A Platform for Computing at the Mobile Edge:

Joint solution with HPE, Saguna and AWS

February 2018

Table of Contents

1 Overview ... 3

2 The Business Case for Multi-Access Edge Computing ... 4

2.1 The need for localized “cloud” services ... 4

2.2 The mobile network value proposition ... 4

2.3 The value of a standards-based solution enabling an eco-system of Edge Applications ... 4

3 Solution Overview ... 6

4 The new generation of Edge Applications ... 8

4.1 Reference Architecture: Smart City Surveillance ... 8

4.2 Reference Architecture: AR/VR Edge Application ... 11

4.3 Reference Architecture: Connected Vehicle (V2X) ... 13

5 Summary ... 15

6 Appendix, Technical description ... 17

6.1 Infrastructure layer ... 17

6.2 MEC layer ... 19

6.3 Application Enablement layer ... 21

7 Contributors ... 25

1 Overview

Imagine a world where cars can alert you to dangerous drivers, help avoid collisions, predict traffic patterns, and autonomously drives fleets of vehicles. Or, consider a an Industrial Revolution (Industrial 4.0), where sensors capture instrument large and small machines, reporting data in real-time, creating intelligent automation and orchestration in industries like manufacturing, agriculture, healthcare and logistics.

Envision city and public services that provide intelligent parking, congestion management, pollution detection and mitigation, emergency response and security. At the same time all of this is happening, internet speeds 10 times faster than today, with latencies at 1/100th of current averages, using a seamless combination of mobile, WiFi and fixed access. This vision is far from being science fiction, today's emerging technologies are emerging in both standardization bodies and industry.

This is the world of 5G.

This new generation of applications fuels technological developments and creates new business opportunities for mobile operators. To realize the business gains from the technological achievements of 5G, edge computing is required. The need to act quickly and near to user, edge computing is required to meet the latency requirements of some 5G applications; manage the potentially exorbitant access network load; and support data localization required by other. By providing a cloud-enabled platform for edge computing, mobile operators can take a leading role in the 5G ecosystem, to open up completely new offerings and revenue streams.

Recognizing the business opportunity and technical demand for a scalable, manageable and future-proof edge computing system, this paper brings together three experienced partners to help operators leverage existing mobile network infrastructure, and establish a platform to enable a new wave of revenue-generating applications for 5G.

2 The Business Case for Multi-Access Edge Computing

2.1 The need for localized “cloud” services

Agility, scalability, elasticity and cost efficiency have made cloud computing a platform of choice for application development and delivery. However, much of the data generated by Internet of Things (IoT) devices may never reach the cloud due to privacy requirements, latency constraints, and costs. IoT applications need local cloud services operating close to the connected devices to improve the economics of telemetry data processing. By analyzing telemetry data at the cloud edge, latency gets minimized time-critical applications, and protects sensitive information locally.

2.2 The mobile network value proposition

Mobile networks have expanded to the point where they offer ubiquitous coverage in most countries worldwide. These networks combine wireless access, broadband capacity and security.

2.3 The value of a standards-based solution enabling an eco-system of Edge Applications

Multi-Access Edge Computing (MEC) transforms mobile communication networks into distributed cloud computing platforms that operate at the mobile access network. Located near end-users and connected devices, MEC enables mobile operators to open their networks to new, differentiated services while providing application developers and content providers access to Edge Cloud benefits.

Key to the implementation of MEC is recognition and adherence to a global standard. In fact, ETSI’s MEC Industry Specification Group (ISG) has defined the first set of standardized APIs and Services for MEC. The standard is backed by a wide range of participants from leading mobile operators and industry vendors, including both HPE and Saguna.

In the following sections, we will outline the key benefits provided by MEC:

2.3.1 Extremely low latency

Even with data transmission at the speed of light, applications that require extremely low latency prevent cannot use traditional internet-based cloud to make decisions fast enough. By moving decision-making closer to the user and their connected devices, MEC can drastically reduce the distance data travels to make local clouds a viable alternative.

2.3.2 Broadband delivery

Video content is typically delivered using TCP streams. When network latency is compounded by congestion, the effective bitrate plummets and users experience annoying delays. By moving the data source, in this case video, MEC provides deterministic & low latency, with minimal jitter. As such, it creates a localized broadband highway for streaming at high bitrates.

2.3.3 Economical & Scalable

In "massive" IoT uses cases, many devices such as sensors or cameras send vast amounts of data upstream, which current backhaul networks cannot support. MEC provides a cloud computing environment at the edge of the network, where IoT data can be aggregated & processed locally, thus significantly reducing upstream data. The infrastructure can easily scale-as-you-grow expanding capacity locally or deploying additional edge clouds in new locations.

2.3.4 Privacy & Security

By deploying the MEC Edge Cloud locally, enterprises can ensure that private data stays on premise. However, unlike server-based, on-premise installations, MEC is a fully automated edge cloud environment with centralized management providing.

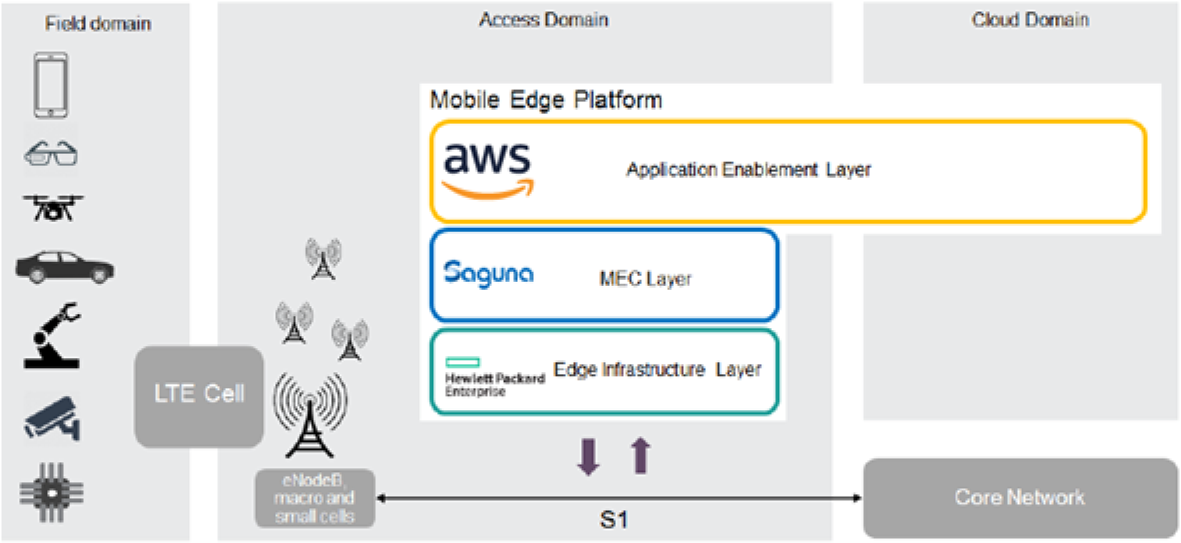
2.3.5 Role of MEC in 5G

Edge Computing is necessary to support the ultra-low latency use cases specified as part of 5G network goals, as it shortens the distance and time between apps and data. MEC is also instrumental in delivering 'faster data' for a more 'connected world' with billions of connected devices, while allowing for cost economization related to transporting enormous volumes of data from user devices and IoT.

