White Paper

Telecom Operators and Cloud Service Providers
5G + AI Edge Computing

Creating a 5G + AI Edge Middle Platform with OpenNESS and Baidu IME

This white paper explores the successful integration of Open Network Edge Services Software (OpenNESS), an open source initiative from Intel with Baidu's Intelligent Mobile Edge (IME), to create a 5G + AI Edge Middle Platform, which provides integrated 5G + AI capabilities to industries and developers. It aims to explore new and innovative ways to empower businesses of all sizes, enabling 5G + AI edge computing capabilities to be quickly deployed in real-world applications, enabling 5G + AI edge computing capabilities to be quickly deployed in real-world applications.

Table of Contents

Introduction: 5G Leads to New Opportunities at the Edge.....1

About OpenNESS.....1

MEP Related Microservices.....2

5G Edge Application Enhancement and Optimization Components.....2

About Baidu Intelligent Mobile Edge (IME).....3

Lightweight Edge Computing PaaS Layer.....3

Service Capability Layer.....4

Service Capability Exposure Layer....4

Combining the Strengths of OpenNESS and IME to Create 5G + AI Edge Middle Platform.....5

Combining the Strengths of OpenNESS and IME.....5

Overview of the 5G + AI Edge Middle Platform.....5

Key Operation Flow of the 5G + AI Edge Middle Platform.....6

Typical Use Cases of the 5G + AI Edge Middle Platform.....7

V2X Car Road Cooperation.....7

5G Video Streaming Acceleration.....7

Forward Looking: Continuous Integration of AI, 5G and Edge Innovations.....8

Introduction: 5G Leads to New Opportunities at the Edge

The number of edge devices and the volume of data are both exploding exponentially, and various AI-enabled use cases are also emerging in areas like manufacturing. AI edge computing is rapidly becoming a key pillar of technological development, both now and for the future. Meanwhile, the rapid deployment of 5G is propelling edge computing to new heights. The traditional "device-cloud" infrastructure is rapidly evolving into a "device-edge-cloud collaboration" infrastructure, with integrated 5G and AI capabilities.

As a powerful complement to cloud computing capabilities, edge computing has proven to be effective in helping businesses act faster on data at its source, minimizing the network latency bottleneck that often occurs in traditional centralized computing models. Industries can take advantage of edge computing to achieve real-time responses, and draw new insights from data that can improve business intelligence, security, and privacy. In addition, the "cloud-edge collaboration" model allows new and innovative use cases that utilize both 5G and AI capabilities by offering the flexibility to deploy different levels of computing power at the edge. In September 2018, Intel and Baidu established a joint 5G + AI Mobile Edge Computing (MEC) Innovation Lab to accelerate the research and development of edge computing technology. The Lab also aims to explore new and innovative ways to empower businesses of all sizes, enabling 5G + AI edge computing capabilities to be quickly deployed in real-world applications.

This white paper explores the successful integration of Open Network Edge Services Software (OpenNESS), an open source initiative from Intel with Baidu's Intelligent Mobile Edge (IME) to create a 5G + AI Edge Middle Platform, which provides integrated 5G + AI capabilities to industries and developers. Middle Platform is a concept adopted by technology giants/leaders in China. It's an enabling platform, which organizes core capabilities, maximizes synergies, and provides integrated business capabilities to internal business units, as well as third-parties.

About OpenNESS

OpenNESS is a MEC software toolkit that enables edge platforms to on-board and manage applications and network functions with cloud-like agility across any type of network. This open source distribution is designed to foster open collaboration and application innovation at the edge of the network and also on-premises, making it easier for cloud and Internet of Things (IoT) developers to engage with a worldwide ecosystem of hardware, software and solutions integrators to develop solutions for 5G and Edge.

OpenNESS is based on the specifications of Multi-Access Edge Computing (MEC) created by the European Telecommunications Standards Institute (ETSI), as well as 3GPP 5G standards. It contains two parts: an Edge Node and a Kubernetes-based Controller. Built on a cloud native, microservices-based architecture, it provides developers modular capabilities and can be adopted in parts to fit a larger solution or as a complete platform.

OpenNESS offers Mobile Edge Platform Management (MEPM), Mobile Edge Platform (MEP) and Data Plane based on ETSI reference architecture. It enables developers to manage and orchestrate edge services at various locations (across the Network Edge and on-premises Edge), through 4G/5G GTP/IP), Wireline (IP) and Wi-Fi (IP) access network technologies, while also handling 3GPP network related services. In addition to MEP related microservices, OpenNESS also offers 5G edge application enhancement and optimization components.

