MobiledgeX

Building On-Demand Edge Clouds with Mixed Infrastructure

Thomas Vits, Technical Partnerships, MobiledgeX
Francesc Guim, Principal Architect and Edge Chief, Intel CTO office NEX
March 29, 2022

Notices and Disclaimers

- Intel technologies may require enabled hardware, software or service activation.
- No product or component can be absolutely secure.
- Your costs and results may vary.
-

Multi-tiers

TIERS	ON-PREMISE EDGE				
	INTELLIGENT SENSOR/GW	INTELLIGENT EDGE	NETWORK EDGE	DC EDGE	PUBLIC CLOUD
NETWORK LATENCIES (Wire Round trip)	< 1ms	< 1ms	1-5ms	5ms + 1-2 ms (every 100kms)	5ms + 5 ms (every 400kms)
DEPLOYMENT REQUIREMENTS	**Compute Available Power:** < 50 W **Form Factor:** Small Box **Thermals:** NEBs **Mgmt.:** Remote	**Compute Available Power:** ~10KW **Form Factor:** Rack(s) **Thermals:** NEBS or Standard DC **Mgmt.:** Remote	**Compute Available Power:** <600 W **Form Factor:** Pizza box **Thermals:** NEBS **Mgmt.:** Remote	**Compute Available Power:** 9KW/rack 1KW sqm **Form Factor:** Rack(s) **Thermals:** NEBS or Standard DC **Mgmt.:** Remote	**Standard Data Center (DC)**
WHERE, WHAT & WHY	Use case KPI *Intelligent Transportation* Data Privacy, Backhaul Traffic Savings Reliability *V2V* Latency *Retail* Same as Int. Transp. *Video Analytics* Same as Int. Transp.	Use case KPI *AR/VR* Latency Backhaul Traffic Savings Network scalability *Retail* Data Privacy, Backhaul Traffic Savings Reliability *RT Streaming* Same as AR/VR *Healthcare* Access to services	Use case KPI *Intelligent Transportation* Data Privacy Backhaul Traffic Savings Reliability, throughput *V2V* Latency *Video Analytics* Same as Int. Transp. *Drone/IoT* Same as Int. Transp. *Rural* Access to services	Use case KPI *Intelligent Transportation*, Data Privacy.. *Video Analytics* Same as Int. Transp. *Drone/IoT* Same as Int. Transp. *Healthcare* Reliability Data privacy CDN & Storage Backhaul traffic Savings Throughput *FaaS* Latency *AR/VR/MR* Latency	Use case KPI *CDN* Backhaul traffic savings Throughput *Storage GW* Same as CDN

Multi-verticality

Content Delivery Networks

CDN from the edge, cache data at the edge,

Examples: Qwilt, Netflix

Resources: Storage, CPUs

Manufacturing

General AI inferencing models for industrial and real-time deployments

Examples: Defect detection, Real Time Monitoring, AGV, Logistics

Resources: Communication, CPUs, storage, FGPAS, AI Acc.

Video/Video Analytics

Live video analytics and video pre-processing, video transcoding

Examples: Traffic video analysis and alarm systems

Resources: CPUs, FPGAs, storage, AI Acc., GPU

NFV

Flexible NFV (specifically VRAN/CRAN)

Examples: De-Centralized Ran, vEPC, UPF

Resources: CPUs, storage, FPGAs, pNIC/sNIC

Speech Recognition

Speech-to-text, User commands,
Biometric Recognition

Examples: Voice assistant, customer support

Resources: Communication, CPUs, GNA-s, FPGA, AI Acc., GPU

AR/VR & Gamming

Image processing and AI for recognition and annotations, video transcoding and rendering

Examples: Google glass, VR gaming, live stadium VR experiences

Resources: CPUs, FPGAs, AI Acc.

ADAS & V2V/V2X

Assist in (autonomous) driving, infotainment

Examples: overtaking systems, V2V comm, navigation, CDNs

Resources: CPUs, storage, AI Acc,

Retail

Shops using storage and AI inferencing for enhanced in-shop user experience. Online retailers using AI and analytics. Ad bidding, etc.

Examples: age/gender recognition, trends analysis

Resources: Communication, CPUs, storage, AI Acc., GPU, FPGA

Medical Applications

AI based health analysis (inferencing), assisting medical appliances

Examples: Bone age inferencing

Resources: CPUs, Communication, storage, AI Acc., GPU, FPGA

Smart Cities & Gov.

Retail, banking, hospitality, education and transportation

Examples: Surveillance, Safety, Healthcare etc.

Resources: CPUs, Communication, storage, GPU, FPGA

Common Framework to Address Different Verticals

EDGE END TO END EDGE ARCHITECTURAL FOUNDATION: INTER-OPERABLE, OPEN & SELF MANAGED

END USERS & DEVICE LOCATION + TRANSPORT TYPE

- STREET USERS

Telco Edge

- VEHICLES
- STREET CAMERAS
- STREET SENSORS
- …

IOT Edge

- RETAIL SHOPS
- PUBLIC LOCATIONS (I.E: LIBRARIES)
- PRIVATE ENTERPRISE
- …

Enterprise Edge

INFRASTRUCTURE (5G/LTE, Wireless, Lora, Wired …)

MULTI-SERVICE (NFV & non-NFV) COMMON SERVICE TAXONOMY

NFV

Flexible NFV

De-Centralized Ran

vEPC, UPF

Resources: CPUs, storage, **FPGAs , pNIC/sNIC**

Latency: To be analyzed

Internet of Things

IoT devices in many fields such as factory automation, process automation, smart grids, V2V

