Pretrained Models
- Sample Apps
- Model Downloader

Deep Learning Workbench

- Calibration Tool
- Model Analyzer
- Benchmark App
- Accuracy Checker
- Aux. Capabilities

Traditional Computer Vision

Optimized Libraries & Code Samples

- OpenCV*
- OpenVX*
- Samples

For Intel CPU & GPU/Intel® Processor Graphics

Tools & Libraries

Increase Media/Video/Graphics Performance

- Intel® Media SDK Open Source version
- OpenCL™ Drivers & Runtimes

For GPU/Intel® Processor Graphics

Optimize Intel® FPGA (Linux* only)

- FPGA RunTime Environment (from Intel® FPGA SDK for OpenCL™)
- Bitstreams

OS Support: CentOS* 7.4 (64 bit), Ubuntu* 16.04.3 LTS (64 bit), Microsoft Windows* 10 (64 bit), Yocto Project* version Poky Jethro v2.0.3 (64 bit), macOS* 10.13 & 10.14 (64 bit)

Intel® Architecture-Based Platforms Support

Intel® Vision Accelerator Design Products & AI in Production/ Developer Kits

An open source version is available at 01.org/openvinotoolkit (deep learning functions support for Intel CPU/GPU/NCS/GNA).

OpenVX and the OpenVX logo are trademarks of the Khronos Group Inc.
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

Inference Workload Optimization through OpenVINO™

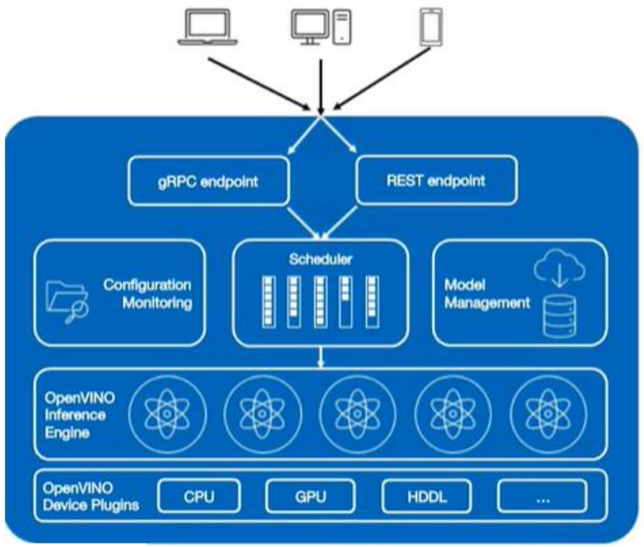
OpenVINO

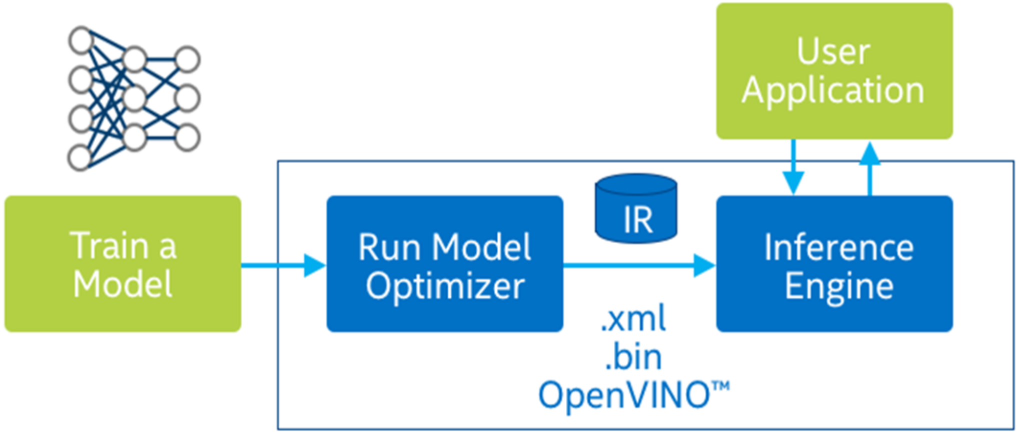
- AI Inference Optimization on Intel Platforms
- Deep Learning models with sample apps (detection/classification/segmentation)

OpenVINO Model Server (OVMS)

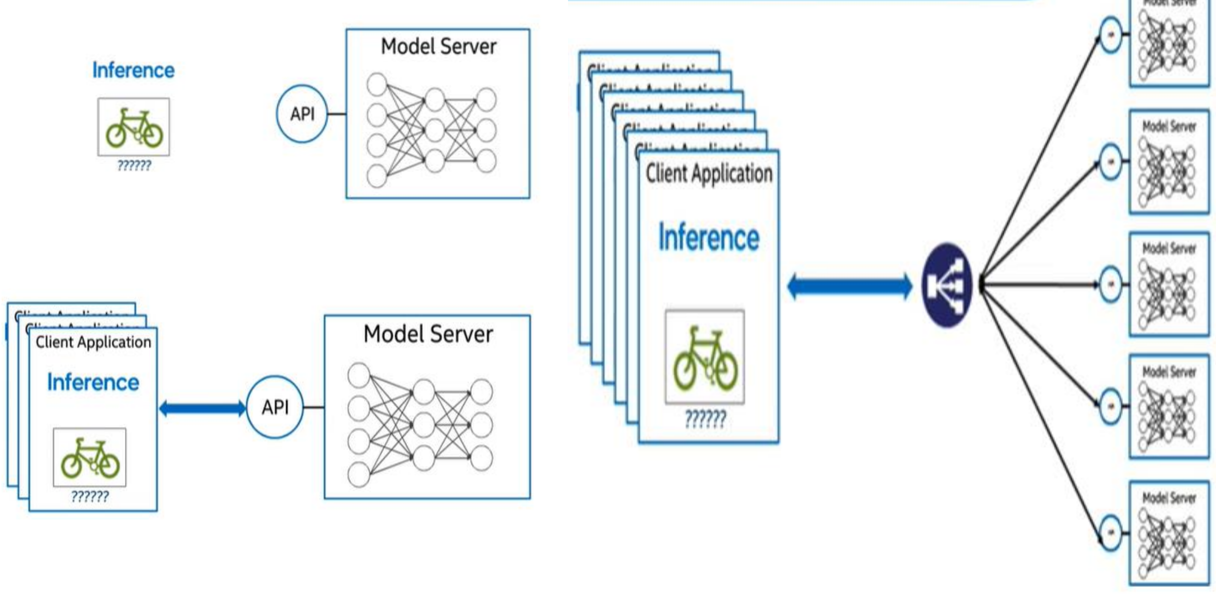
- Production grade inference server
- Ease edge AI applications deployment & scaling

OpenVINO Model Server (OVMS)





OpenVINO Inference Optimization



What is Intel® Smart Edge?

Edge-native Kubernetes Certified Distributed Computing Platform

certified kubernetes

- Enables deployment and management of container-based workloads with cloud-like ease, resiliency and security at the edge.
- Runs demanding workloads like AI, media, and software-defined networking functions, powered by pre-validated blueprints and solutions provided by Intel and a robust partner ecosystem.
- Powers diverse use cases across industries and delivers performance, security, manageability and sustainability.