MEP Related Microservices

- Edge Application Agent (EAA): Provides edge service discovery, subscription and notification functions, similar to ETSI Mp1 interface.
- Domain Name System (DNS) service: Supports ETSI MEP DNS function requirements and edge application DNS resolution and forwarding.
- Edge Interface Service (EIS): Attaches additional network interfaces to edge hosts in order to provide external communication for applications. The interfaces can be configured and managed through the Controller.
- MEP enhancement service: Includes video transcoding, host BIOS firmware remote upgrades, and others.

5G Edge Application Enhancement and Optimization Components

- 5G network functions: Includes Core Network Configuration Agent (CNCA), Application Function (AF) and 5G Operation and Management (OAM).
- Enhanced platform awareness (EPA): Provides supports for huge pages, Node Function Discovery (NFD), Intel Resource Management Daemon (RMD), Topology Manager and more.
- Accelerator card plug-in: Supports Intel's High Density Deep Learning (HDDL-R) accelerator card, Intel Visual Cloud Accelerator Card – Analytics (VCAC-A), and Intel FPGA Programmable Acceleration Card N3000.
- Container Network: Provides a flexible and high-performance network interface for Linux containers. Several open source CNIs are supported including SR-IOV, Userspace, Kube-OVN, Calico/eBPF, Weave, Bond and Multus.
- Telemetry: A family of services that collectively achieves monitoring and aggregation of platform metrics such as CPU utilization, temperature, power consumption, etc. Supported services include Prometheus/Grafana, Node Exporter, collectd, OpenTelemetry, and Telemetry Aware Scheduler (TAS). TAS enables intelligent scheduling/de-scheduling decisions based on platform metrics and alerts.