Resources: Communication, CPUs, storage, **Movidius, SPH**

Latency: Factory automation : 0.25ms to 10ms

Smart grids: 3-20ms / Process automation: 50-100ms

Autonomous

Assist in autonomous driving

Examples: overtaking systems, V2V comm, navigation

Resources: CPUs, storage, **Mobileye, SPH**

Latency: Ideally <20ms, up to 100 ms

AR/VR & Gaming

Process images (image recognition) from devices, wearables and annotate useful information

Examples: Google glass

Resources: CPUs, **FPGAs, ATS**

Latency: seamless - <20 ms , sensitive- <25ms Tolerable 50-100ms

Data caching & Storage GW

Cache data at the edge for faster loading at user end

Using Edge as main storage for the devices

Examples: cache popular videos in a region, Netflix

Resources: Storage, CPUs

Latency: Not latency bound

Video/Video Analytics

Live video analytics and video pre-processing, video transcoding

Examples: Traffic video analysis and alarm systems

Resources: CPUs, **FPGAs,** storage, **ATS, SPH for Inference?**

Latency: To be analyzed

FaaS

Perform web page related pre-processing at the edge and send page to user device

Examples: web page rendering, ad block, content evaluation

Resources: CPUs, **FPGAs,**

Latency: Not slower than current web page load times

Speech Recognition

Speech-to-text, User commands,
Biometric Recognition

Resources: Communication, CPUs, **GNA-s, FPGA, SPH**

Latency: To be analyzed

Medical Applications

Assist medical appliances through connectivity and analysis
Examples: Tele surgery
Resources: CPUs, Communication, storage, **ATS, SPH**
Latency: Tele surgery <150ms without haptic, <10ms with haptic feedback

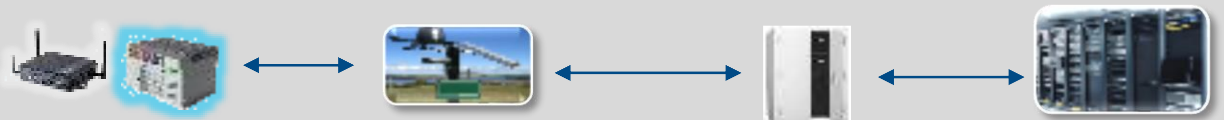
Enterprise

IMBD, Specific Enterprise WL (i.e: Linked In)

Resources: CPUs, Communication, storage, **FPGA**

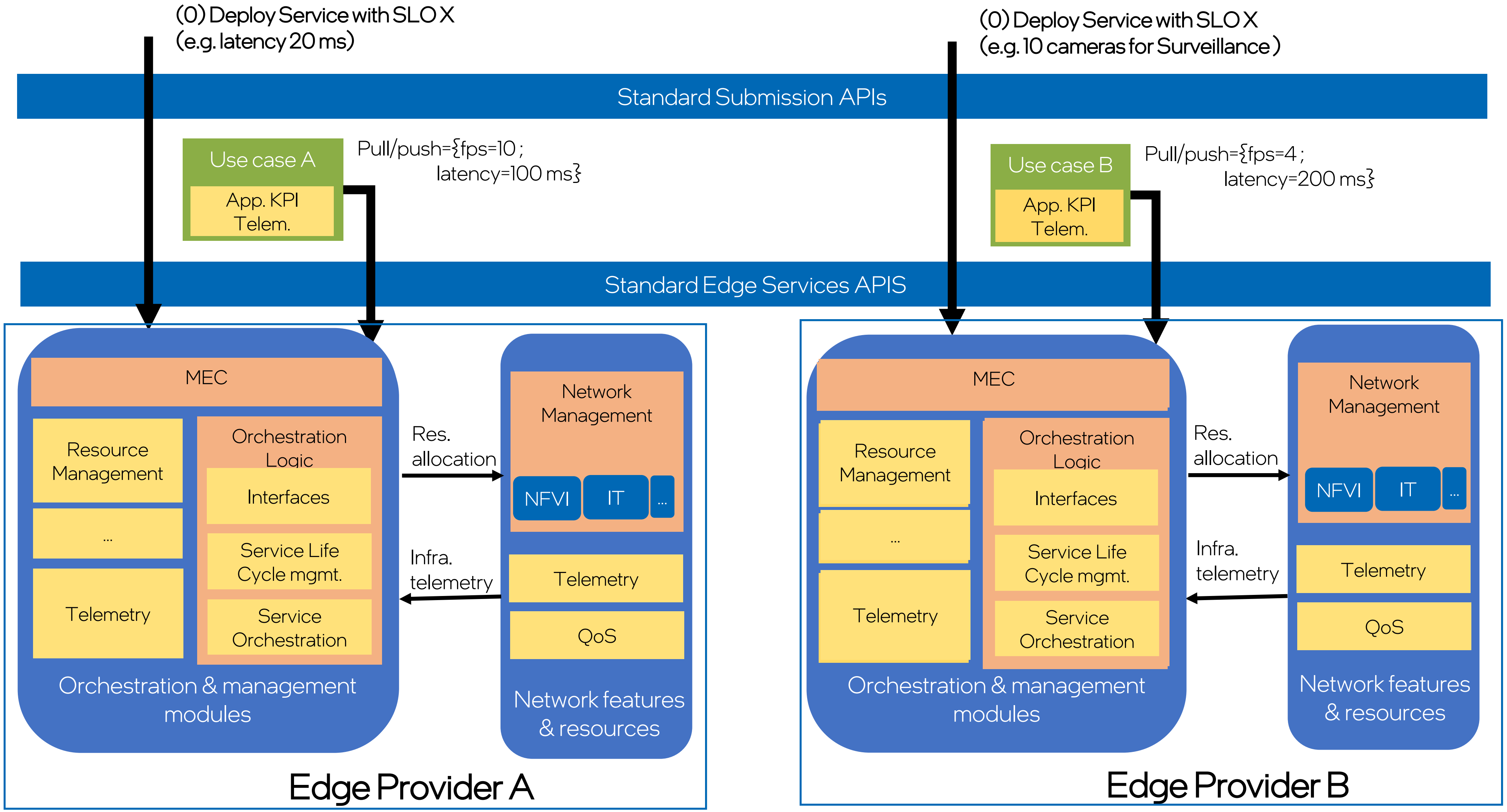
Latency: <10 ms

MULTI-TIER COMMON HW EDGE TAXONOMY



ON-PREM / FAR EDGE ↔ FAR EDGE ↔ NEAR EDGE ↔ DATA CENTER EDGE

API Consistency and Service Oriented Architectures



(0) Deploy Service with SLO X
(e.g. latency 20 ms)

(0) Deploy Service with SLO X
(e.g. 10 cameras for Surveillance)

Standard Submission APIs

Use case A

App. KPI
Telem.