Intel Smart Edge for Developers
Develop with us

- Developer Experience Kits* for Edge Video Analytics, ORAN xApps
- Apache 2.0 license
- Community Support
- Available via Github, Intel® Developer Catalog & Intel® DevCloud

Intel Smart Edge for Builders
Build with us

- Builder Experience Kits* for Private Wireless, SASE, Access Edge, Near Edge
- Intel royalty-free license
- Intel Basic Support
- Available via Github

Intel Smart Edge for Enterprises
Buy from us

- Turn-key commercial software for private wireless: Edge Controller and Edge Node
- Intel paid license
- Intel Premium Support
- Available via Github

Partner Market Ready Solutions, Partner Commercial Applications, Intel Reference Applications

Consumable as integrated platforms or composable building blocks

Cloud-Native & Modular Approach

Intelligent, Security & Multi-Access Edge Solutions

Kubernetes Plugins for Intel IPs (eASIC, FPGA, QAT, NICs, GPU)

Modularity

Horizontal building blocks (Intel SDKs and SW toolkits)

Micro-Services-Based

AI-Inference-aaS

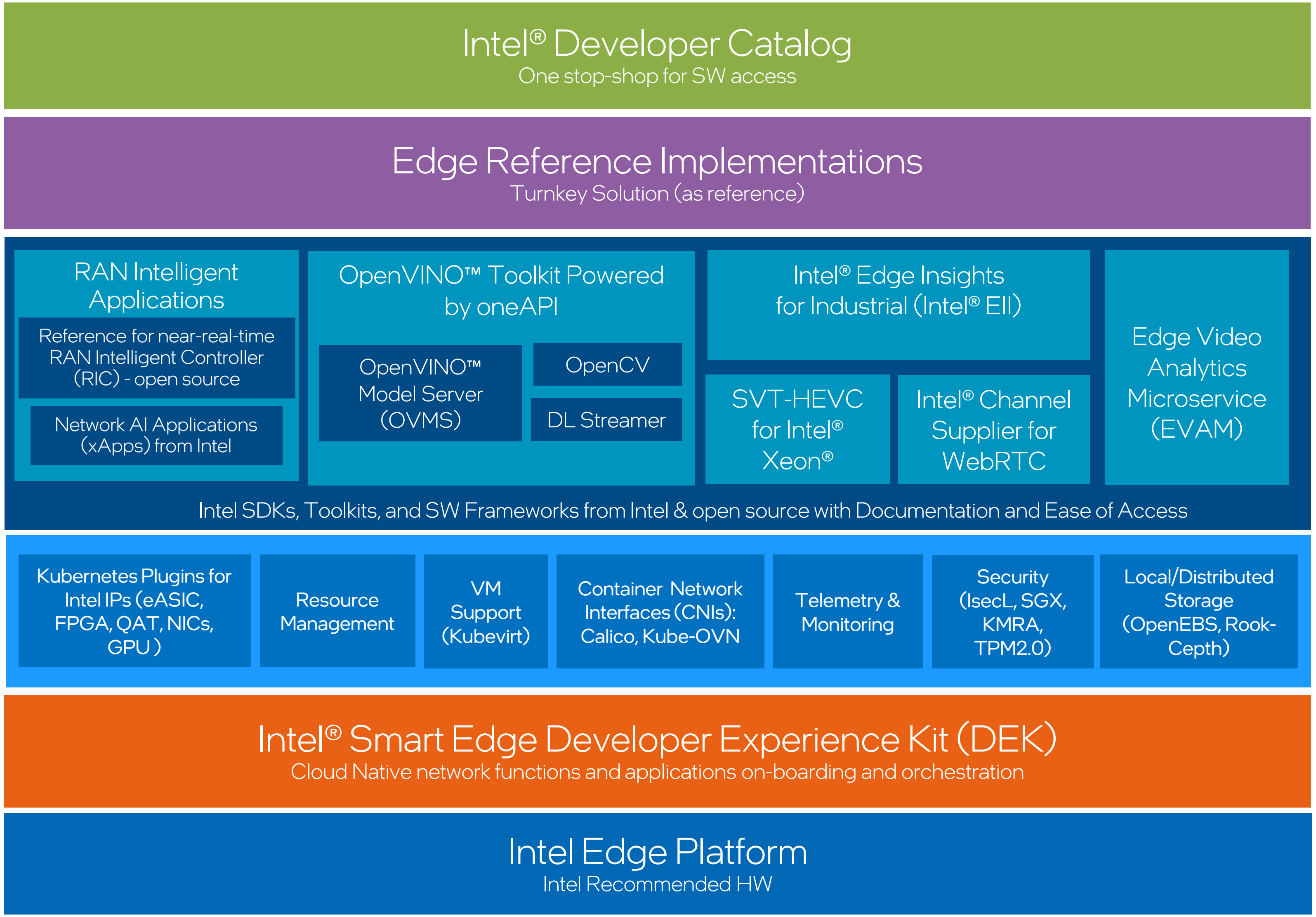
Real-time-communication (RTC)-aaS

Data-Ingestion-aaS

Data-Insights-aaS

RAN Intelligence - aaS

....more to come...



Intel® Developer Catalog
One stop-shop for SW access

Edge Reference Implementations
Turnkey Solution (as reference)

RAN Intelligent Applications
- Reference for near-real-time RAN Intelligent Controller (RIC) - open source
- Network AI Applications (xApps) from Intel

OpenVINO™ Toolkit Powered by oneAPI
- OpenVINO™ Model Server (OVMS)
- OpenCV
- DL Streamer

Intel® Edge Insights for Industrial (Intel® EII)

SVT-HEVC for Intel® Xeon®

Intel® Channel Supplier for WebRTC

Edge Video Analytics Microservice (EVAM)

Intel SDKs, Toolkits, and SW Frameworks from Intel & open source with Documentation and Ease of Access

Kubernetes Plugins for Intel IPs (eASIC, FPGA, QAT, NICs, GPU)

Resource Management

VM Support (Kubevirt)

Container Network Interfaces (CNIs): Calico, Kube-OVN

Telemetry & Monitoring

Security (IsecL, SGX, KMRA, TPM2.0)

Local/Distributed Storage (OpenEBS, Rook-Cepth)

Intel® Smart Edge Developer Experience Kit (DEK)
Cloud Native network functions and applications on-boarding and orchestration

Intel Edge Platform
Intel Recommended HW

Developer Outreach

Intel® Developer Catalog

Intel® DevCloud (soon)

Recommended HW

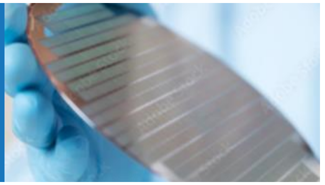
Intel Edge Platforms

Intel® Smart Edge Open: Reference Implementations

Reference Implementations Only Available in the Intel® Developer Catalog

PCB Defect Detection

Optimized video streams ingestion, edge AI inference , optimized apps on-boarding



Telepathology

AI inference, Inference scaling, reduce large storage needs for medical images, optimized apps on-boarding

Telehealth

Real-time communication, Video decode/encode SW acceleration, AI inference, optimized apps on-boarding

Immersive Media

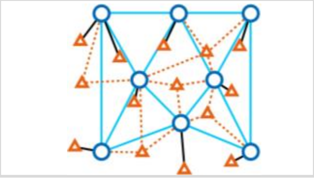
Real-time communication, 360 video decode/encode SW acceleration, 360 frames AI inference, optimized apps on-boarding

Wireless-Ready Smart Intersection

Reduced TCO, Edge AI Inference, AI models Optimization, optimized apps on-boarding

Intelligent Connection Management

Deep reinforcement learning (DRL) algorithm, graph neural network (GNN) model to implement networks



Characteristics of MEC Platforms for AI

- Open standard interfaces
- Ability to gather required data from across applications, end-to-end services, platform and hardware capabilities
- Ability to filter through for required data sets
- Secure access and secure transport
- Easy life cycle management supporting ease of operation