It is important to note, that MEC is currently deployed in today's 4G networks. By deploying this standard-based technology in their existing networks, communication service providers can reap the benefits today while creating an evolutionary path to their next generation 5G network.

Solution Overview

To enable next generation apps and services, AWS, Saguna, and HPE have created a platform that enables an applications eco-system at the network edge. The platform enables application developers to easily create new and exciting edge applications using the broad palette of AWS services. It also allows mobile operators to effectively deploy MEC and operate edge applications within the fabric of their mobile network using Saguna's MEC V-RAN software running on HPE's Edgeline hardware.


Field domain
Access Domain
Cloud Domain
Mobile Edge Platform
aws
Application Enablement Layer
Saguna
MEC Layer
Hewlett Packard Enterprise
Edge Infrastructure Layer
LTE Cell
eNodeB, macro and small cells
S1
Core Network

The proposed Mobile Edge Solution consists of three main layers as illustrated above:

Edge Infrastructure Layer

Based on powerful x86 compute platform designed to address specific challenges of mobile network edge environment, Edge Infrastructure Layer provides compute, storage and networking resources at edge locations. It supports wide range of deployment options spanning from RAN base station sites to backhaul aggregation sites to regional branch offices.

MEC Layer

MEC Layer enables placement of an application within a mobile access network, providing services that can include mobile traffic breakout and steering, registration and certification services, and radio network information services. It also provides optional integration point with mobile core network services such as charging and lawful intercept.

Application Enablement Layer

This layer provides the tools and framework to build, deploy and maintain edge-assisted applications. Application Enablement Layer spans from edge locations to the centralized cloud, allowing application placement locally at the edge (e.g. latency-critical or bandwidth-hungry components) while keeping other application functions centralized in the cloud.

The flexible design inherent in the proposed Mobile Edge Platform allows for scaling the edge component of solution, allowing it to fit the needs of concrete use cases. The edge component of the solution can be deployed at the furthest edge of mobile network (e.g., co-located with eNodeB equipment at RAN site), enabling low-latency for bandwidth-demanding application components to be deployed in closest to end devices. The Edge Component can also be deployed at any traffic aggregation point between base station and mobile core, serving traffic from multiple base stations. This flexible placement of edge components is enabled by scalability of underlying components – from infrastructure layer up to MEC and Application enablement layer.

The Mobile Edge Platform provides tools to build, deploy and manage edge-assisted applications:

- Development libraries and frameworks that span from edge to cloud. Includes function-as-a-service at the edge and cloud, AI frameworks for creating and training models in the cloud with seamless deployment and inference at the edge, communication brokerage between edge application services and cloud. These development libraries and frameworks expose well-defined APIs and already have wide adoption in the developer community, smoothing the learning curve and accelerating time-to-market for edge-assisted applications and use cases
- Tools to automate deployment and life-cycle management of edge application component throughout massively distributed edge infrastructure
- Infrastructure services such as virtual infrastructure services at the edge, traffic steering policies at the edge, DNS services, radio awareness services, integration of edge platform into overall NFV framework of mobile operator providing ability to deploy and manage the underlying edge infrastructure at scale
- Diverse compute resources, fitted to particular need of edge application such as CPU, GPU for acceleration of graphics-intensive or AI workloads, FPGA accelerators, cryptographic and data compression accelerators, etc.

This unique combination of functionalities enables rapid development of edge applications, deployment and management of the edge infrastructure and applications at scale, and overall fast time-to-market with edge-enabled use cases.

3 The new generation of Edge Applications

A Mobile Edge Platform enables new application behaviors, revolutionizing traditional patterns. By adding the ability to run certain components and application logic at the mobile network edge, in close proximity to the user devices/clients, the Mobile Edge Platform allows us to re-engineer the functional split between Client and Application Server, enabling new generation of application experiences.

The list below is brief, but not exhaustive, preview to the possibilities in industrial, automotive, public and consumer domains.

- Industrial 4.0
  - Next-Generation Augmented Reality Wearables, Smart Glasses
  - IoT for Automation
  - Predictive Maintenance
  - Asset Tracking
- Automotive
  - Driverless Car
  - Connected Vehicle to Vehicle (V2X)
- Future Cities
  - Surveillance Cameras
  - Smart Parking
  - Emergency Response Management
- Consumer Enhanced Mobile Broadband
  - Next Generation AR/VR & Video Analytics
  - Social Media High-Bandwidth Media Sharing
  - Live Events Streaming at Crowded Venues
  - Gaming

In the sections below, we dive deeper to illustrate how some of the applications can be implemented. Specifically, we illustrate the reference architectures for:

- Smart City Surveillance
- AR/VR at the Edge
- Connected V2X

3.1 Reference Architecture: Smart City Surveillance

Internet of Things technologies has the capability to allow municipalities to increase safety, security, and overall quality of life for residents, while reducing operational costs. One of the most prominent trends is leveraging video for real-time situational analysis, something often called “video-as-a-sensor”. Modern advancements in video recognition technology allow detection of a wide variety of objects (e.g. people, vehicles, people belongings), situational recognition (e.g., traffic jam, fight, trespassing, and abandoned objects), and classify recognized objects (e.g. recognize faces, license plates).