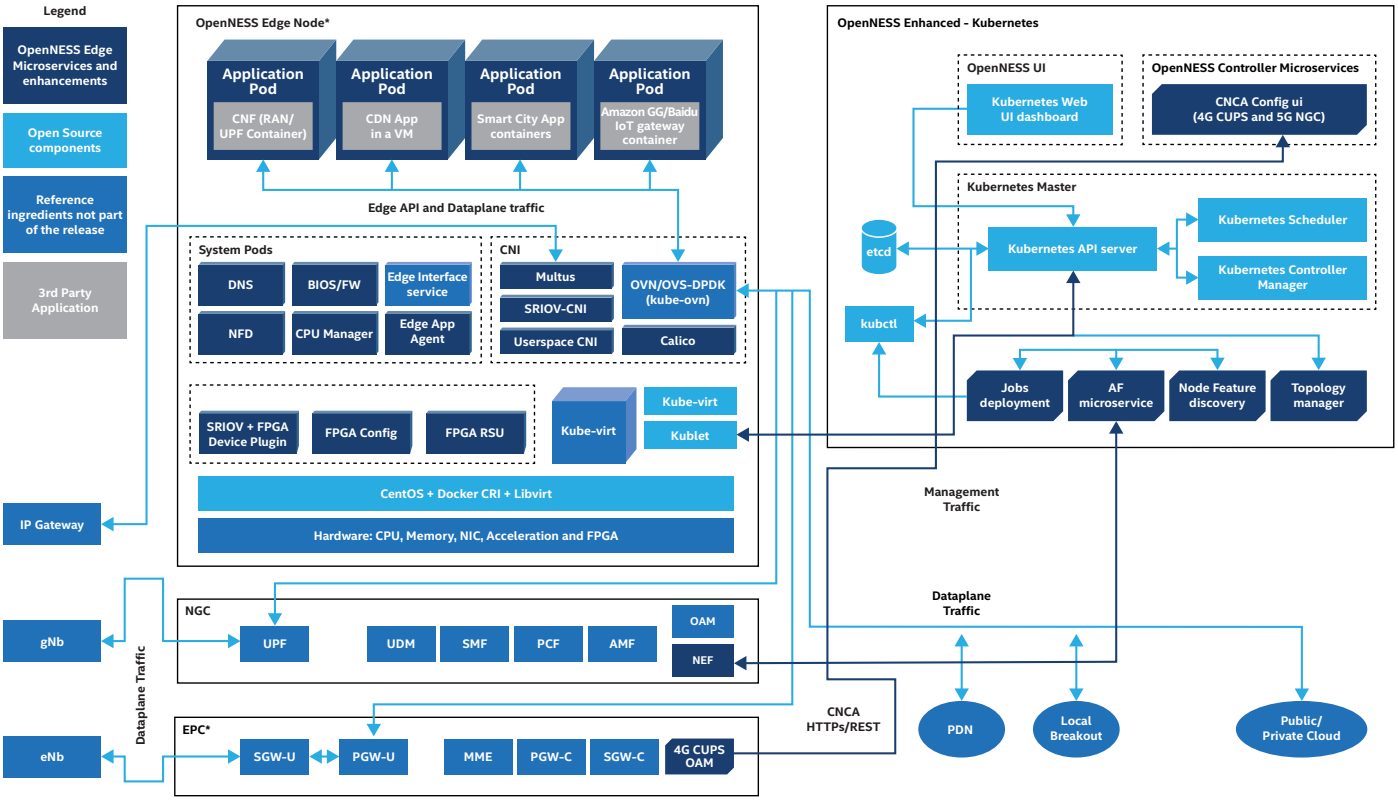

Figure 1. OpenNESS Network Architecture

About Baidu Intelligent Mobile Edge (IME)

Baidu IME is dedicated to meet the requirements of cloud-edge collaboration and multi-cluster management in 5G edge computing, as well as the needs of edge applications to access 5G + AI capabilities. The platform is based on Baidu's open source, unified-access edge platform IME Stack, on top of which additional 5G + AI collaboration capabilities are added. It is a native 5G + AI edge computing infrastructure.

IME has two main functions: the first is to extend Baidu's cloud AI capabilities to the edge and become a lightweight, efficient and secure AI microservice platform for edge applications; the second is to expose packaged 5G acceleration capabilities to developers, thereby helping them better integrate AI applications with 5G.

IME's architecture consists of three main layers: a lightweight edge computing PaaS layer, a service capability layer and a service capability exposure layer.