Challenges in Deploying AI at the Edge

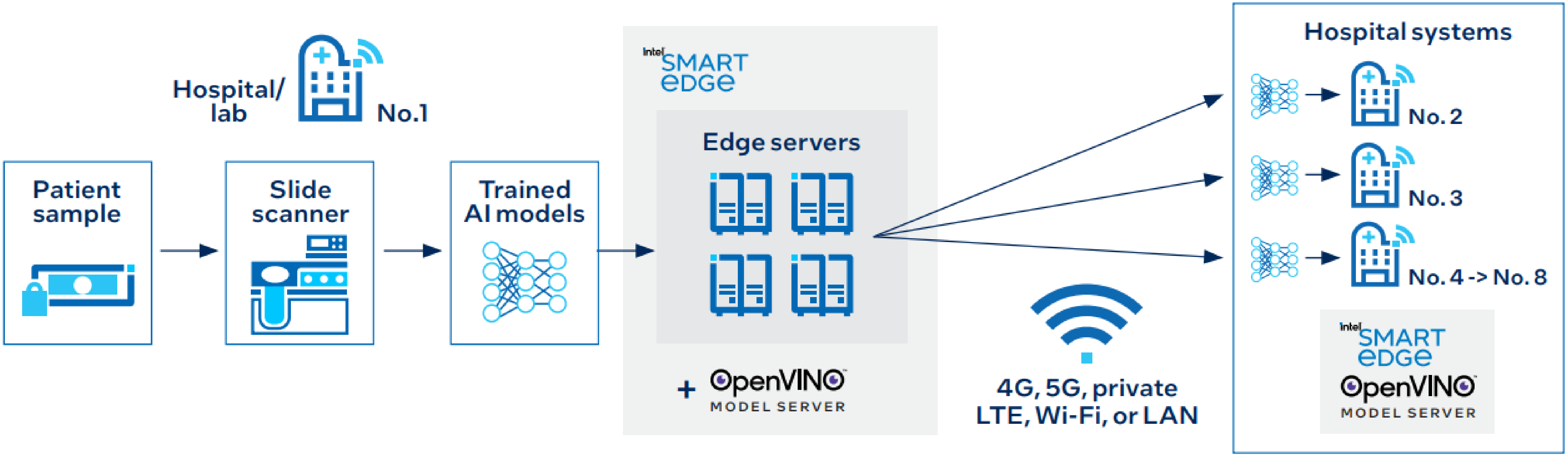
- Resource Constraints
- High Network Performance
- Policy Management
- Seamless scale between Edge to Cloud
- Security & Privacy
- Application onboarding
- Life cycle management
- Public and Private Cloud
- Hardware Abstraction & Utilization
- AI & ML Models for Edge
- Automation/Operation Autonomy

AI in Healthcare

Example Use Cases

Digital Pathology with Edge Computing

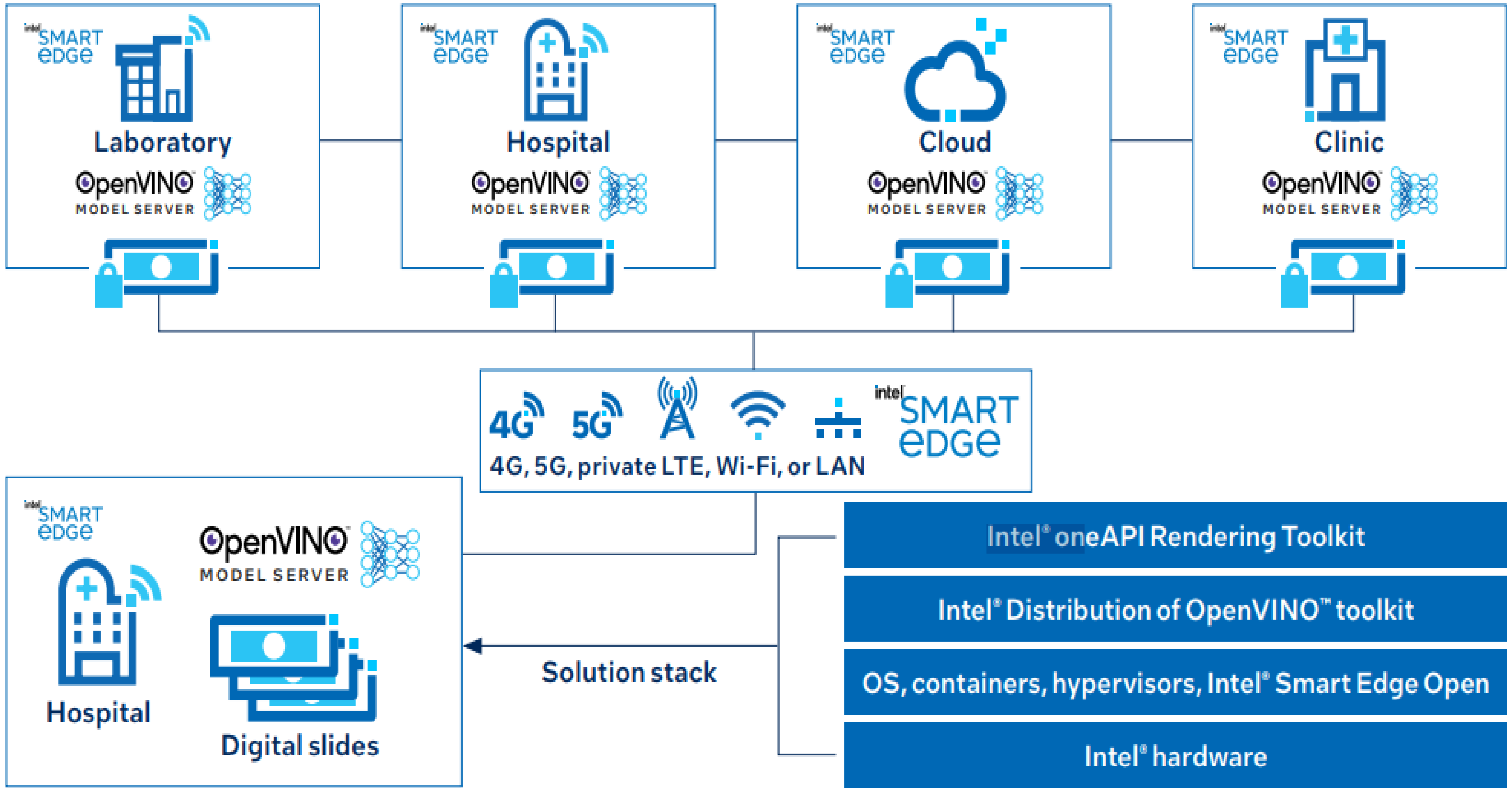
- Roughly half of pathologists worldwide live in USA with ~4.3% of world's population!
- Digital pathology converts physical glass slides into digitized images
- AI enhances operational efficiency by deploying AI models for image resolution, classification & segmentation



Source: https://www.intel.com/content/www/us/en/healthcare-it/resources/digital-pathology-business-brief.html

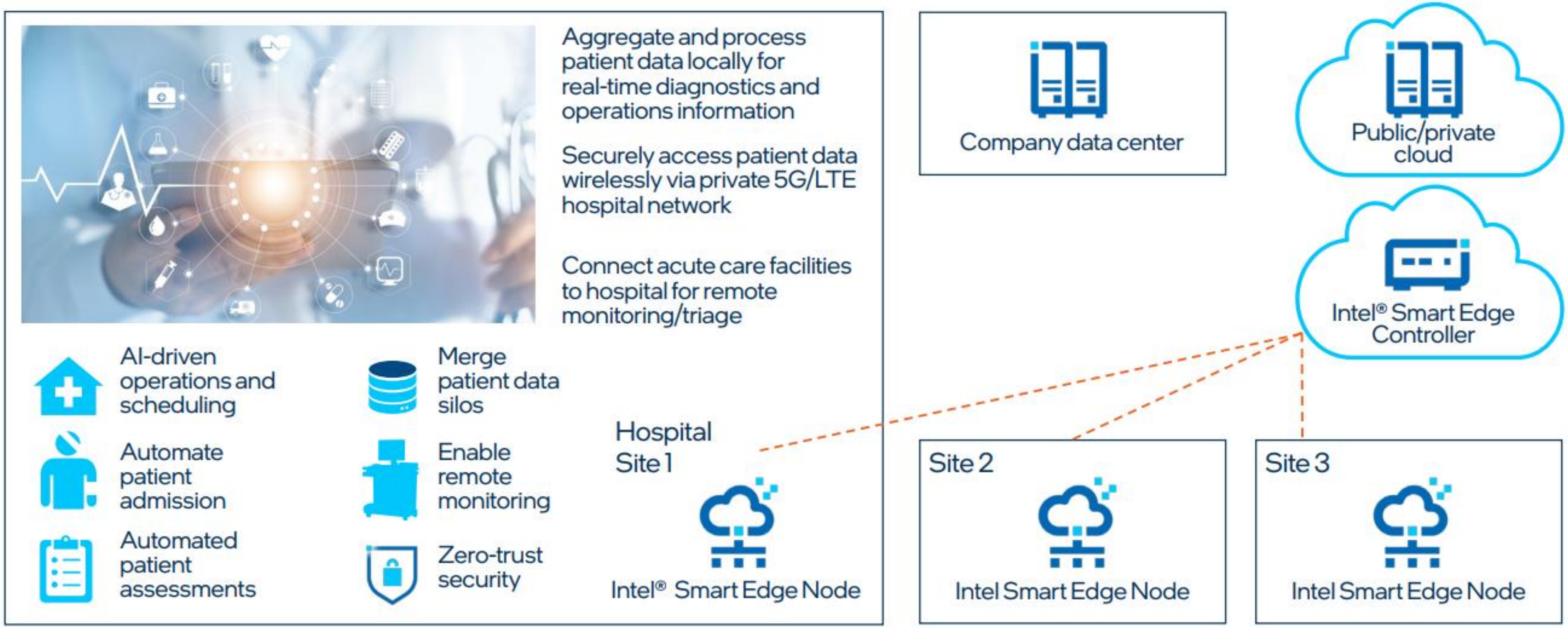
AI Model Manageability & Service Orchestration - At Scale