The Mobile Edge Platform offers robust and cost-efficient smart city surveillance systems by enabling:

Efficient video processing at the edge

Computer vision systems require high quality video input (especially for extracting advanced attributes) and hardware acceleration of inference models. The Mobile Edge Platform provides hosting environment at the very edge of a network, offloading backhaul networks and cloud connectivity from bandwidth-hungry high resolution video feeds and allowing low-latency actions based on recognition results (e.g., opening gates for recognized vehicles or people, controlling traffic with adaptive traffic lights). The Mobile Edge Platform provides industry standard GPU resources to accelerate video recognition and to host AI models deployed at the edge

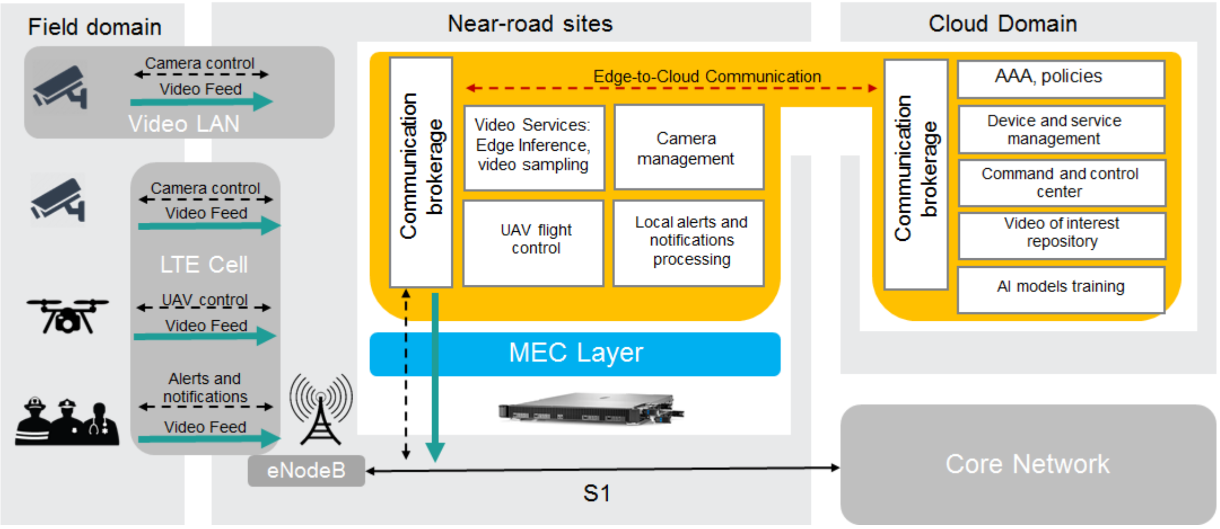
Flexible access network

End-to-end smart city surveillance system might leverage different means to generate video input – such as existing fixed surveillance cameras, mobile wearable cameras (e.g., for law enforcement services or first responders) and drone-mounted mobile surveillance. Video generated from a diverse set of endpoints requires flexibility from access network – fixed video networks, as well as mobile cellular network with native mobility support for wearable or UAV-mounted cameras. Additionally, automated drone-mounted system require low-latency access to control drone flight, which might require end-to-end latencies within millisecond. The Mobile Edge Platform provides means to use robust low-latency cellular access with native mobility support, as well as incorporate existing fixed video networks.

Flexible video recognition models

Video recognition AI models usually require extensive training on sample sets of objects and events, as well as periodic tuning (or development of models for extracting some new attributes). These compute-intensive tasks are usually performed at a centralized cloud level, leveraging highly scalable lower cost cloud compute resources. However, seamless deployment of the trained models at the edge is a complex operational task. The Mobile Edge Platform provides a seamless development and operational experience, starting from creation, training, and tuning of AI models in the cloud, deploying it into edge locations for inference, and ultimately managing the lifecycle of the deployed models.

The diagram below illustrates example of Smart City Surveillance application:

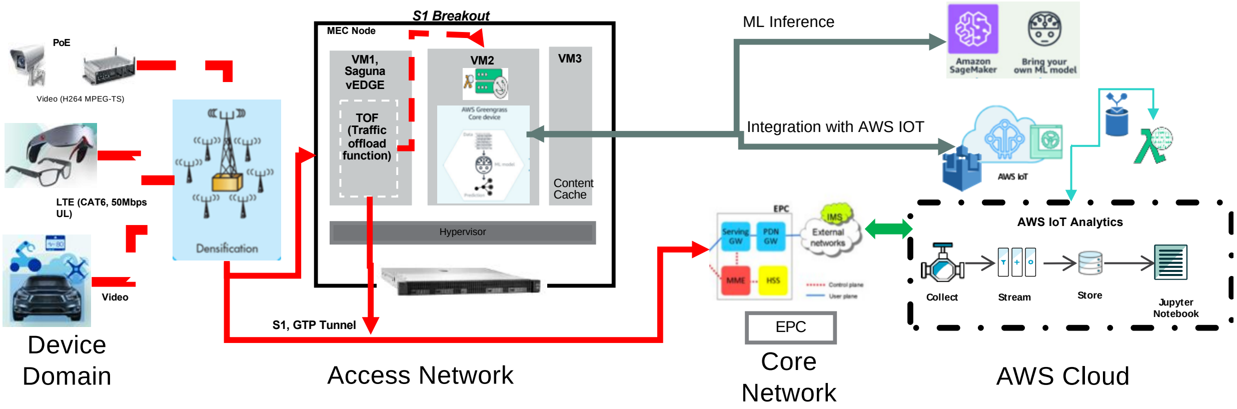


The Future City Surveillance solution consists of three main domains:

- Field domain with diverse ecosystem of devices producing video - e.g. body-worn cameras from first responder units, drones, fixed video surveillance systems, wireless fixed cameras. Video feeds are ingested into the Mobile Edge Platform via cellular connectivity and leveraging existing video networks
- Edge sites located in close proximity to the video-generating devices, hosting latency-sensitive services (UAV flight control, local alerts processing), bandwidth-hungry compute intensive applications (edge inference) and gateway functionalities for video infrastructure control (camera management). Video services extract target attributes from the video streams and share them in format of descriptive metadata with local alerting services and cloud services. Video services at the edge can as well produce low-resolution video proxy or sampling video for transferring video-of-interest to cloud for historical purposes.
- Cloud domain hosts centralized non-latency critical functions, such as device and service management functions, AAA and policies, command and control center functions. It also handles compute-intensive non-latency critical tasks like AI model training.

A more detailed view of the solution is shown below, which also details adds MEC applications with machine learning and inference models applied at the edge via:

- Model training (for surveillance patterns of interest e.g. facial recognition, person counts, dwell time analysis, heat maps, activity detection) via Deep Learning AMIs on the AWS Cloud
- Deployment of trained models to the MEC platform's application-container AWS Greengrass via Amazon Sagemaker
- Application of inference logic (e.g. alerts or alarms based on select pattern detection) via AWS Greengrass ML Inference.