Pull/push={fps=10 ;
latency=100 ms}

Use case B

App. KPI
Telem.

Pull/push={fps=4 ;
latency=200 ms}

Standard Edge Services APIS

MEC

Resource Management

...

Telemetry

Orchestration Logic

Interfaces

Service Life Cycle mgmt.

Service Orchestration

Orchestration & management modules

Res. allocation

Infra. telemetry

Network Management

NFVI

IT

...

Telemetry

QoS

Network features & resources

Edge Provider A

MEC

Resource Management

...

Telemetry

Orchestration Logic

Interfaces

Service Life Cycle mgmt.

Service Orchestration

Orchestration & management modules

Res. allocation

Infra. telemetry

Network Management

NFVI

IT

...

Telemetry

QoS

Network features & resources

Edge Provider B

Background

Opportunity

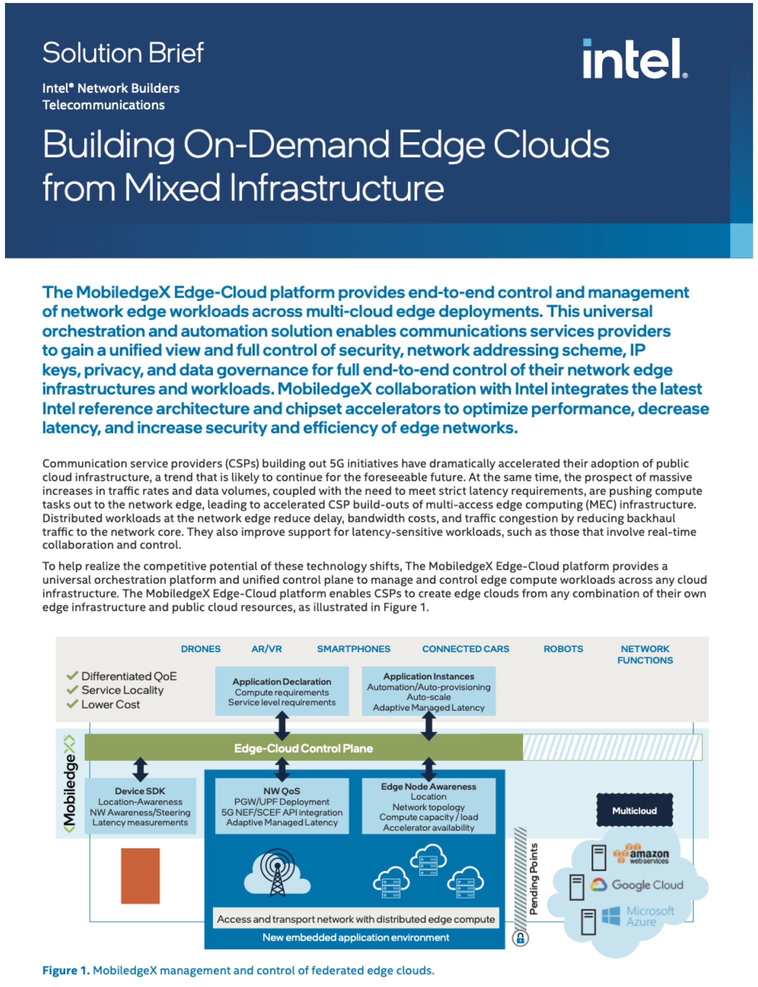
Communication service providers (CSPs) building out 5G initiatives have dramatically accelerated their adoption of public cloud infrastructure, a trend that is likely to continue for the foreseeable future.

At the same time, the prospect of massive increases in traffic rates and data volumes, coupled with the need to meet strict latency requirements, are pushing compute tasks out to the network edge, leading to accelerated CSP build-outs of multi-access edge computing (MEC) infrastructure.

Distributed workloads at the network edge reduce delay, bandwidth costs, and traffic congestion by reducing backhaul traffic to the network core. They also improve support for latency-sensitive workloads, such as those that involve real-time collaboration and control.

Solution

To help realize the competitive potential of these technology shifts, the MobiledgeX Edge-Cloud platform provides a universal orchestration platform and unified control plane to manage and control edge compute workloads across any cloud infrastructure. The MobiledgeX Edge-Cloud platform enables CSPs to create edge clouds from any combination of their own edge infrastructure and public cloud resources, as illustrated in Figure 1.



Solution Brief

intel.

Intel® Network Builders
Telecommunications

Building On-Demand Edge Clouds from Mixed Infrastructure

The MobiledgeX Edge-Cloud platform provides end-to-end control and management of network edge workloads across multi-cloud edge deployments. This universal orchestration and automation solution enables communications services providers to gain a unified view and full control of security, network addressing scheme, IP keys, privacy, and data governance for full end-to-end control of their network edge infrastructures and workloads. MobiledgeX collaboration with Intel integrates the latest Intel reference architecture and chipset accelerators to optimize performance, decrease latency, and increase security and efficiency of edge networks.