- OpenVINO™ Model Server centralizes AI model management & helps optimize AI model inference
- Intel® Smart Edge Open accelerates Edge service deployments at scale
- Intel® OneAPI Rendering Toolkit enhances whole slide image visualization



Source: https://www.intel.com/content/www/us/en/healthcare-it/resources/digital-pathology-business-brief.html

Connected Healthcare with Intel® Smart Edge



Aggregate and process patient data locally for real-time diagnostics and operations information

Securely access patient data wirelessly via private 5G/LTE hospital network

Connect acute care facilities to hospital for remote monitoring/triage

AI-driven operations and scheduling

Automate patient admission

Automated patient assessments

Merge patient data silos

Enable remote monitoring

Zero-trust security

Hospital Site 1

Intel® Smart Edge Node

Company data center

Public/private cloud

Intel® Smart Edge Controller

Site 2

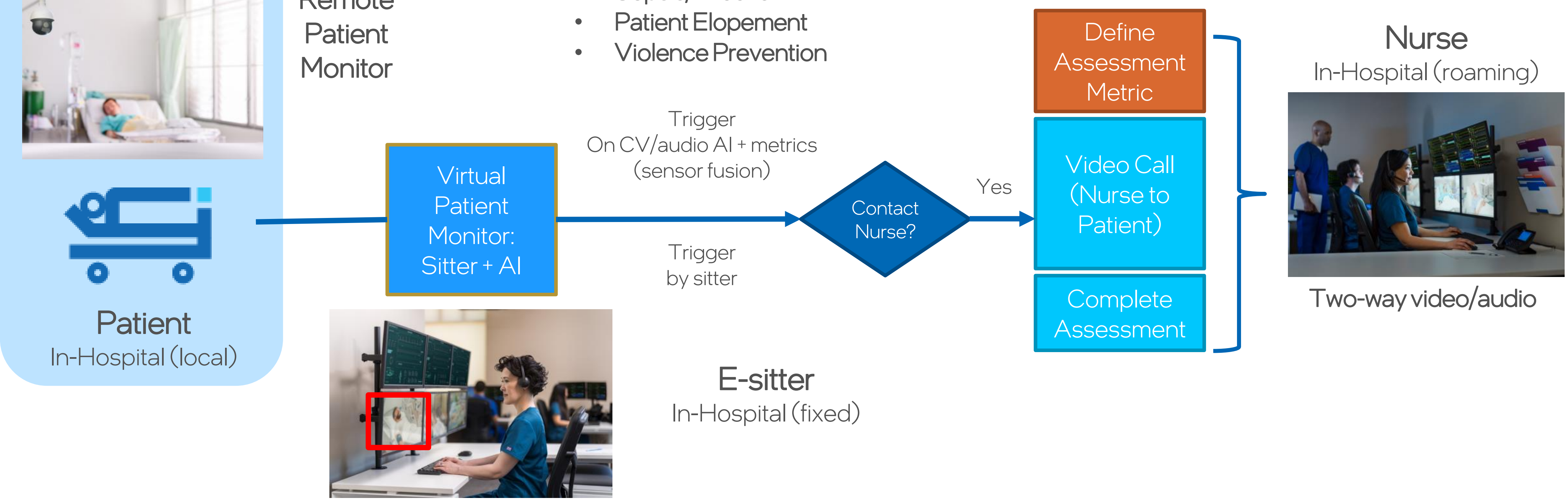
Intel Smart Edge Node

Site 3

Intel Smart Edge Node

Source: https://www.intel.com/content/www/us/en/healthcare-it/resources/smart-hospital-technical-brief.html

Remote Patient Monitoring w/ Multi-modal AI + Telemedicine



In-Room 2-way video/audio tablet and fixed IP camera

Patient

In-Hospital (local)

Remote Patient Monitor

Virtual Patient Monitor: Sitter + AI

Two-way video/audio

AI in Real-time to Detect:

- Fall Risk
- Sepsis/Infection
- Patient Elopement
- Violence Prevention

Trigger
On CV/audio AI + metrics
(sensor fusion)

Trigger
by sitter

E-sitter

In-Hospital (fixed)

Contact Nurse?

Yes

AI for Automation

- Automated patient assessment
- Auto-populate charts/EMR
- Voice-activated controls
- Hand/motion activated controls

Define Assessment Metric

Video Call (Nurse to Patient)

Complete Assessment

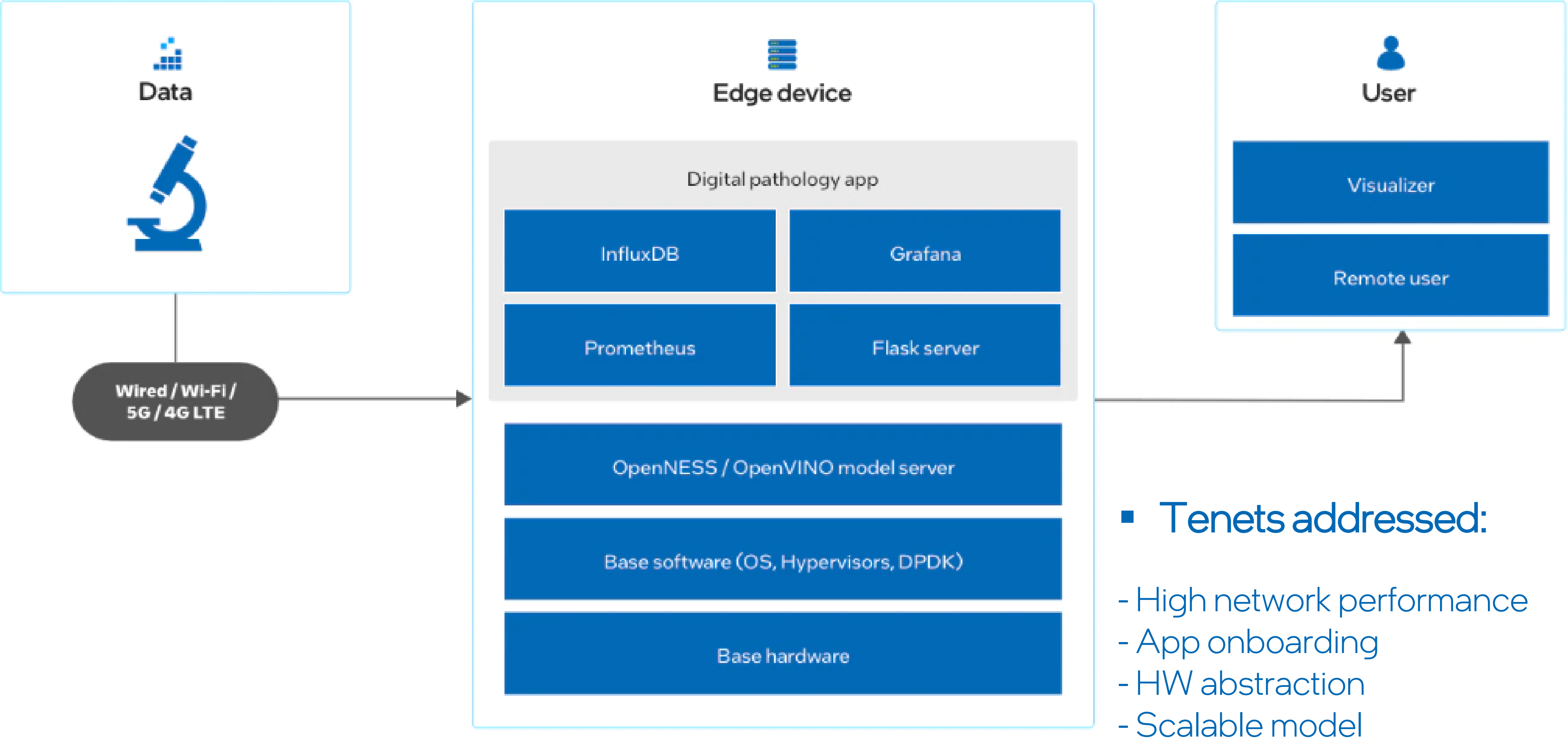
Nurse

In-Hospital (roaming)