This design approach, based on the Mobile Edge Platform,enables cost efficient way of building and operating Smart City Surveillance system, with edge processing for bandwidth-hungry and latency sensitive services.

3.2 Reference Architecture: AR/VR Edge Application

One of the use-cases benefiting most from a Mobile Edge Platform is Augmented Reality (AR). There are multiple ways in which a Mobile Edge Platform unlocks true potential of AR.

Next generation AR wearables

In its current form, immersive AR experiences require heavy processing at the Client side (e.g. calculating head and eye positions and motion tracking, rendering of high-quality 3D graphics, running video recognition models). Ability to run these computations at the AR device side (head-mounted display, smart glasses, smartphone) in large part define characteristics of those devices - cost, size, weight, battery life and overall aesthetic appeal.

With the compute limitations of today's devices, the heaviest computational tasks need to be offloaded from devices to a remote server or cloud. However, truly immersive AR experiences require catering to human physiological metrics – the AR content and the surrounding physical world might not align if latency is above 10ms of end-to-end latency. To achieve this latency, offload to traditional centralized cloud is required.

Mobile Edge Platform provides powerful compute at the very edge of a network, enabling to offload latency-critical functions from AR device to the network, enabling next generation of lightweight, compact devices with long battery life and native mobility.

Mission-critical operations

AR experiences have already proven their value in workforce enablement applications - with remote collaboration applications, AR assisted maintenance in industrial space, etc.

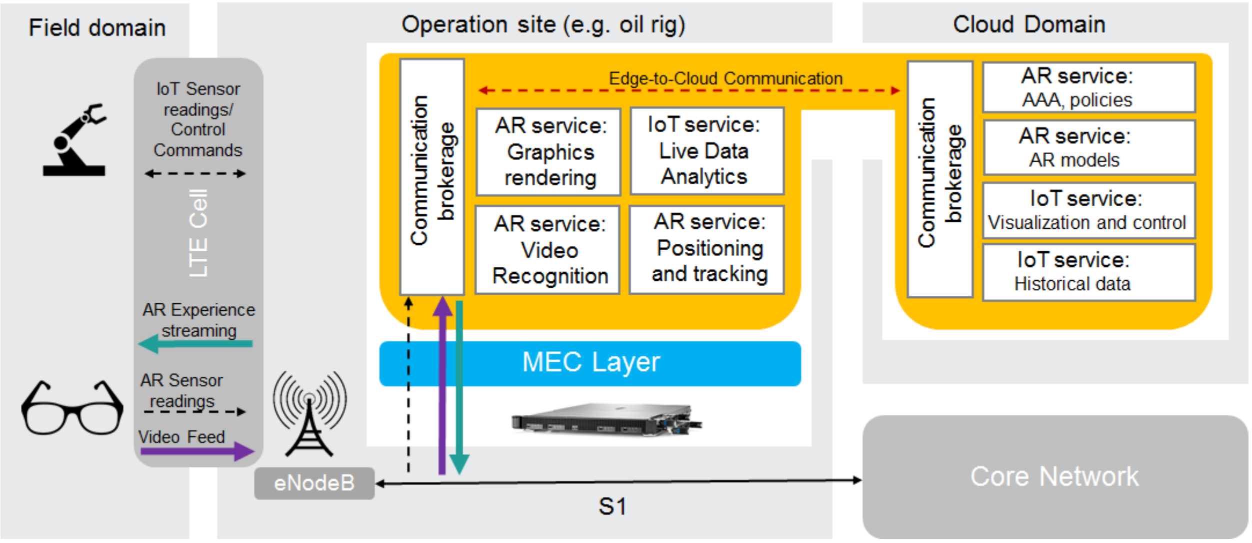
AR experiences become an important part of mission-critical operations, especially in hazardous conditions such as oil extraction sites, refineries, mines or AR-assisted healthcare. Those use cases require high reliability from AR application, even when global connectivity from Client to Server side is degrading or broken. Additionally, the Mobile Edge Platform provides capabilities to operate AR in anoffline mode, with critical components deployed both locally in close proximity to devices and globally in the cloud as a fallback option.

Localized data processing

In many cases, AR experience blends in data coming from relevant local sources (for example, live sensor readings from operational equipment to an AR maintenance application). As the data is critical, it often requires often-scarce bandwidth to ingest large data volumes into the cloud and is governed by data security or privacy frameworks. It is required to provide localized data processing and ingest into AR experience.

Mobile Edge Platform allows ingestion from local data sources into AR experiences, as well as execute commands from the AR experience to the local data sources (e.g. perform equipment maintenance tasks).

The diagram below shows the example architecture for an AR use-case, powered by Mobile Edge Platform.


Field domain
Operation site (e.g. oil rig)
Cloud Domain
IoT Sensor readings/ Control Commands
LTE Cell
AR Experience streaming
AR Sensor readings
Video Feed
eNodeB
Communication brokerage
Edge-to-Cloud Communication
AR service: Graphics rendering
IoT service: Live Data Analytics
AR service: Video Recognition
AR service: Positioning and tracking
MEC Layer
Communication brokerage
AR service: AAA, policies
AR service: AR models
IoT service: Visualization and control
IoT service: Historical data
Core Network
S1

The edge-assisted AR application consists of three main domains:

- Ultra-thin client (e.g. head-mounted display), generating sensor readings of head and eye position, location and other relevant data such as live video feed from embedded cameras
- Edge services, execute latency-critical functions (positioning and tracking, graphics rendering), bandwidth-hungry functions (e.g. computer vision models for video recognition) as well as local data (processing of IoT sensor readings from localized equipment)
- Cloud services, part of AR backend hosted in traditional centralized cloud. These services execute centralized functions (e.g. authentication and policies, command and control center, AR models repository), resource-hungry non latency critical (computer vision model training) and horizontal cross-enterprise functions (e.g. data lakes, integration points with other enterprise systems, etc.)

This design approach offloads heavy computations, making client devices cost-efficient, lightweight and battery-efficient. The devices can be also be operated offline, as local data can be ingested to local systems from external sources and control actions. Offline operation will also save on WAN connectivity costs, and also averts issues with data localization guidelines. Working as an integrated part of the mobile network, this use case natively supports global mobility, telco-grade reliability and security.