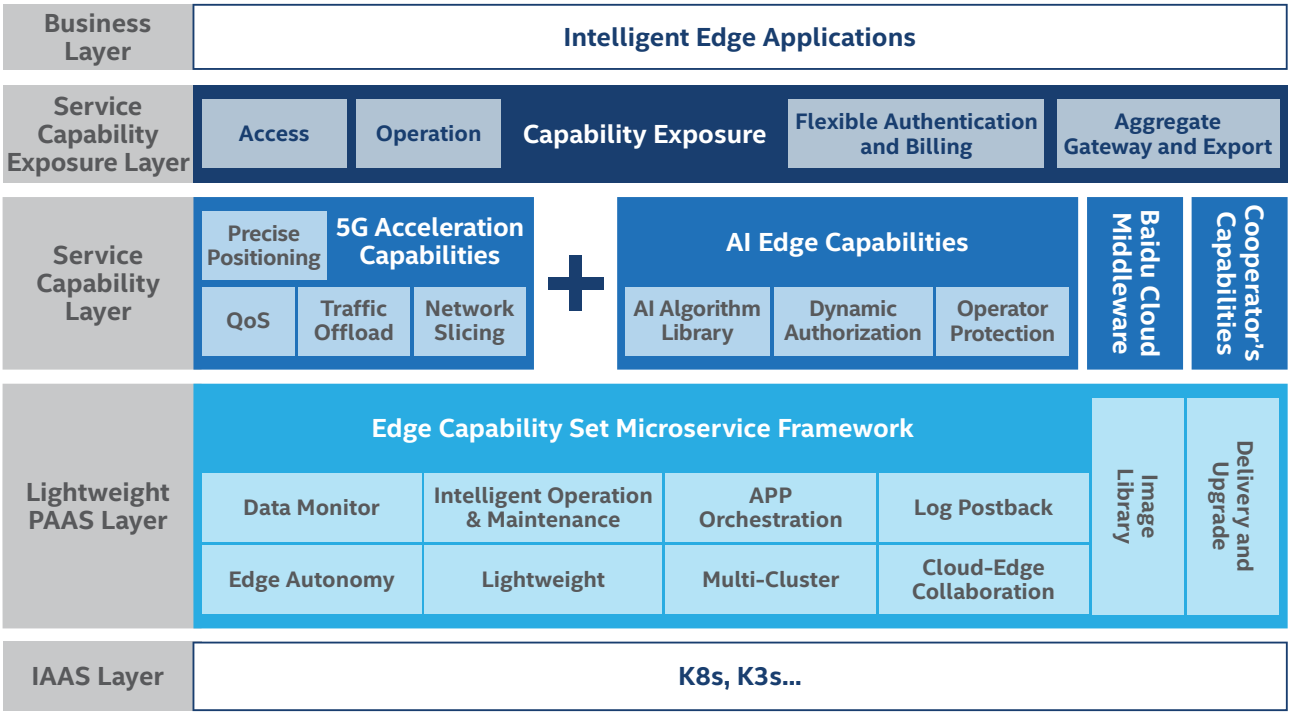

Figure 2. IME Architecture

Lightweight Edge Computing PaaS Layer

- Multi-cluster management and scheduling: For scenarios where a single K8s cluster can only support limited nodes, two clusters can be connected in a north-south configuration through the Cluster Controller and Cluster Shim to form a cluster tree topology to achieve unlimited cluster expansions. Through the Cluster Controller, all the clusters can be integrated for unified scheduling, thereby achieving optimal use of edge resources on a broader level. IME is also highly autonomous and can support autonomous operation in the case of poor network conditions. Currently, IME supports K8s and K3s clusters , but it can also expand to support any type of cluster through the Cluster Shim. IME can also effectively resolve the management and scheduling challenges of large-scale mobile edge clusters while maintaining compatibility with different operators' networks.
- Center cloud-edge coordination: Includes resource coordination, computing power coordination, and data coordination. Among them, the coordination of resources and computing power is achieved through hierarchical cluster management and unified scheduling; while coordination in data is achieved through timely synchronization between stateful data. The Cluster Controller and Cluster Shim can establish a data tunnel for cloud-edge collaboration. One deployment in the cloud will enable deployment to a large number of edge nodes and vice versa.
- A lightweight microservice architecture: Microservices are agile, compact, and facilitate rapid development, and are suitable for rapid iteration of edge computing scenarios. By referencing to Istio implementation, IME is designed to be lightweight and interface with different IaaS (K8s/K3s). The framework can provide network functions such as authentication, throttling, circuit breaker, and Transport Layer Security (TLS).