Two-way video/audio

Check it Out – Telepathology Reference Implementation with Intel® Smart Edge Open

- Network Optimization and AI Inferencing Management for Telepathology
- Enables digital pathology through lab analysis automation
- Automated network abstraction, which helps avoid complex data routing and traffic shaping and gives confidence in efficient data sharing and AI model utilization
- Reduced 'hands-on' management for data routing as well as AI model optimization within the IT infrastructure



- **Tenets addressed:**

- High network performance
- App onboarding
- HW abstraction
- Scalable model
- Platform & application security

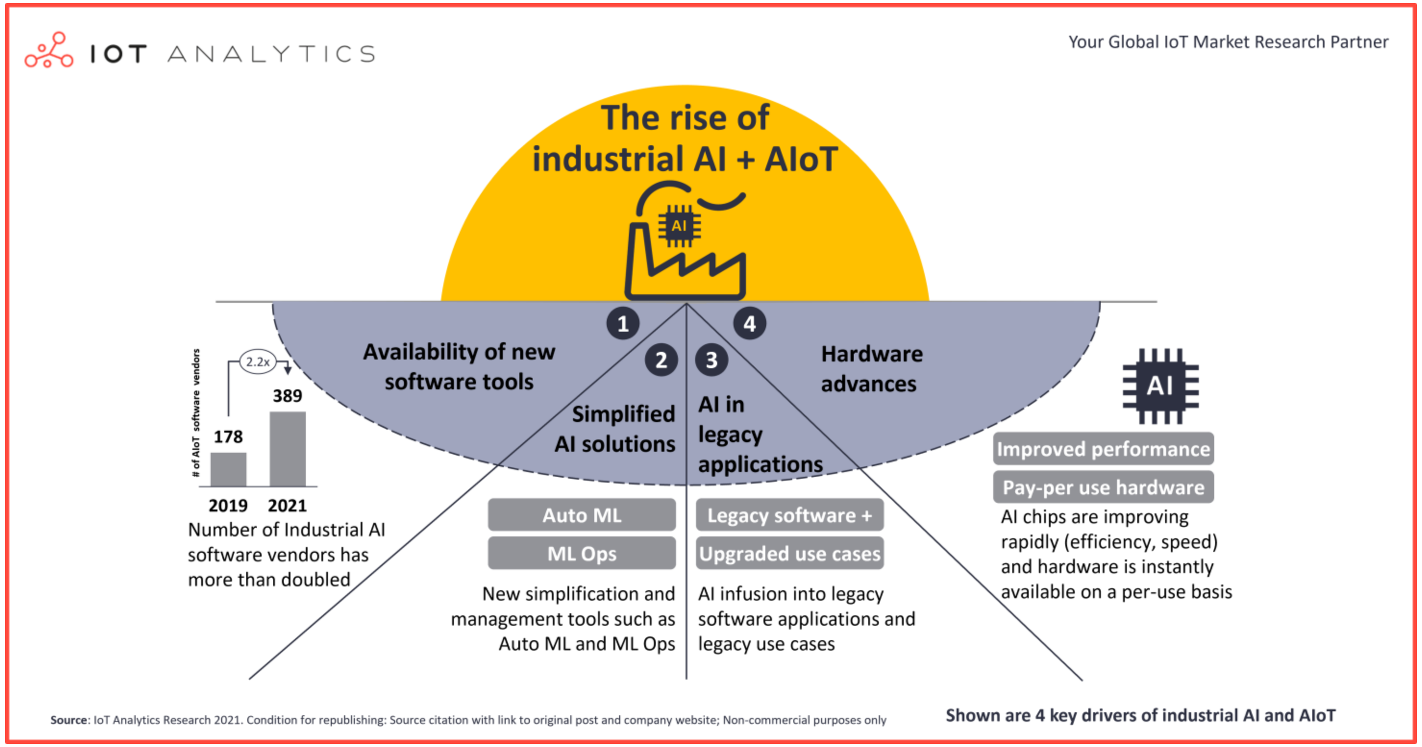
Source: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/network-optimization-ai-inferencing-telepathology.html

AI in Industrial IoT

Example Use Cases

Advances in Industrial Ecosystem

- Digital Transformation
- IT/OT Convergence
- Multiple sectors – Oil & Gas, Manufacturing, Utilities, logistics, etc.
- Varying requirements based on the industry – Automation, vision processing, scalability, data processing, time sensitive, predictive analytics, etc.

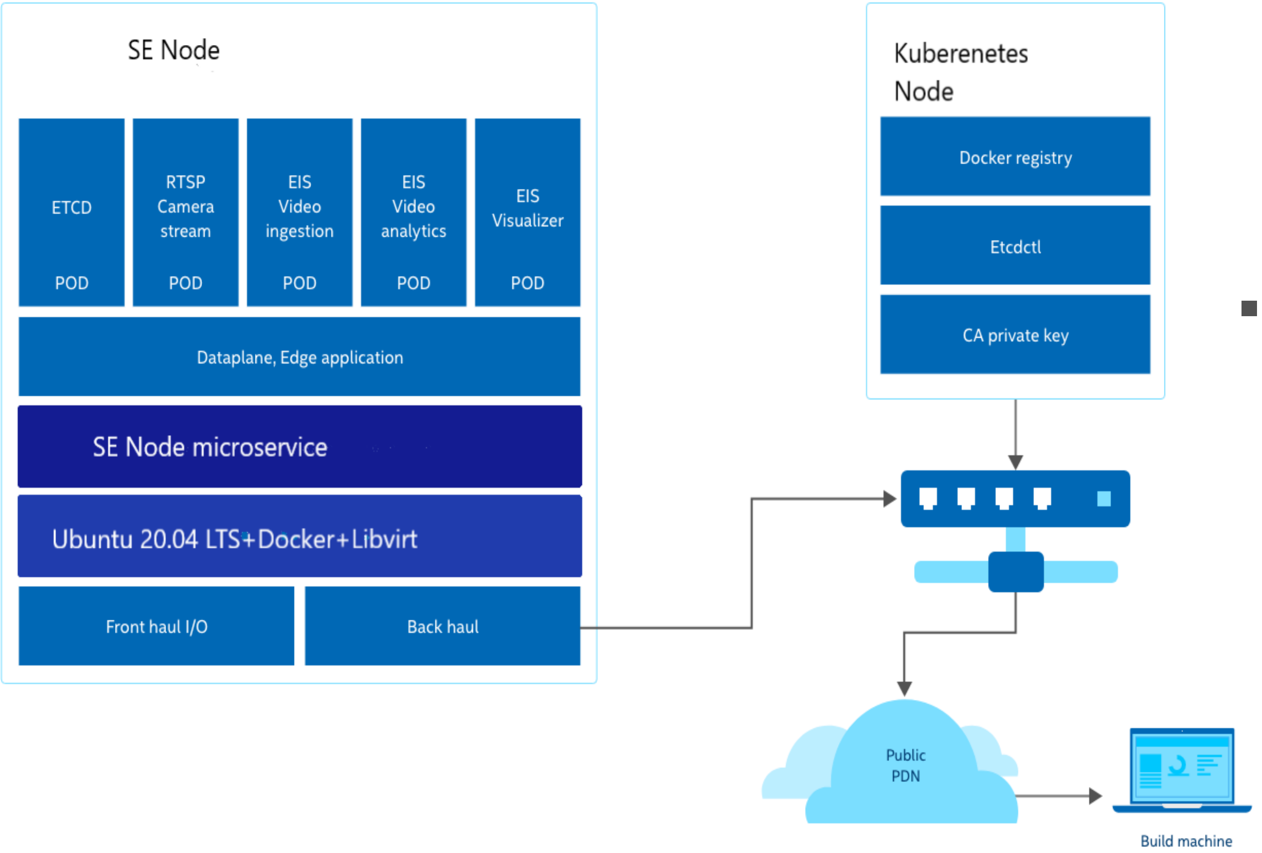

IOT ANALYTICS
Your Global IoT Market Research Partner
The rise of industrial AI + AIoT
AI
1
Availability of new software tools
# of AIoT software vendors
2.2x
178
389
2019
2021
Number of Industrial AI software vendors has more than doubled
2
Simplified AI solutions
Auto ML
ML Ops
New simplification and management tools such as Auto ML and ML Ops
3
AI in legacy applications
Legacy software +
Upgraded use cases
AI infusion into legacy software applications and legacy use cases
4
Hardware advances
AI
Improved performance
Pay-per use hardware
AI chips are improving rapidly (efficiency, speed) and hardware is instantly available on a per-use basis