3.3 Reference Architecture: Connected Vehicle (V2X)

Connectivity between vehicles, pedestrians, roadside infrastructure and other elements of environment promises tectonic shifts in transportation. However, connectivity is just one piece of the puzzle. Fulfilling the promise of Vehicle to Vehicle communications (V2X) can be only be realized with a new generation of applications, powered by edge computing:

Transportation safety

V2X promises the ability to coordinate actions between vehicles sharing the road (sometimes referred to as a Cooperative Cruise Control). Information exchange between vehicles, including things like intention to change speed or trajectory, can significantly improve safety of autonomous driving. By its nature, car traffic is dynamic, so decisions must be made in near real-time, with decision-making needing to occur within milliseconds. Massively distributed nature of road infrastructure, near real time decision making and requirements for high-speed mobility makes the Mobile Edge Platform perfect fit to host the distributed logic of cooperative driving

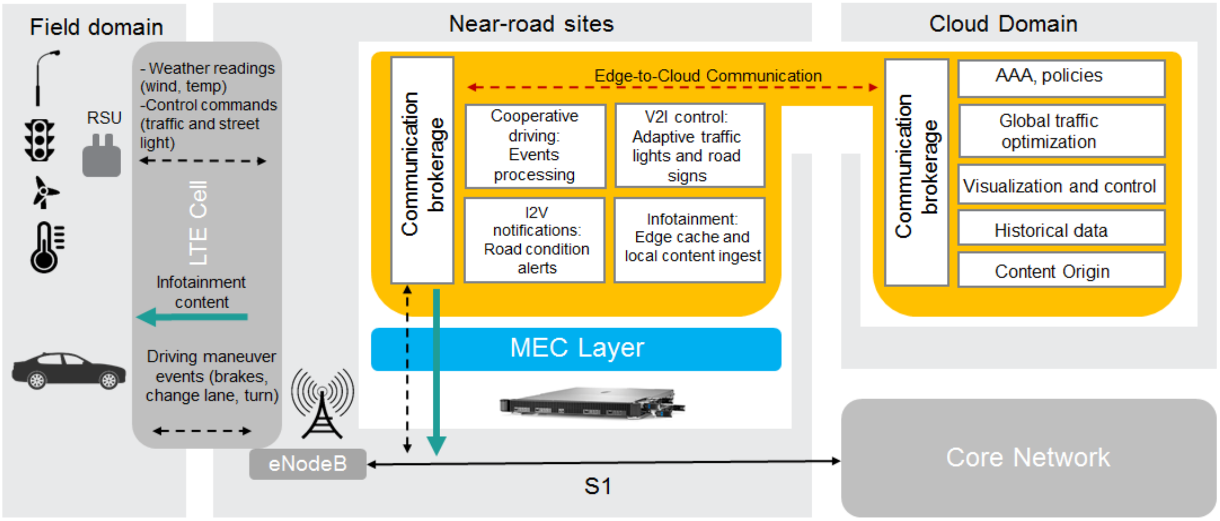
Transportation efficiency

Cooperative driving promises not only increased safety on a road, but also a significant boost in efficiency. With coordinated vehicle maneuvers, overall capacity of road infrastructure can increase without significant road construction. Higher transportation efficiency can be further supported by Vehicles-to-Infrastructure solutions. Vehicles can communicate with roadside equipment for speed guidance, coordinated traffic lights, and parking reservations. While some information can be transmitted via short-range communications (e.g., vehicle to parking lot), coordinated actions of distributed infrastructure (e.g., traffic light changes across multiple intersections) require hosting the logic at the Mobile Edge Platform.

Transportation experience

With autonomous driving technologies controlling vehicle operation, increasing importance is given to car infotainment systems. The Mobile Edge Platform can cache and massively distribute content that is highly localized and context-aware. Localized, context-aware content can guide vehicles and their passengers to nearby refueling stations, restaurants based upon their profile, and attractions, with the ability to provide promotions.

The diagram below shows example of V2X edge-enabled architecture:


Field domain
Near-road sites
Cloud Domain
RSU
- Weather readings (wind, temp)
-Control commands (traffic and street light)
LTE Cell
Infotainment content
Driving maneuver events (brakes, change lane, turn)
eNodeB
Edge-to-Cloud Communication
Communication brokerage
Cooperative driving: Events processing
V2I control: Adaptive traffic lights and road signs
I2V notifications: Road condition alerts
Infotainment: Edge cache and local content ingest
MEC Layer
Communication brokerage
AAA, policies
Global traffic optimization
Visualization and control
Historical data
Content Origin
Core Network
S1

V2X solution consists of three main domains:

- Field domain- Vehicles generate data about intended driving maneuvers (e.g. braking, lane change, turn, acceleration) and also listens for the intentions of surrounding vehicles. The vehicle can also acquire road infrastructure signals from sensors and actuators relevant to driving (wind and temperature sensors along road, street lighting, traffic lights- all of which can be communicated through stationary roadside gateways
- Near-road sites- Located in alongside infrastructure (e.g. respective RAN eNodeB sites), Near-road Sites host latency-sensitive and highly localized services. Examples services include analyzing driving maneuvers, coordinating vehicle-to-vehicle notifications, communicating infrastructure sensor data, and generating commands to road infrastructure (e.g. coordinated switching of traffic lights across several intersections) as well as caching highly localized infotainment content
- Cloud domain- The cloud hosts time-insensitive functions, such as AAA and device policy control, historical data, management function and centralized infotainment content origin.

The V2X design coordinates low-latency data and command exchange between vehicles and infrastructure, while service less urgent and centralized content from the cloud.