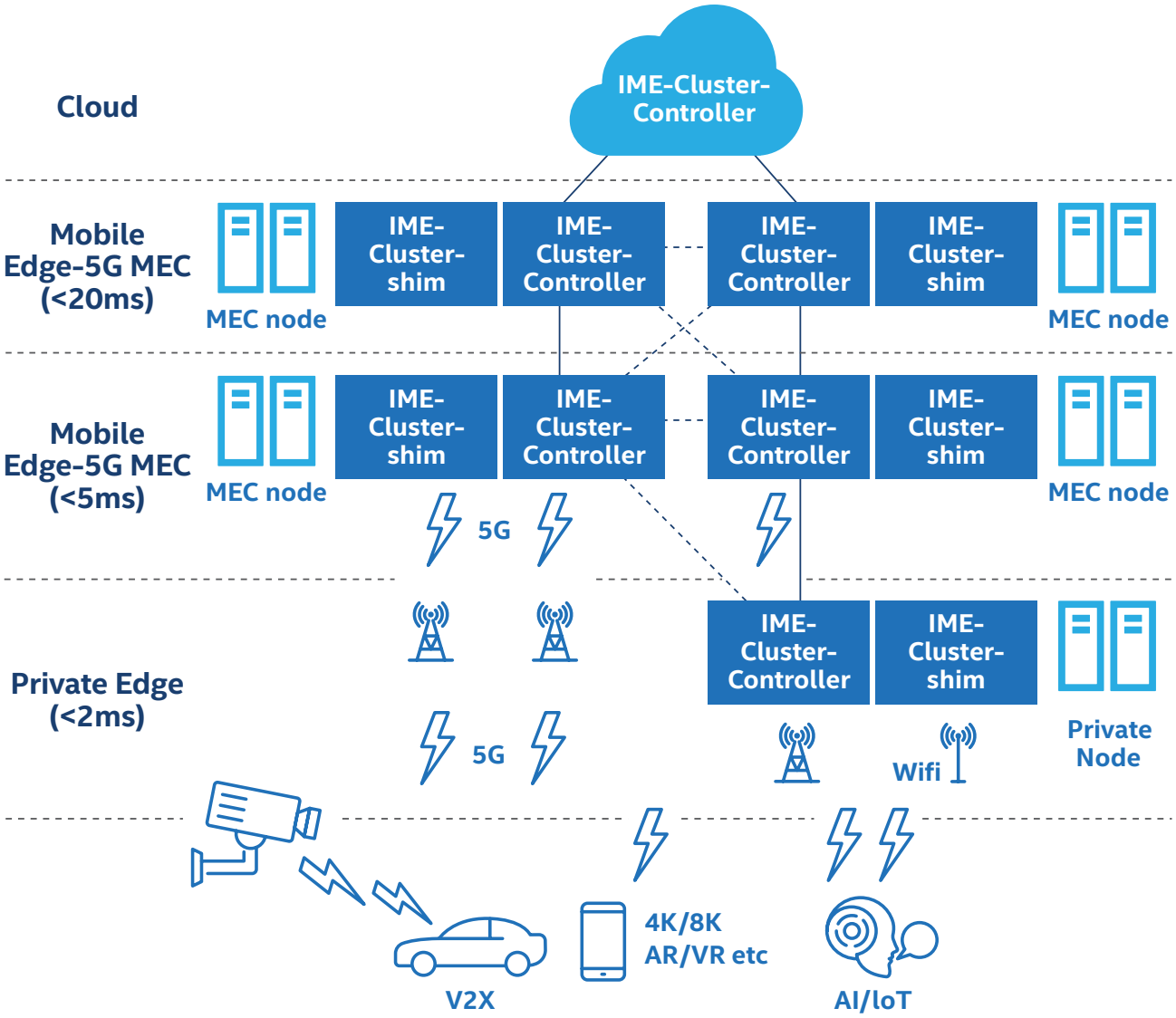

Figure 3. IME's Multi-Cluster Management Architecture

Service Capability Layer

Provides three levels of service capability to satisfy the needs of various business scenarios and applications:

- Fundamental service capability: Basic AI algorithms, basic 5G network capabilities (with reference to AF standards and support for multiple operator networks) and on-demand middleware services that are all available to be utilized directly.
- Composite service capability: Several fundamental capabilities are combined into composite capabilities to address the requirements of specific or more complex business scenarios and applications.
- Service Orchestration: Customized orchestration templates are available to allow rapid development and deployment to address the requirements of typical business scenarios and applications.

Service Capability Exposure Layer

Provides three types of capabilities for the capability products mentioned above:

- API Gateway: Third-party service capability registration API, unified service capability exposure API, external TLS termination, internal TLS establishment, and circuit breaker API.
- Billing: Flexible and personalized billing API to satisfy various business needs.
- Operational: invocation requests, user profile maintenance API and many more.

Combining the Strengths of OpenNESS and IME to Create 5G + AI Edge Middle Platform

Intel and Baidu share common goals for OpenNESS and IME, such as enabling AI edge computing capabilities with low latency and high reliability, and promoting 5G + AI capabilities that can be deployed in commercial and business scenarios. To achieve these goals, Intel and Baidu worked collaboratively on a 5G + AI Edge Middle Platform solution. This Middle Platform uses ETSI reference architecture as a foundation and incorporates mature PaaS technology to provide 5G + AI capabilities to edge applications.

Combining the Strengths of OpenNESS and IME

- OpenNESS: Provides components such as CNCA and AF to configure a 5G core network control plane (Network Exposure Function or NEF) and data plane (User Plane Function or UPF) to manage 5G network resources and expose 5G capabilities.
- IME: Acts as a 5G + AI edge computing enablement framework, allowing Baidu's cloud capabilities to deploy rapidly at the edge to achieve "cloud-edge" integration, and expose AI capabilities.

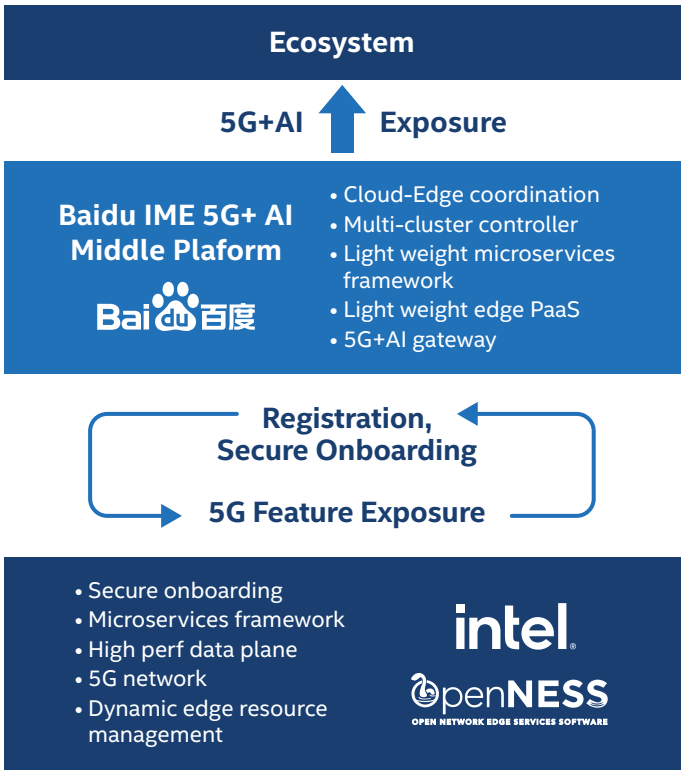

Figure 4. Combining the Strengths of OpenNESS and IME

Overview of the 5G + AI Edge Middle Platform

Based on the design methodology of the 5G + AI Edge Middle Platform (which consists of OpenNESS and IME), OpenNESS is an implementation of a MEC platform. It supports the interaction with 5G networks and the exposure of 5G capabilities. On the other hand, IME acts as a super PaaS APP of MEC. It provides IME applications with a lightweight microservice-based operating environment and 5G + AI capabilities (which it obtained through MEP). IME's Cluster Controller and Cluster Shim are responsible for docking with MEPM, and handling operations such as updating and deploying AI and 5G acceleration capabilities. The Cluster Controller also manages multiple MEC clusters, as well as cloud-edge collaboration.