Source: IoT Analytics Research 2021. Condition for republishing: Source citation with link to original post and company website; Non-commercial purposes only

Shown are 4 key drivers of industrial AI and AIoT

Image source: https://iot-analytics.com/rise-of-industrial-ai-aiot-4-trends-driving-technology-adoption/

Intel® Smart Edge PCB Defect Detection Reference Implementation



- Helps deploy a solution for Printed Circuit Board (PCB) defect detection using AI for product quality checks and enabled by Intel Edge insights for Industrial (EII) and Intel® Smart Edge Open Developer Experience Kit platform.
- Supports two types of defect detection: missing components and short circuits due to solder bridge formed during the assembly process.
  - **Tenets addressed:**

- High network performance
- Vision processing infrastructure
- HW abstraction
- Scalable model
- Platform & application security
- Building blocks using EII

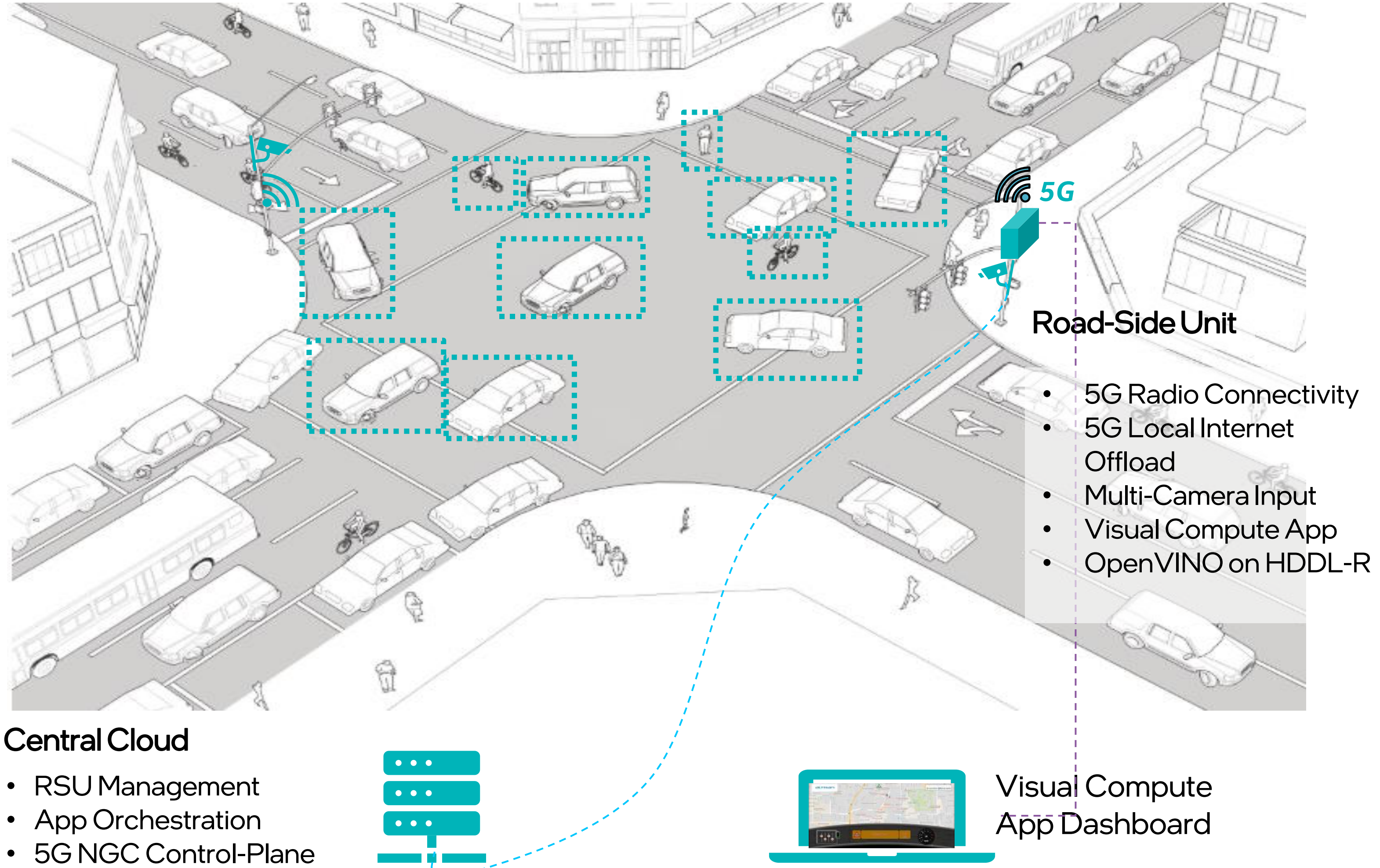
Source: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/wireless-network-ready-pcb-defect-detection.html

AI in Smart Cities

Example Use Cases

5G Smart Road-Side Infrastructure Platform

Foundation Kit for Visual Compute + 5G Smart Road-Side Infrastructure



Road-Side Unit

- 5G Radio Connectivity
- 5G Local Internet Offload
- Multi-Camera Input
- Visual Compute App
- OpenVINO on HDDL-R

Central Cloud

- RSU Management
- App Orchestration
- 5G NGC Control-Plane

Visual Compute App Dashboard

Road-Side Unit (RSU)

- Hardware Foundation Kit based on Intel Hardware Platform Single-Socket, Intel® Xeon® D, Intel® Xeon® Scalable processor
- The RSU Software Platform integrates Capgemini Engineering ENSCONCE Edge PaaS
- The RSU Software Platform will have integrated support for OpenVINO™
- Intel HDDL-R accelerator based VPU Offload for Intel OpenVINO Apps
- Intel PAC N3000 FPGA for 5G Layer-1 offload
- The RSU can also provide 5G Connectivity supported through Capgemini Engineering 5G gNodeB L2/3 and Intel® FlexRAN L1/FPGA
- Capgemini Engineering 5G NGC UPF for local data offload