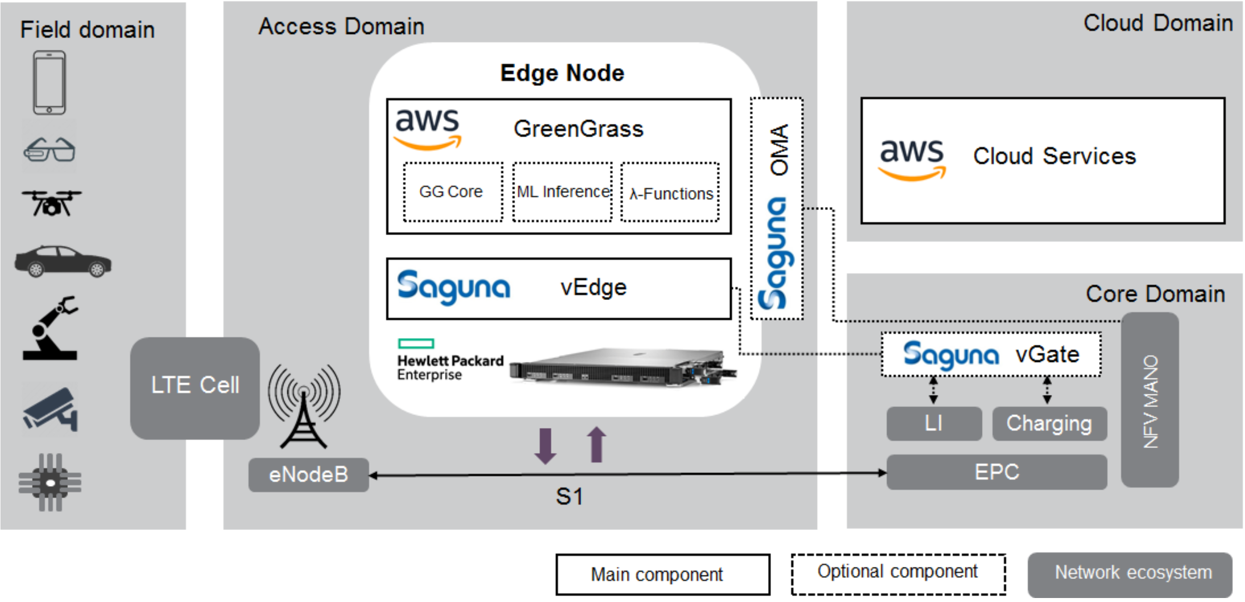
4 Summary

Technological and market developments are creating a new wave of applications that rely upon the instantaneous communication and analysis found in modern mobile networks using edge access infrastructure. While the “killer-app” may not be available, this paper illustrates the reasons to develop an application enablement eco-system and a platform to serve multiple edge use-cases.

5 Appendix, Technical description

This chapter gives a more detailed overview into functional components of the proposed Mobile Edge Platform solution, as well as technical characteristics of each component.

Figure below illustrates functional diagram of the Mobile Edge Platform:



5.1 Infrastructure layer

Physical infrastructure for MEC Node is based on edge-optimized converged platform: Hewlett Packard Enterprise EdgeLine EL4000.

This MEC solution can place workloads within any segment of mobile access network - for example at a RAN site, backhaul aggregation hub, CRAN hub. Since traditional datacenter equipment cannot operate in roadside conditions, the industrialized EdgeLine EL4000 is ideal for this MEC solution:

Compute Density

EdgeLine EL4000 hosts up to four hot-swappable compute cartridges in 1U chassis, providing up to 64 Intel™ Xeon-D cores with optimized price/core and watt/core characteristics. That design provides up to 3 times higher compute density than a typical datacenter components, while additionally reducing power consumption. The power consumption and environmental characteristics allow an operator to place an EdgeLine EL4000 in a MEC node at the deepest edge of access network down to a RAN site, where space and power constraints make other general purpose compute platforms infeasible.

Workload-specific compute

Diversity of MEC use cases requires different types of compute resources. The EdgeLine EL4000 platform provides diverse compute and hardware acceleration capabilities, allowing to co-locate workloads in the same chassis with different compute needs:

- General Workloads- Intel™ x86 processors serve general workloads. Typical examples include Virtual Network Functions, virtualized edge application enablement platform, control applications requiring fast control actions
- Built-in GPU- Built-in GPU accelerate graphics processing. Applications benefitting from a built-in GPU is video transcoding at the edge for MEC-assisted content distribution, and rendering 3D graphics for AR/VR streaming application
- Dedicated GPU- Dedicated GPU cards, accelerate deep learning algorithms. Enabled by strategic partnership with NVIDIA, EdgeLine platform can be equipped with hardware acceleration for machine learning inference at the edge. Video analytics and computer vision can greatly benefit from deep learning hardware acceleration. Additionally, machine learning inference at the edge for anomaly detection and predictive maintenance should also use a dedicated GPU
- Built-in cryptographic acceleration- Intel™ QuickAssist Technology accelerates cryptographic or compressed data
- PCIe extension slots- Up to four PCI cards can be placed in a single chassis. Options include specialized plug-in units such as dedicated FPGA boards, neuromorphic chips, etc. The importance of such specialized hardware acceleration is being actively evaluated for many network function workloads (such as RAN baseband processing) as well as applications (efficient deep learning inference).

Physical and operational characteristics

MEC Nodes should be capable of operating in sites traditionally used for hosting telco infrastructure (e.g., radio base station equipment at RAN sites, access routers at traffic hubs, etc.). The operating environments differ from traditional datacenters, limited by space, climate control, and limited physical accessibility. EdgeLine EL4000 is optimized to operate in such environments, with operational characteristics comparable to the telco purpose-built appliances:

Parameter	RAN Baseband appliance	Typical datacenter platform	EdgeLine EL4000
Operating temperature	+0 to+50°C	+10 to +35°C	0 to +55°C
Non-destructive shock tolerance	30 G	2 G	30 G
MTBF, expected	30-35 years	10-15 years	>35 years

On top of enhanced operational characteristics, EdgeLine EL4000 exposes open iLO interface for the management of highly distributed infrastructure of MEC Nodes. The iLO interface is compliant with RedFish industry standard, exposing infrastructure management functions via a simple RESTful service.

5.2 MEC layer

The MEC Platform layer is based on Saguna Open-RAN solution and consists of following functions:

- Saguna vEdge function, located within MEC Node
- Optionally, Saguna vGate function, located at the core network site
- Optionally, Saguna OMA function, located within MEC Node or aggregation point of several MEC Nodes

5.2.1 Saguna Open-RAN components overview

Saguna vEdge resides in the MEC Node and enables services and applications to operate inside the mobile RAN by providing MEC services such as registration and certification, Traffic Offload Function (TOF), real-time Radio Network Information Services (RNIS), optional DNS services. The virtualized software node is deployed in the RAN on a server at a RAN site or aggregation point of mobile backhaul traffic. It may serve single or multiple eNodeB base stations and small-cells and can easily be extended to support WiFi and other communications standards in heterogeneous network (HetNet) deployments.

Saguna vEdge taps the S1 interface (GTP-U and S1-AP protocols) and steers the traffic to appropriate local or remote end point based on configured policies. Saguna vEdge implements local LTE traffic steering in number of modes (inline steering, breakout, tap).

It has a communication link connecting it to the optional Saguna vGate node using Saguna's OPTP (Open RAN Transport Protocol). It exposes open REST APIs for managing the platform as well as providing platform services to the MEC-assisted applications.