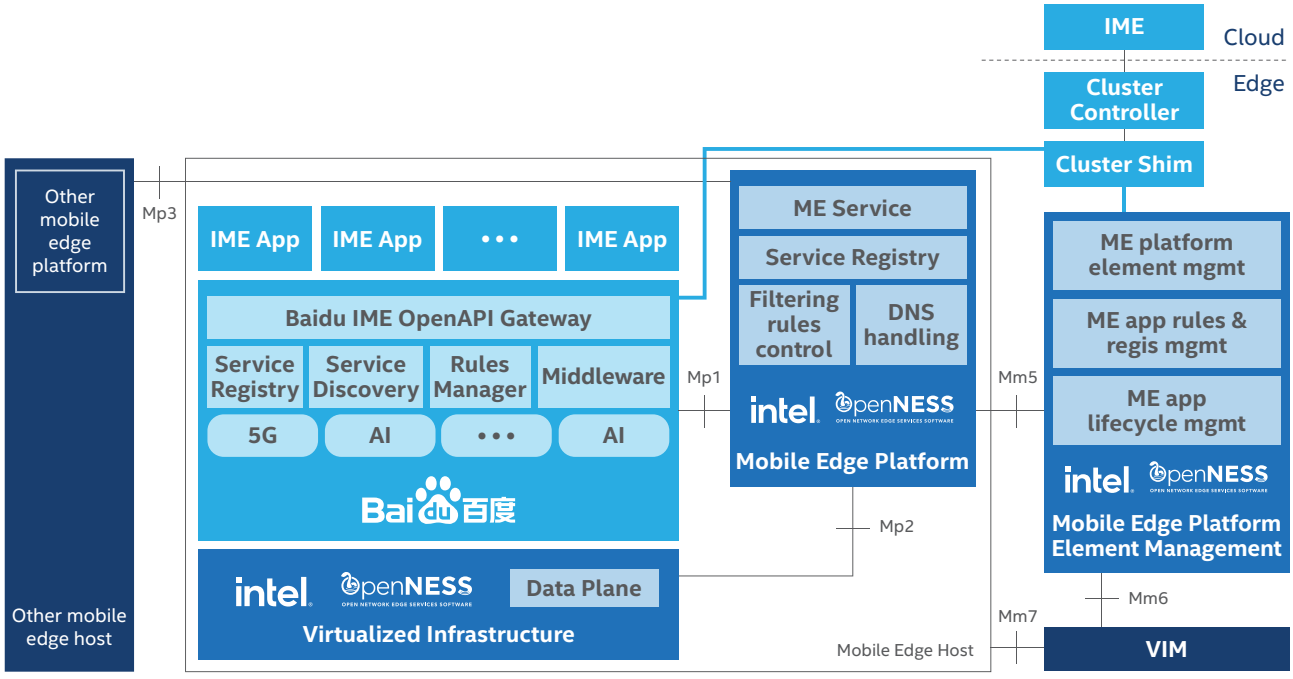

Figure 5. Positioning of OpenNESS and IME

Following the design philosophy above, the structure of the 5G + AI Edge Middle Platform is as follows:

- IaaS and part of the PaaS functions: Provided by OpenNESS.
- PaaS functions: OpenNESS provides MEP and takes care of the interactions with 5G core networks; IME handles multi-cluster management, cloud-edge collaboration and provides a lightweight microservice framework for applications.
- 5G + AI capabilities middle platform: Provided by IME.