Communication service providers (CSPs) building out 5G initiatives have dramatically accelerated their adoption of public cloud infrastructure, a trend that is likely to continue for the foreseeable future. At the same time, the prospect of massive increases in traffic rates and data volumes, coupled with the need to meet strict latency requirements, are pushing compute tasks out to the network edge, leading to accelerated CSP build-outs of multi-access edge computing (MEC) infrastructure. Distributed workloads at the network edge reduce delay, bandwidth costs, and traffic congestion by reducing backhaul traffic to the network core. They also improve support for latency-sensitive workloads, such as those that involve real-time collaboration and control.

To help realize the competitive potential of these technology shifts, The MobiledgeX Edge-Cloud platform provides a universal orchestration platform and unified control plane to manage and control edge compute workloads across any cloud infrastructure. The MobiledgeX Edge-Cloud platform enables CSPs to create edge clouds from any combination of their own edge infrastructure and public cloud resources, as illustrated in Figure 1.

Figure 1. MobiledgeX management and control of federated edge clouds.

Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries.

Telco Edge Native Service Application Categories

Network

- 4G, 5G, 6G
- Developer consumption experience

Flow

- Pipeline automation
- CV, ML, AI
- Services along the flow

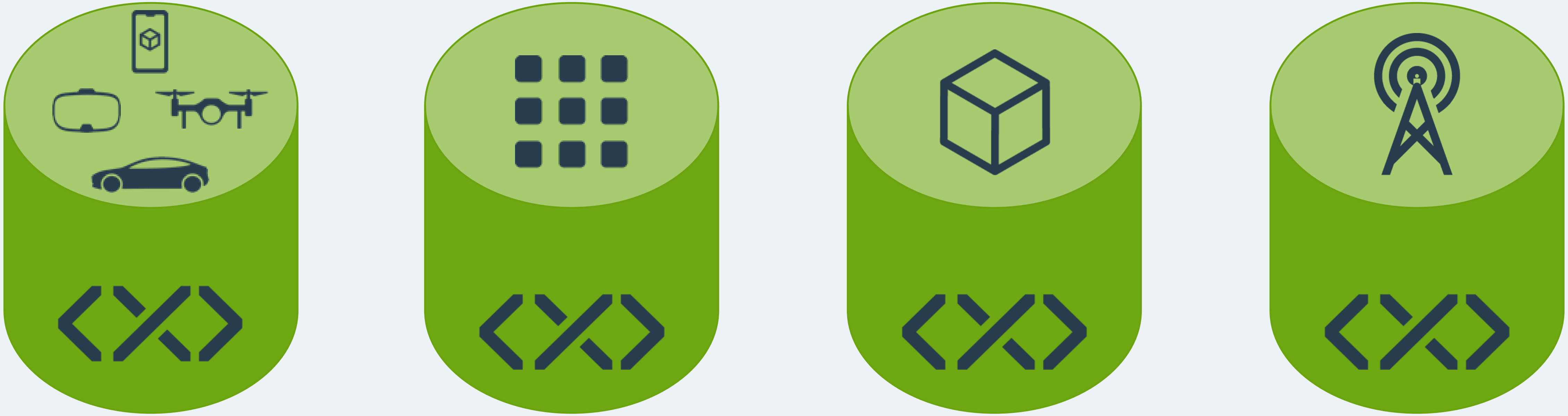
Immersive

- Metaverse
- Pervasive
- Multi-model, multi-device
- Personalization
- Interactive
- Local content
- AR/MR

Autonomous

- M2M
- Devices that move or execute independent actions

Value Proposition



EDGE APPLICATIONS

Easy Deployment and Management for Developer Applications through Common Interface for Edge Enhanced and Edge Native Applications

HYBRID CLOUD ORCHESTRATION

Orchestrate and Manage Application Workloads across infrastructures including OpenStack, VMWare, and Hyperscalers (e.g. Google Anthos)

NETWORK FUNCTIONS

Cloud Native Network Function deployment simplifies operations and management of Network Functions such as UPF