Saguna vGate is an optional component which resides in the core network. It is responsible for preserving core functionality for RAN-generated traffic: Lawful interception (LI), charging and policy control. The Saguna vGate also enables mobility support for session generated by an MEC-assisted application.
Operating in a Virtual Machine, Saguna vGate is adjacent to the enhanced packet core (EPC). It has a communication link connecting it to the Saguna vEdge nodes using Saguna's OPTP (Open RAN Transport Protocol), and Mobile network integrations for LI and charging functions.

Saguna OMA (Open Management & Automation) is an optional subsystem, residing in the MEC Node or at aggregation point of several MEC Nodes. It provides a management layer for the MEC Nodes and integrates it into cloud NFV environment (NFVO, VIM and OSS systems).

Saguna OMA provides two management modules: The Saguna Open-RAN Mobile Edge Platform Manager (MEPM) and Virtualized Network Function Manager (VNFM).
The VNFM provides Life-Cycle-Management and monitoring for MEC Platform (Saguna vEdge) and MEC-assisted applications. This is a standard layer of management required within NFV environments. It resides at the edge to manage the local MEC environment.

The MEPM provides an additional layer of management required for operating and prioritizing MEC applications. It is responsible for managing the rules and requirements presented by each MEC application rules and resolving conflicts between different MEC-assisted applications.

The OMA (Open Management & Automation) node operates on a virtual machine (VM), manages on-boarded MEC-assisted application via its workflow engine using Saguna and 3rd party plugins. The Saguna OMA is managed via REST API.

Saguna Open-RAN services

As a MEC Platform layer, Saguna Open-RAN provides the following services:

Mobile Network Integration Services

- *Mobility* with Internal Handover support for mobility events between cells connected to the same MEC Node and External Handover support between two or more MEC Nodes as well as from cells connected to MEC Node to cells which are not connected to any MEC Node
- *Lawful Interception (LI)* for RAN-based generated data. It supports X1 (Admin), X2 (IRI) and X3 (CC) interfaces and is pre-integrated with Utimaco and Verint LI systems
- *Charging* support using CDR generation for application-based charging (based on 3GPP TDF-CDR) and charging triggering based on time, session and data. Supported charging methods are File based (ASN.1) and GTP'

- *Management*, vEdge REST API for MEC services discovery and registration, ME Platform Manager (MEPM) and Virtualized Network Function Manager (VNFM) allow to efficiently operate MEC solution and integrate it into existing NFV environment

Edge Services

- *Registration* for MEC-assisted applications. The MEC Registration service provides dynamic registration and certification of MEC applications, registration to other MEC services provided by the MEC Platform, setting the MEC application type
- *Traffic Offload Function* routes specific traffic flows to the relevant applications as configured by the user. The TOF also handles tunneling protocols such as GPRS Tunneling Protocol (GTP) for Long Term Evolution (LTE) network, Standard A10/A11 interfaces for 3GPP2 CDMA Network and handles plain IP traffic for WiFi/DSL Network
- *DNS* provides DNS caching service by storing recent DNS addresses locally to accelerate the mobile Internet and DNS server functionality, preconfiguring specific DNS responses for specific domains. This enables causing the UE to connect to a local application for specific TCP sessions
- *Radio Network Information Service*, provided per Cell and per Radio Access Bearer (RAB). The service is vendor-independent and can support eNodeBs from multiple RAN vendors simultaneously. It supports standard ETSI queries (e.g. cell info) and notifications mechanism (e.g. RAB establishment events). Additional information based on Saguna proprietary model provides real-time feedback on cell congestion level and RAB available throughput using statistical analysis
- *Instant Messaging* with Short Message Service (SMS) provided as a REST API request. It offers smart messaging capabilities, for example sending SMS to UEs on specific area (e.g. sports stadium) or sending SMS to UE when entering / exiting specific area (e.g. shop)

Mobile Edge Applications

- *Throughput guidance application* is using the internal RNIS algorithm to deliver throughput guidance for specific IP addresses of the server side or according to domain names of the servers. The application can be configured with the period of such Throughput Guidance update per target.

5.3 Application Enablement layer

Application Enablement layer consists of Amazon Web Services Greengrass hosted at the MEC Node side, working in conjunction with Amazon Web Services cloud.
AWS Greengrass is designed to support IoT solutions that connect different types of devices with the cloud and each other as well as to run local functions and parts of applications at the network edge. Devices that run Linux and support ARM or x86 architectures can host the

Greengrass Core. The Greengrass Core enables the local execution of AWS Lambda code, messaging, data caching, and security.

Devices running AWS Greengrass Core act as a hub that can communicate with other devices that have the AWS IoT Device SDK installed, such as micro-controller based devices or large appliances. These AWS Greengrass Core devices and the AWS IoT Device SDK-enabled devices can be configured to communicate with one another in a *Greengrass Group*. If the Greengrass Core device loses connection to the cloud, devices in the Greengrass Group can continue to communicate with each other over the local network. A Greengrass Group represents localized assembly of devices, for example it may represent one floor of a building, one truck, or one home.

Greengrass builds on AWS IoT and AWS Lambda, and can also access other AWS services. It is built for offline operation and greatly simplifies the implementation of local processing. Code running in the field can collect, filter, and aggregate freshly collected data and then push it up to the cloud for long-term storage and further aggregation. Further, code running in the field can also take action very quickly, even in cases where connectivity to the cloud is temporarily unavailable.

Greengrass has two constituent parts, the Greengrass Core (GGC) and the IoT Device SDK. Both of these components run at on-premises hardware, out in the field.

Greengrass Core is designed to run on devices that have at least 128 MB of memory and an x86 or ARM CPU running at 1 GHz or better, and can take advantage of additional resources if available. It runs Lambda functions locally, interacts with the AWS Cloud, manages security & authentication, and communicates with the other devices under its purview.

The IoT Device SDK is used to build the applications on devices connected to the Greengrass Core device (generally via a LAN or other local connection). These applications will capture data from sensors, subscribe to MQTT topics, and use AWS IoT device shadows to store and retrieve state information.