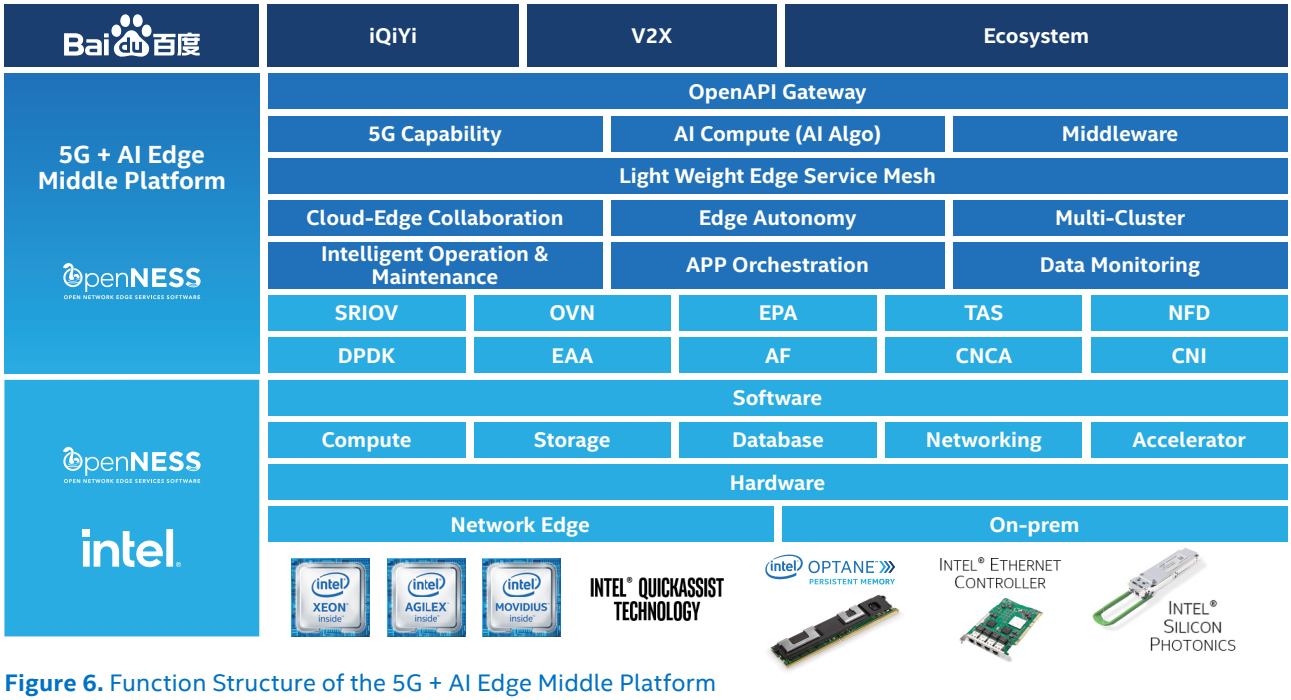

Figure 6. Function Structure of the 5G + AI Edge Middle Platform

Key Operation Flow of the 5G + AI Edge Middle Platform

1. **Initialization between the controllers:** Based on IME's multi-cluster management mechanism, the Cluster Shim will first initiate service registration requests to the OpenNESS Controller, in order to form a 5G + AI Edge Middle Platform.
2. **OpenNESS 5G AF authentication and registration:** OpenNESS 5G AF interfaces with 3GPP 5G networks to complete the registration and authentication. OpenNESS will also acquire 5G network functions and expose 5G capabilities to applications.
3. **Capability registration of the Edge Middle Platform:**
   a. The IME Cluster Shim obtains service and endpoint information (in the case of K8s) for IME components from the responses of the deployed OpenNESS Controller.
   b. IME 5G and AI capabilities register with the IME Open API Gateway. 5G service capabilities exposure by AF will also be registered with the IME Open API Gateway through OpenNESS controller.
4. **Completion of establishment. Application services start:**
   a. Applications are now exposed with 5G and AI capabilities through the IME Open API Gateway.
   b. Through RESTful API, AF will directly expose the service information to the IME Cluster Shim, IME Open API Gateway followed by applications.