NETWORK EXPOSURE

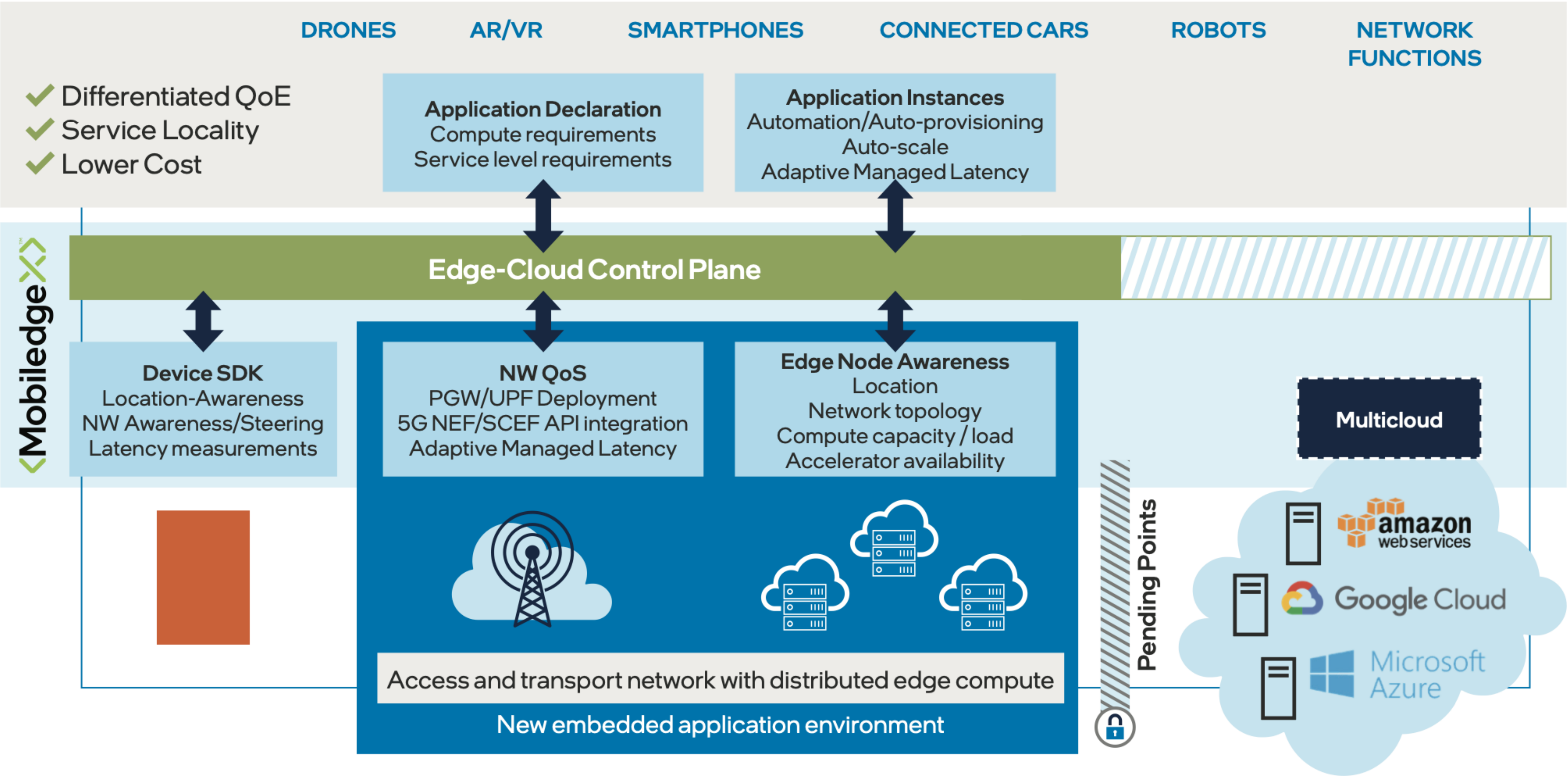
Common API/SDK to expose Network Available Enhancements such as QoS Prioritization and Latency Management Features

Partnership Overview

- The MobiledgeX Edge-Cloud platform provides end-to-end control and management of network edge workloads across multi-cloud edge deployments.
- Our universal orchestration and automation solution enables communications services providers to gain a unified view and full control of security, network addressing scheme, IP keys, privacy, and data governance for full end-to-end control of their network edge infrastructures and workloads.
- MobiledgeX collaborates with Intel to integrate the latest Intel reference architecture and chipset accelerators for optimize performance, decreased latency, and optimum security and efficiency of edge networks.

Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries.

Infrastructure



DRONES AR/VR SMARTPHONES CONNECTED CARS ROBOTS NETWORK FUNCTIONS

- ✓ Differentiated QoE
- ✓ Service Locality
- ✓ Lower Cost

Application Declaration
Compute requirements
Service level requirements

Application Instances
Automation/Auto-provisioning
Auto-scale
Adaptive Managed Latency

Edge-Cloud Control Plane

Device SDK
Location-Awareness
NW Awareness/Steering
Latency measurements

NW QoS
PGW/UPF Deployment
5G NEF/SCEF API integration
Adaptive Managed Latency

Edge Node Awareness
Location
Network topology
Compute capacity / load
Accelerator availability