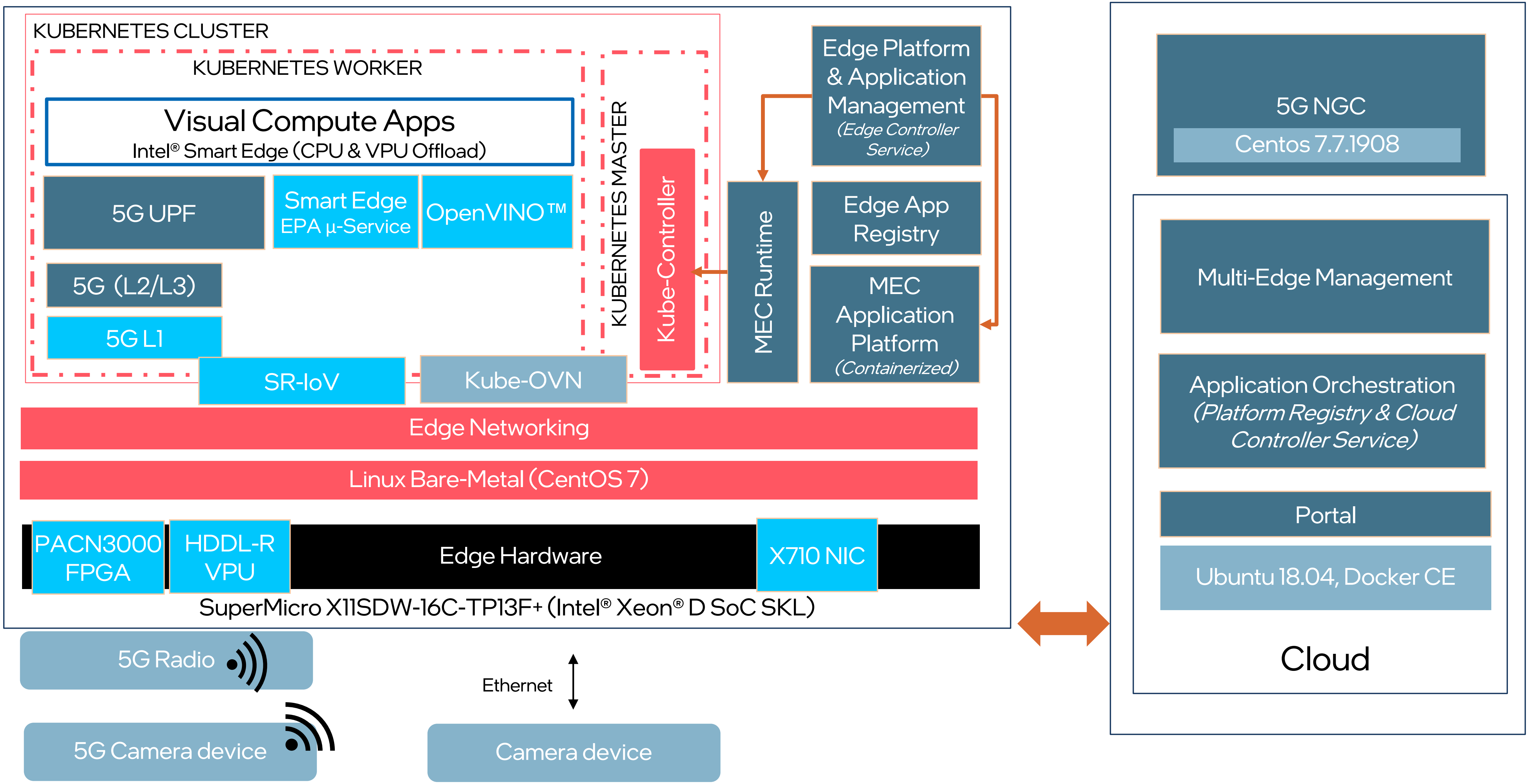
Central Cloud

- The Central Cloud would consist of Capgemini ENSCONCE Cloud PaaS (Platform as a Service)
- ENSCONCE Customer Portal
- Capgemini Engineering 5G NGC

Edge- Applications

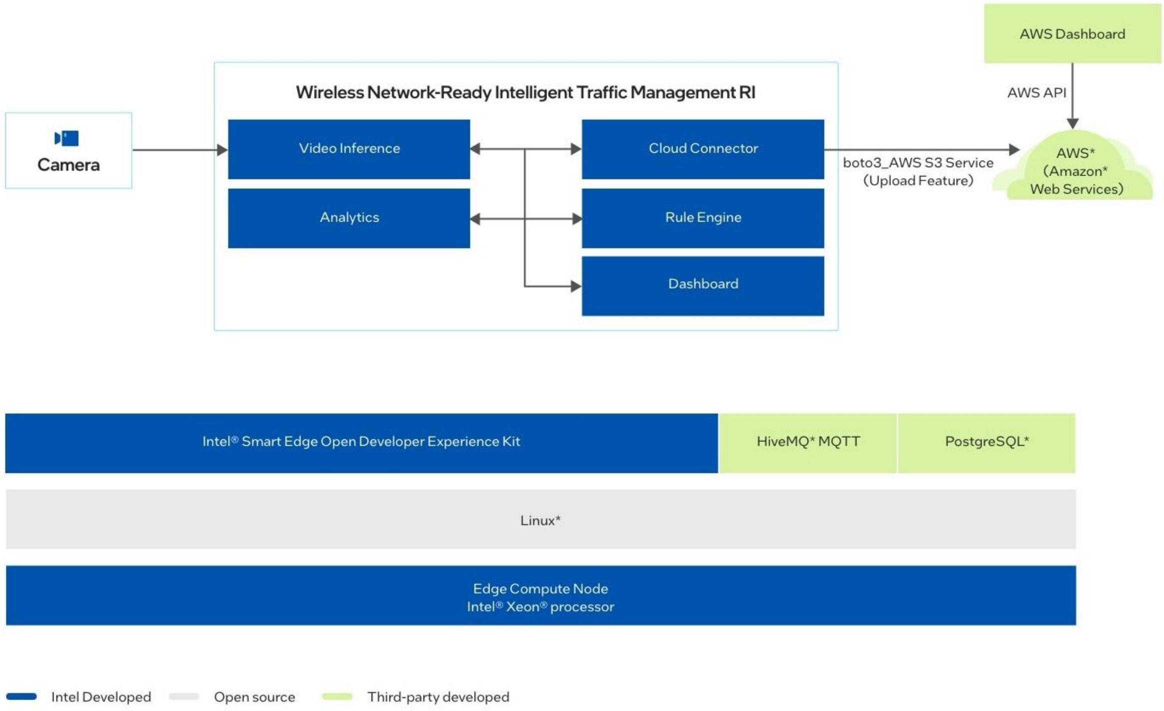
- Run Visual Compute Inference Applications using Intel OpenVINO and HDDL-R Accelerators
- Stream Traffic Meta-Data for V2X Applications

5G Smart Connected Platform : Architecture



Intel® Smart Edge Intelligent Traffic Management Reference Implementation

Easy to Scale Solutions



AWS Dashboard

AWS API

Wireless Network-Ready Intelligent Traffic Management RI

Camera

Video Inference

Analytics

Cloud Connector

Rule Engine

Dashboard

boto3_AWS S3 Service (Upload Feature)

AWS* (Amazon* Web Services)