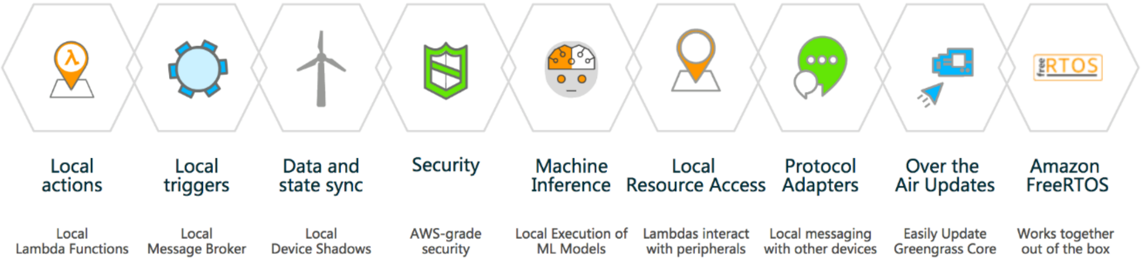
AWS Greengrass Features include:

- Local Support for AWS Lambda: AWS Greengrass includes support for AWS Lambda and AWS IoT Device Shadows. With Greengrass you can run AWS Lambda functions right on the device to execute code quickly
- Local Support for AWS IoT Device Shadows: AWS Greengrass also includes the functionality of AWS IoT Device Shadows. The Device Shadow caches the state of your device, like a virtual version, or “shadow,” of each device that tracks the device’s current versus desired state
- Local Messaging and protocol adapters: AWS Greengrass enables messaging between devices on a local network, so they can communicate with each other even when there is no connection to AWS. With Greengrass devices can process messages and deliver them to another device or to AWS IoT based on business rules user defines. Devices that communicate via the popular industrial protocol, OPC-UA, are supported by the

AWS Greengrass protocol adapter framework and the out-of-the-box OPC-UA protocol module. On top of that AWS Greengrass provides protocol adapter framework to implement support for custom, legacy, and proprietary protocols.

- Local Resource Access: AWS Lambda functions deployed on an AWS Greengrass Core can access local resources that are attached to the device. This allows to use serial ports, USB peripherals such as add-on security devices, sensors and actuators, on-board GPUs, or the local file system to quickly access and process local data
- Local Machine Learning Inference: allows to run locally a machine learning model, built and trained in the cloud. With hardware acceleration available in MEC infrastructure layer, that feature provides powerful mechanism for solving any machine learning task at the local edge, e.g. discovering patterns in data, building computer vision systems, running anomaly detection and predictive maintenance algorithms.

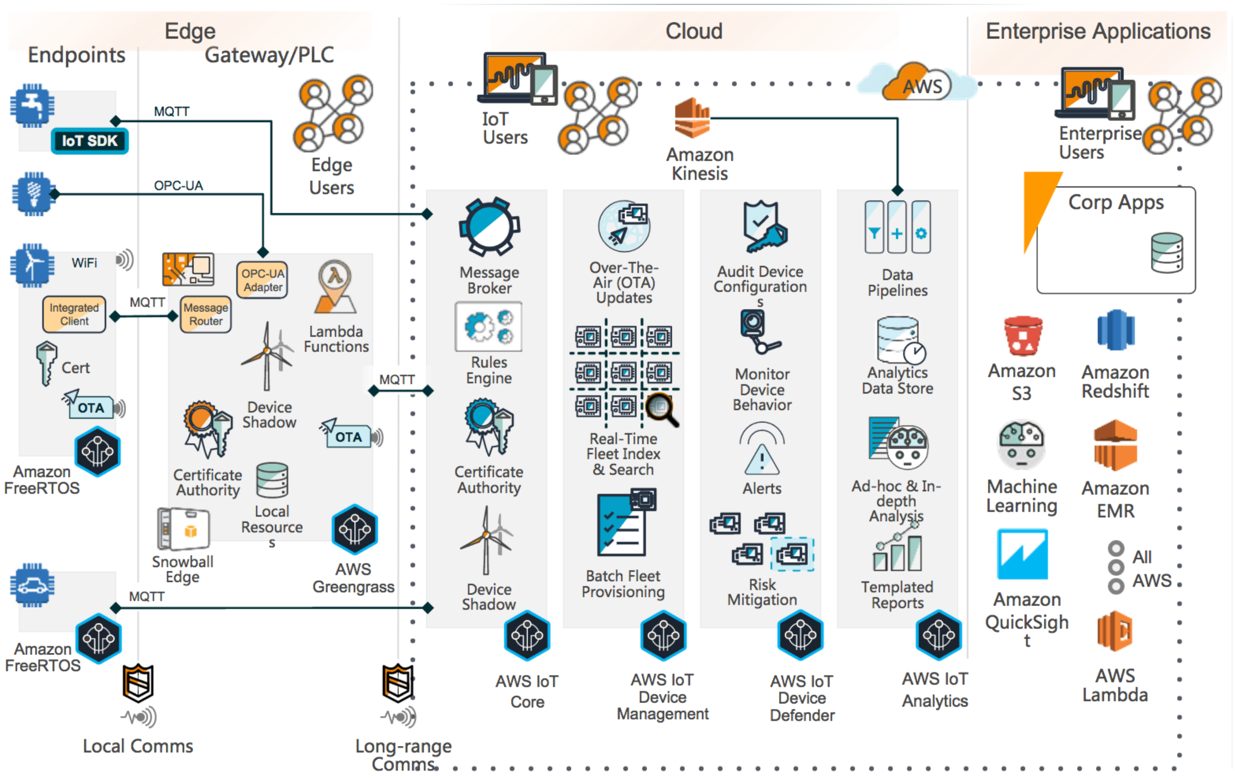
AWS Greengrass has a growing list of features with those currently available illustrated below.



AWS Greengrass on the MEC node acts as a pivot point, integrating the MEC platform with the AWS IoT solution and growing suite of AWS Cloud Services, providing a powerful application enablement environment for developing, deploying and managing MEC-assisted applications at scale.

The figure below illustrates the current portfolio of AWS services enabling a seamless IoT pipeline: from end-points connecting via Amazon FreeRTOS or the IoT SDK through MQTT or OPC-UA, to edge gateways that host AWS Greengrass and Lambda functions providing data-processing capabilities at the edge, up to cloud-hosted AWS IoT Core, AWS Device Management, AWS Device Defender and AWS IoT Analytics services, as well as enterprise applications.

IoT with AWS



aws

Hewlett Packard
Enterprise

Saguna

6 Contributors

Amazon Web Services (aws.amazon.com):

- Shoma Chakravarty, WW Technical Leader, Telecom
- Tim Mattison, Partner Solution Architect

HPE (hpe.com/info/edgeline)

- Alex Reznik, Enterprise Solution Architect and ETSI MEC Chair
- Rodion Naurzalin, Lead Architect, IoT and Converged Edge Systems

Saguna (saguna.net)

- Tally Netzer, Marketing Leader
- Danny Frydman, CTO