Access and transport network with distributed edge compute

New embedded application environment

Pending Points

Multicloud

Cross-Operator Federation

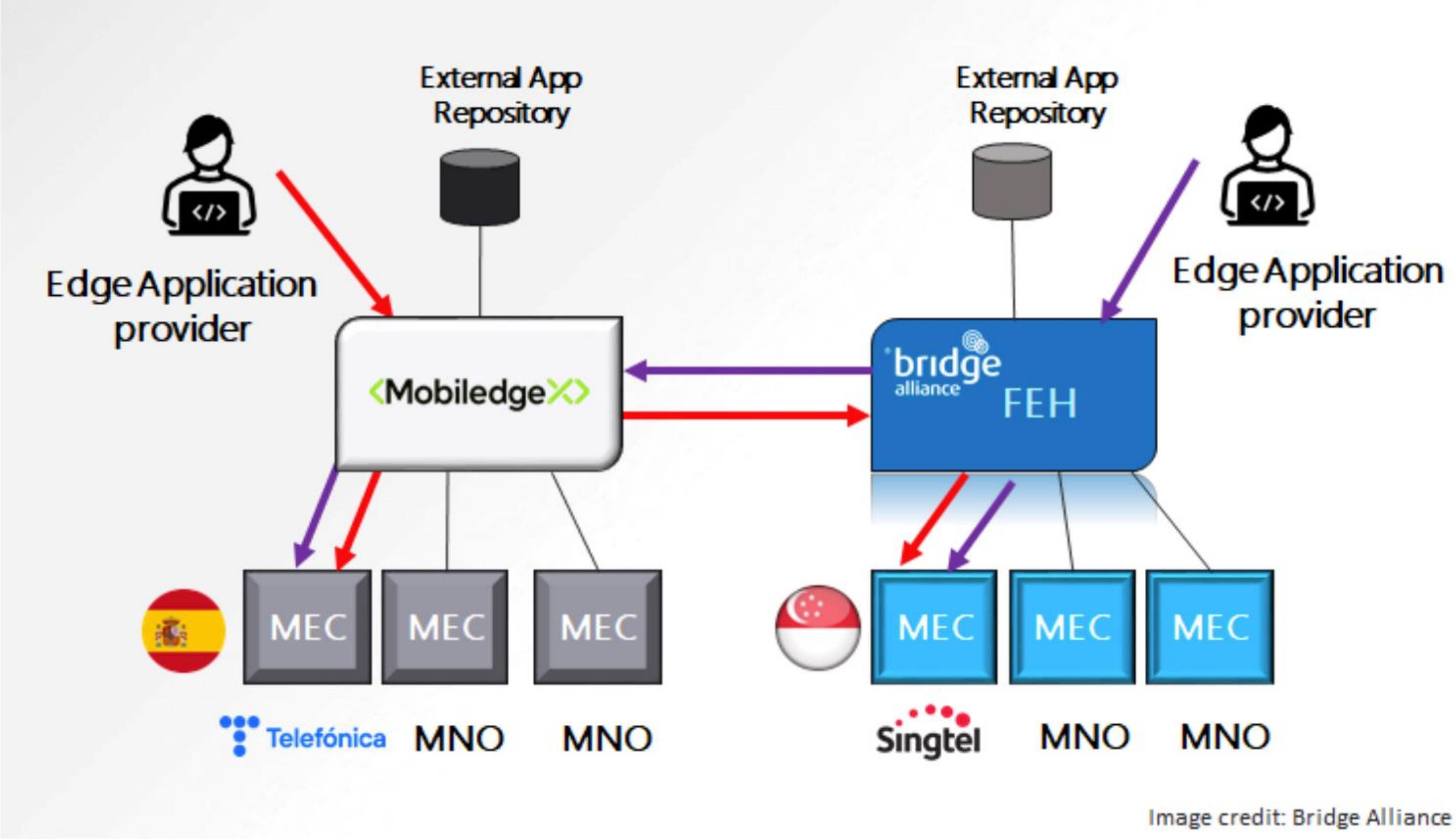


Image credit: Bridge Alliance

https://www.gsma.com/foundry/tec-pre-commercial-trial-edge-compute-service/

Benefits

- ***Flexibility.*** CSPs can deploy federated edge clouds based on their own edge cloud resources, public cloud instances, or any combination, optimizing use of owned infrastructure with unlimited scalability.
- ***Control.*** The MobiledgeX Edge-Cloud solution provides cohesive, unified oversight and control over aggregated edge resources, applications, and workloads, without integration challenges across infrastructure providers.
- ***Profitability.*** New revenue streams can be accelerated to harness and monetize new access to network services available on the network edge. CSPs can now participate across the entire edge computing value chain instead of just providing bandwidth.

Use Case: Foot Traffic Analysis

Problem

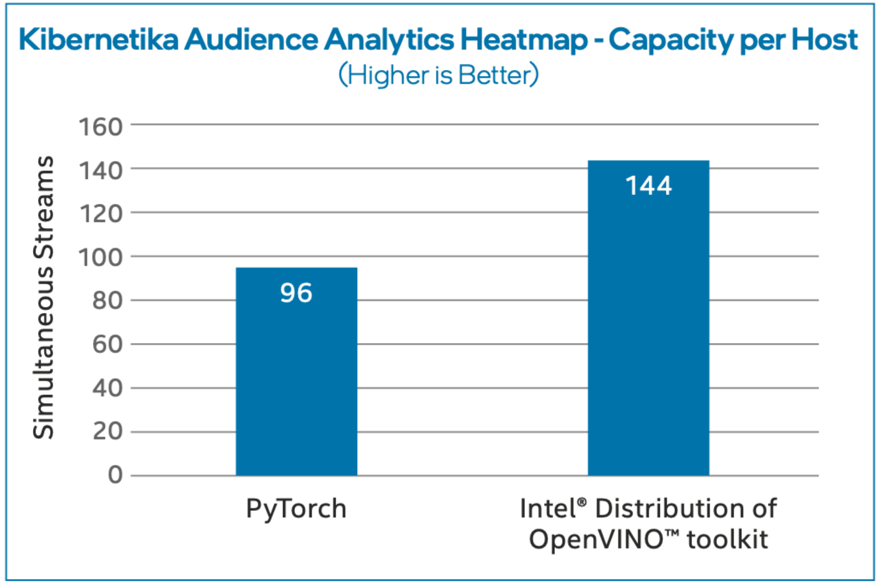
Device offloading consists of shifting compute-intensive workloads such as video analytics from endpoints such as cameras to edge clouds for processing. Data science and AI/ ML solution developer Kibernetika deploys its Heatmap Foot Traffic Analysis solution using MobiledgeX Edge-Cloud on 3rd Generation Intel Xeon Scalable processors. The solution connects to multiple cameras in a retail space and can use any combination of cloud infrastructure: on-prem at the customer location, on CSP-operated edge infrastructure, or on public clouds.

Solution

The Kibernetika solution analyzes customer foot-traffic patterns from the camera video streams, combined with transaction data from point-of-sale devices, to quantify in-store customer behavior in depth. It uses these analytics to generate visualizations use such as heatmaps and track maps that deliver insights to retailers. Store operators can apply that understanding to help guide store layouts and product placement for optimized sales. Analytics from the solution have other retail applications as well, such as managing customer waiting time, detecting shoplifters, and monitoring product levels on shelves.

Scenario: Foot Traffic Analysis

The video analytics required to generate audience analytics heatmaps are particularly compute intensive, and scalability of the solution is vital in retail environments that may include many cameras at each of many sites. Kibernetika optimizes its Machine Teaching technology for Intel architecture by adopting toolkits developed alongside Intel architecture. Intel® Distribution of OpenVINO^{TM} toolkit provides capabilities to build, optimize, and run deep-learning inference models on platforms based on Intel® processors and accelerators. The toolkit enables deep learning inference based on convolutional neural networks (CNNs) at the network edge, accelerating time to market using a library of readily adopted computer functions and pre-optimized kernels.