Intel® Smart Edge Open Developer Experience Kit

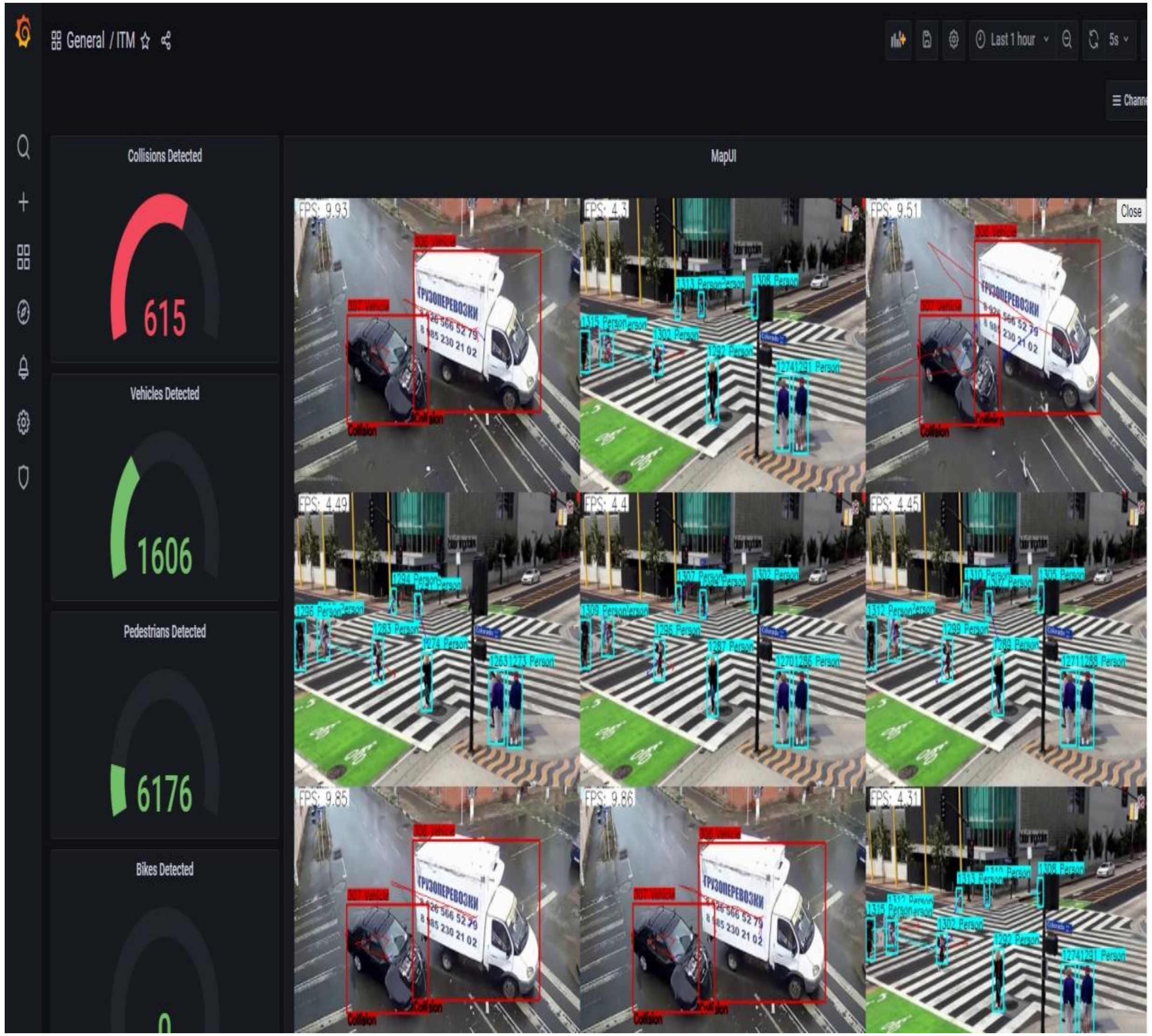
HiveMQ* MQTT

PostgreSQL*

Linux*

Edge Compute Node
Intel® Xeon® processor

Intel Developed · Open source · Third-party developed



General / ITM

Last 1 hour · 5s

Collisions Detected: 615

Vehicles Detected: 1606

Pedestrians Detected: 6176

Bikes Detected: 0

MapUI

- **Tenets addressed:**
  - Low latency communication
  - Vision processing infrastructure
  - HW abstraction
  - Edge to Cloud Scalable model
  - Platform & application security

Source: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/wireless-network-ready-intelligent-traffic-management.html

Solutions to Scale

Infrastructure & Partnerships

Intel Edge & Networking Testbed

To accelerate the edge ecosystem, we have set up a 5G & MEC end-to-end Testbed to offer an environment to verify solutions and to provide innovative solutions to accelerate the commercial deployments of edge

This testbed will enable customers to see real results with applications running on MEC and working with 5G.

Automated customer ready pre-install of DEK

Enabling PWEK and other EKs (over the air radio)

Increased edge accelerator support (ACC, GPU, QAT etc.)

Edge use case support (cameras, sensors etc.)

Experience Centers and Support for edge demos

Engage with us to explore possibilities!

Hillsboro
- External Connectivity ✓
- Radio ✓

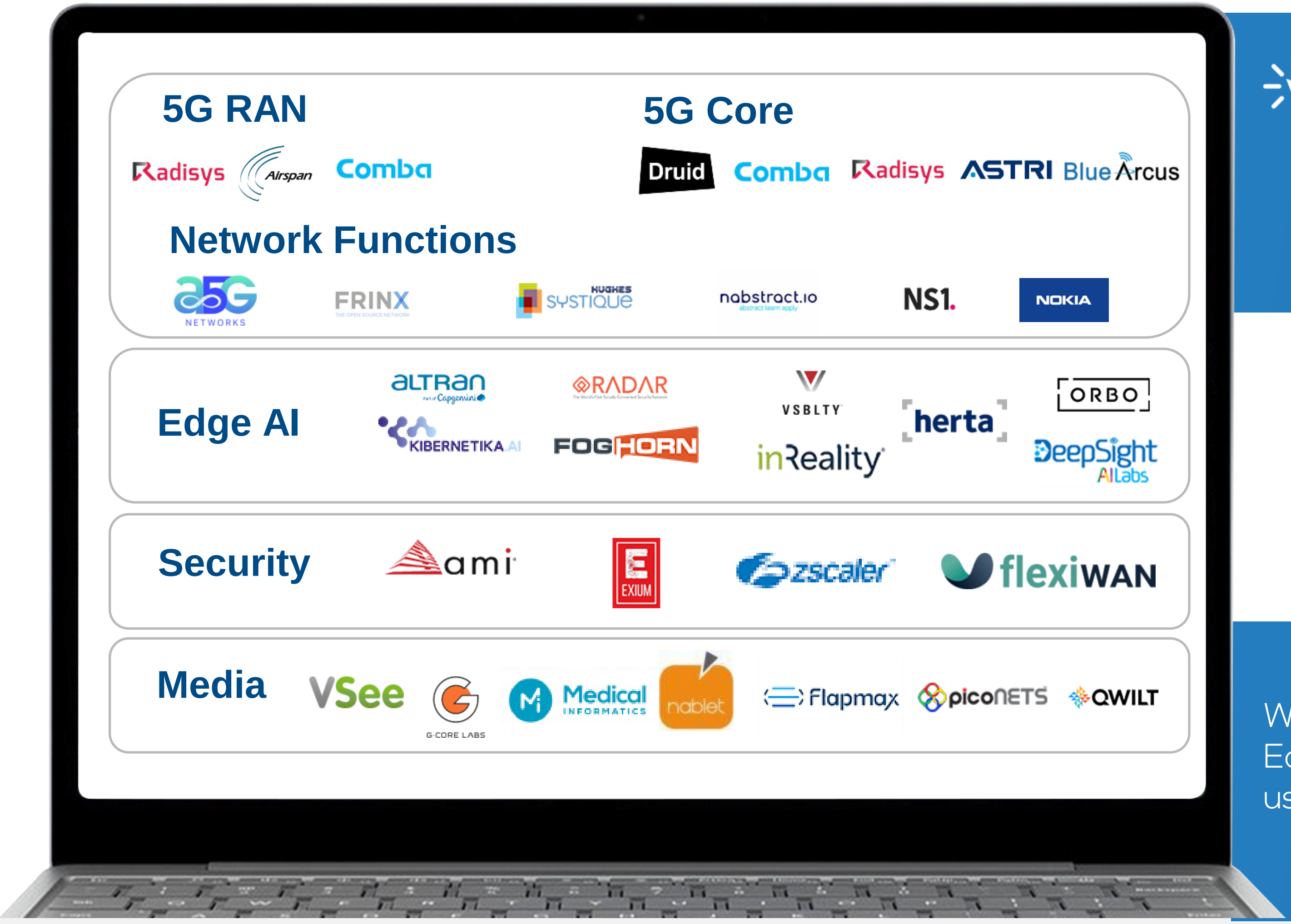
Beijing
- Radio ✓, IoT ✓

Rio Rancho
- Main Scaling Engine ✓
- External Connectivity ✓

Bangalore
- External Connectivity ✓
- Radio ✓

Intel® Smart Edge: Commercial Edge Applications

ISV developed commercial applications that are pre-integrated, tested and can be deployed on the Intel® Smart Edge platform



5G RAN

Radisys · Airspan · Comba

5G Core

Druid · Comba · Radisys · ASTRI · BlueArcus

Network Functions

a5G Networks · FRINX · Hughes Systique · nabstract.io · NS1. · NOKIA

Edge AI

altran · RADAR · VSBLTY · herta · ORBO · KIBERNETIKA.AI · FOGHORN · inReality · DeepSight AILabs

Security

ami · EXIUM · zscaler · flexiWAN

Media

VSee · G-CORE LABS · Medical Informatics · nablet · Flapmax · picoNETS · QWILT

Find us at:
[Smart Edge Commercial Applications Portal]

Accelerate Time To Market

Discover a Growing Ecosystem

Scale with a Robust Edge Portfolio

We have done the work and made it easy for Edge Infrastructure Builders to deploy innovative use cases intel®