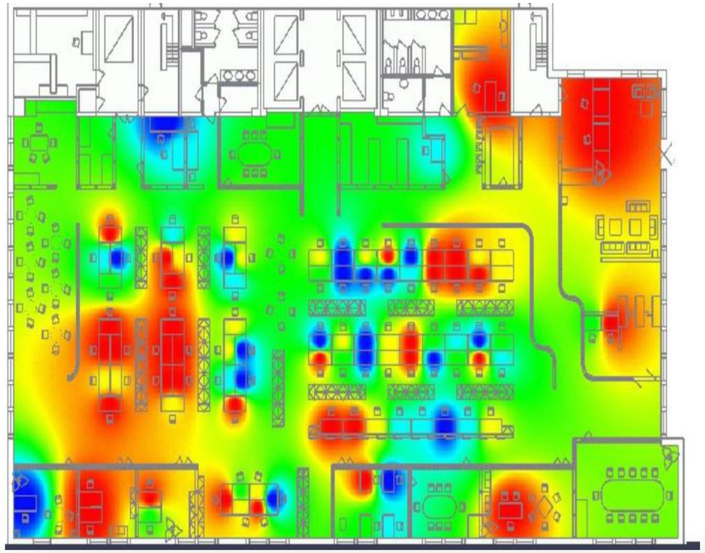
Kibernetika Audience Analytics Heatmap - Capacity per Host

(Higher is Better)

Figure 2. Capacity (simultaneous streams per server) for Kibernetika heatmap generation.[2]

A Kibernetika test team quantified that value, finding that the test server could support 144 camera streams for the audience analytics heatmap workload based on the Intel Distribution of OpenVINO toolkit, compared to 96 streams for the machine-learning workload based on the PyTorch framework, as shown in Figure 2.2

Use Case: Foot Traffic Analysis



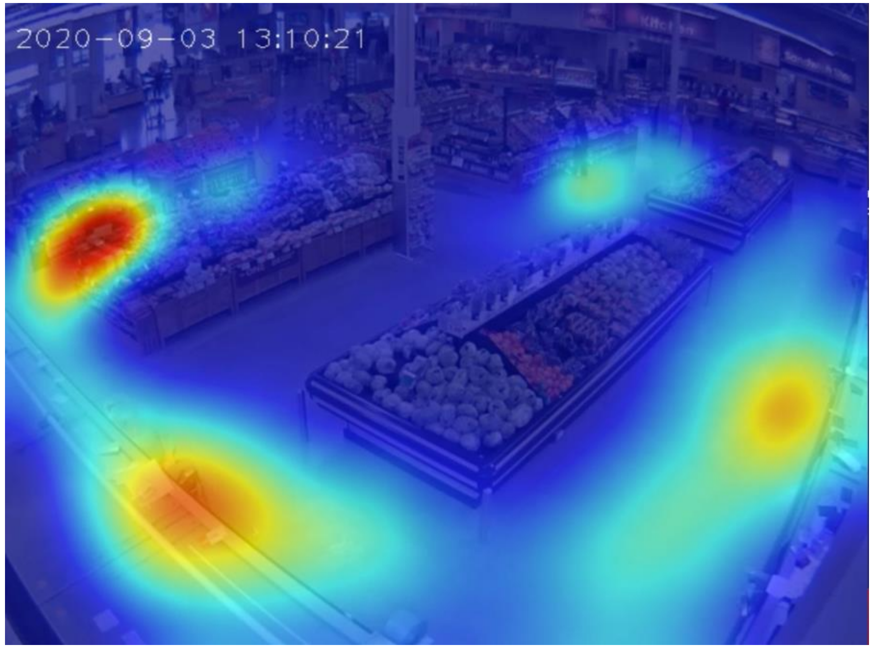
Data Center and Cloud

People Tracking & Heatmap Application
Hosted by MobiledgeX

On-Premise/Outdoor Edge

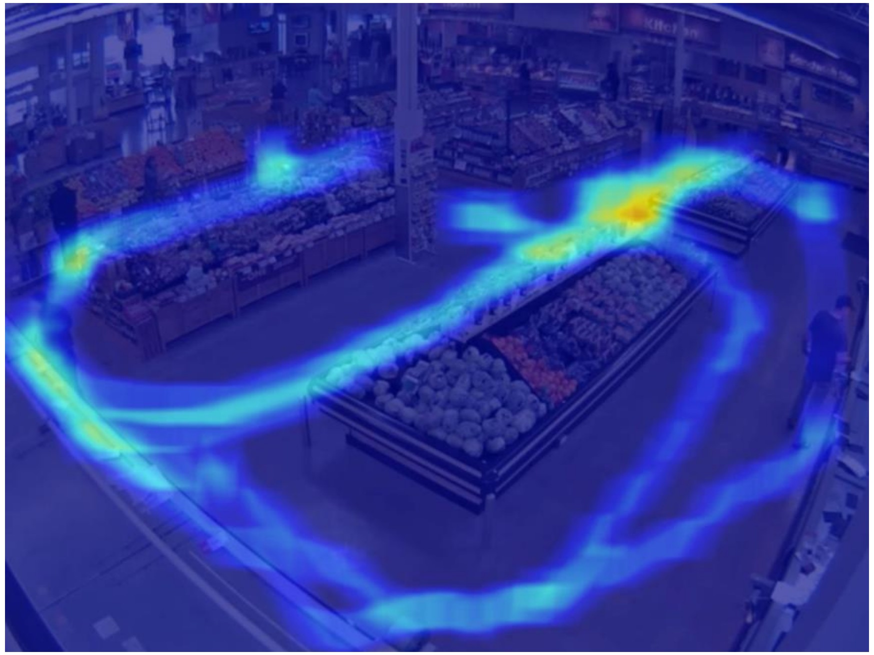
Foot traffic optimization
People tracking and counting

Example 1

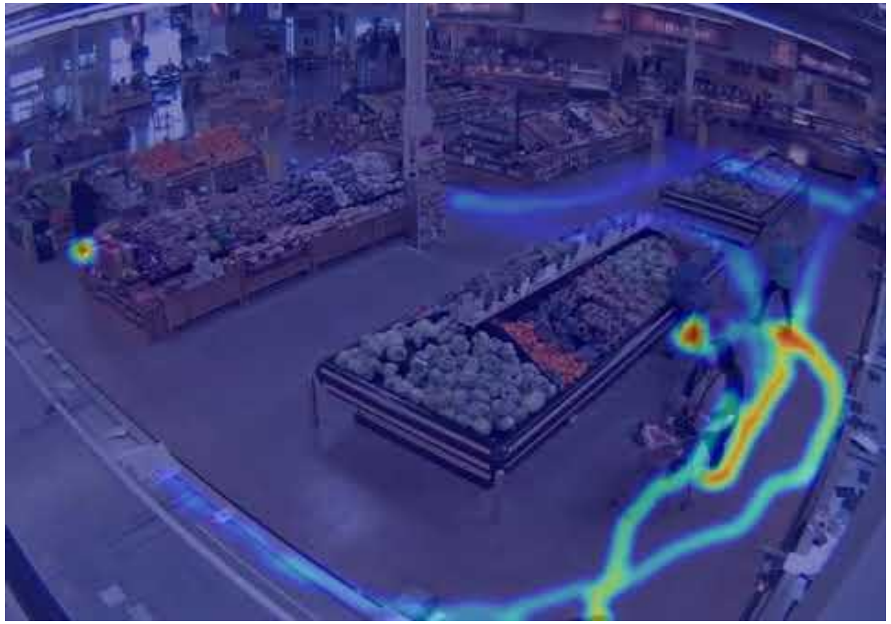


Heatmap

KIBERNETIKA.AI



Trackmap



Use Case: 5G Unified Control Plane

Problem

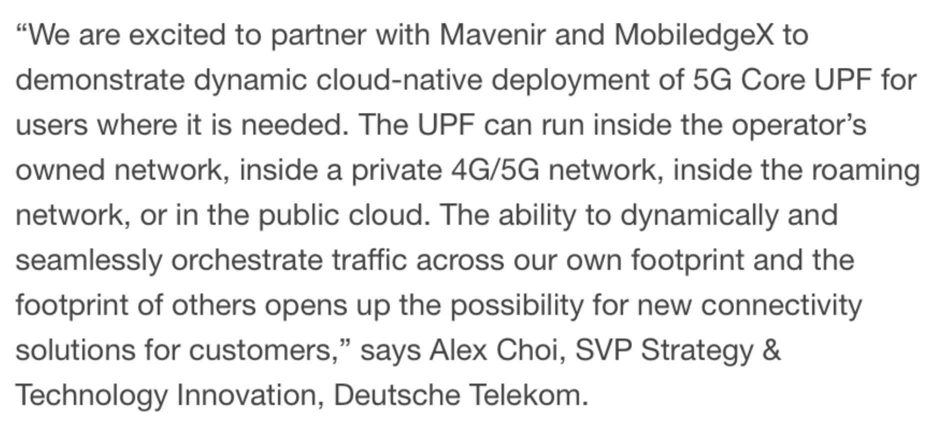
Since the introduction of mobile packet data in 3G, all data traffic, irrespective of source, destination, and requesting application, has been tunneled back to a limited number of fixed (packet gateway) anchor points inside the mobile networks. This is increasingly inefficient, non-performant, and prone to bottlenecks.

Solution

DT, Mavenir and MobiledgeX demonstrated and validated the reference design for deploying cloud-native 5G User Plane Function (UPF) to support Multi-Access Edge Computing (MEC) applications, tailored to each use case need for proximity and performance. The reference design is fully cloud-native, abstracted away from underlying cloud architecture, infrastructure vendor and operating owner.

Scenario: Video Application Traffic

A video application is initiated on a 5G phone in Location A. This request is sent to a 5G core that selects UPF placement in Location A. Tunneling between the two is established. The video request is then passed to the local edge content delivery network for delivery. An identical request is made in Location B. Traffic is then established to Location B and local fulfillment occurs. The MobiledgeX Edge-Cloud 3.0 platform dynamically places UPF workloads on Intel® architecture-based edge infrastructure that is tuned to workload-specific requirements for proximity and performance. The MNO dynamically controls traffic including secure placement, lifecycle management and traffic routing of the UPF and video application at the telco edge.



“We are excited to partner with Mavenir and MobiledgeX to demonstrate dynamic cloud-native deployment of 5G Core UPF for users where it is needed. The UPF can run inside the operator’s owned network, inside a private 4G/5G network, inside the roaming network, or in the public cloud. The ability to dynamically and seamlessly orchestrate traffic across our own footprint and the footprint of others opens up the possibility for new connectivity solutions for customers,” says Alex Choi, SVP Strategy & Technology Innovation, Deutsche Telekom.

Summary

- The ability to define federated edge clouds based on any combination of owned edge assets and public cloud infrastructure is an ongoing monetization opportunity for CSPs.

- The MobiledgeX Edge-Cloud solution is a flexible, universal orchestration platform and control plane for this topology, optimized for performance on Intel architecture.

- This combination allows CSPs to offer edge cloud services with optimized TCO and service levels while retaining full control.

- On-demand edge clouds have potential to deliver new revenue while improving utilization of capital investments.

Q&A

- Thomas Vits, Technical Partnerships, MobiledgeX

[thomas.vits.external@mobiledgex.com]

- Francesc Guim, Principal Architect and Edge Chief, Intel CTO office NEX

[francesc.guim@intel.com]