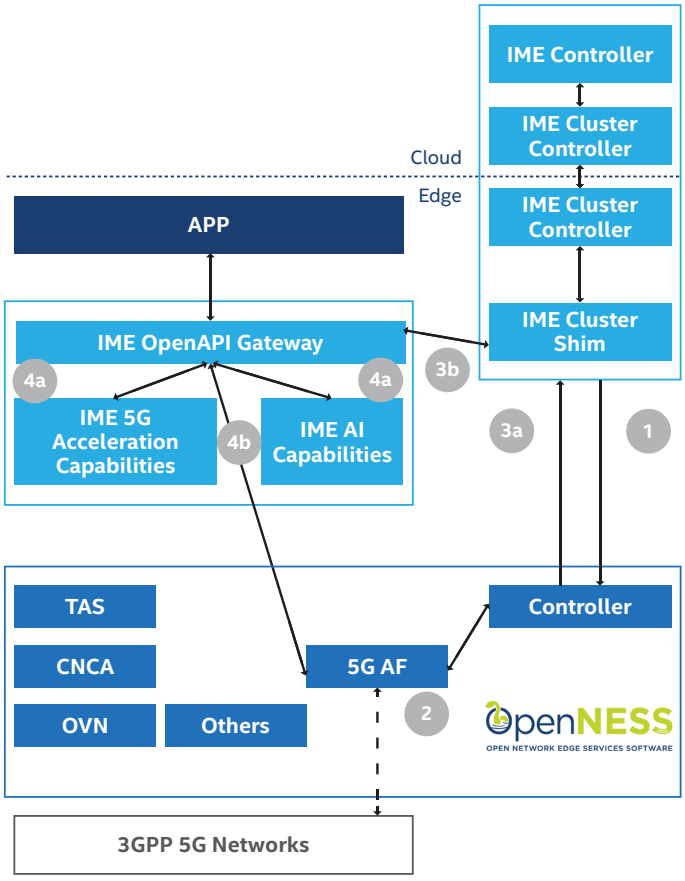

Figure 7. Key Operation Flow of the 5G + AI Edge Middle Platform

Typical Use Cases of the 5G + AI Edge Middle Platform

V2X Car Road Cooperation

Vehicle to Everything (V2X) refers to the communications between vehicles and the outside world and is a key technology in the development of Intelligent Transportation Systems (ITS). The 5G + AI Edge Middle Platform created by OpenNESS and IME provides reliable edge computing support for Baidu's Apollo V2X open source solution (also released in Akraino[1], an open source community). This solution gathers and integrates multi-sensory data from the road and provides computing power, real-time performance and deep learning capabilities at the edge (roadside). It leverages multi-scenario semantic intelligence, live high-definition maps, cloud-edge coordination, flexible communication access, and the ability to reach millions of mobile devices of road users on the move. Its mission is to meet the future demands of large-scale deployment of autonomous vehicles and collaborative cloud-vehicle infrastructure solutions.

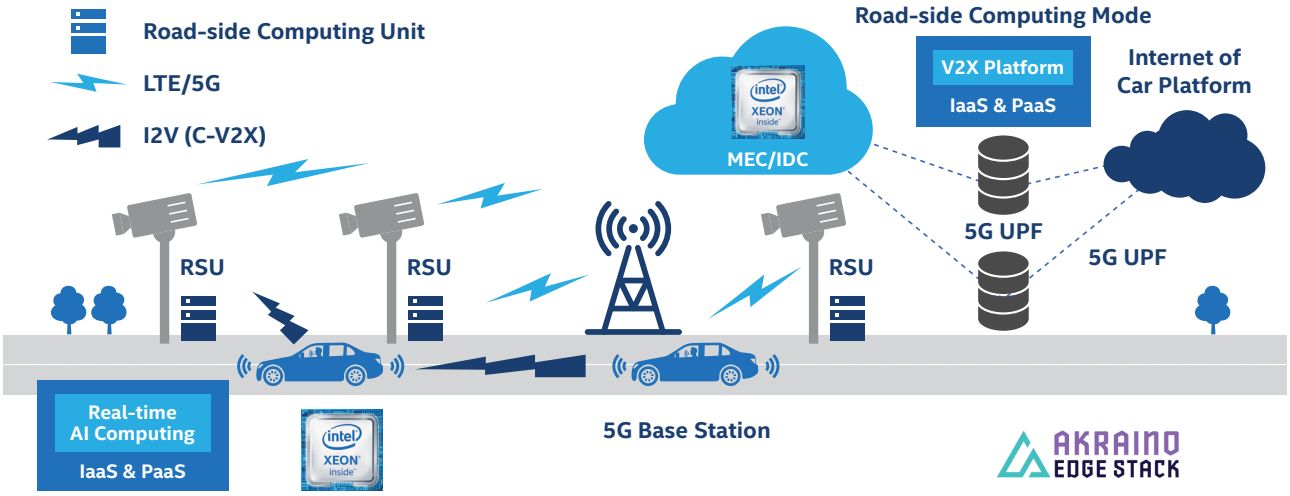

Figure 8. V2X System Design

5G Video Streaming Acceleration

The 5G + AI Edge Middle Platform, created by OpenNESS and IME, has achieved several 5G interconnect functionalities - including Quality of Service (QoS) acceleration, user equipment location querying, UPF breakout application, etc. Based on these capabilities, the Middle Platform is able to accelerate high-definition (HD) videos at the edge based on user equipment location information. The video distribution center's Global Server Load Balance (GLSB) function can identify CDN resource status at the edge, user location and subscription level (VIP or standard), and then achieve 5G video acceleration at the edge for VIP users, creating a higher-quality video experience. Currently, this solution is providing valuable edge acceleration support to Baidu's video services, including QiShu Player (from iQIYI) and VR experiences.

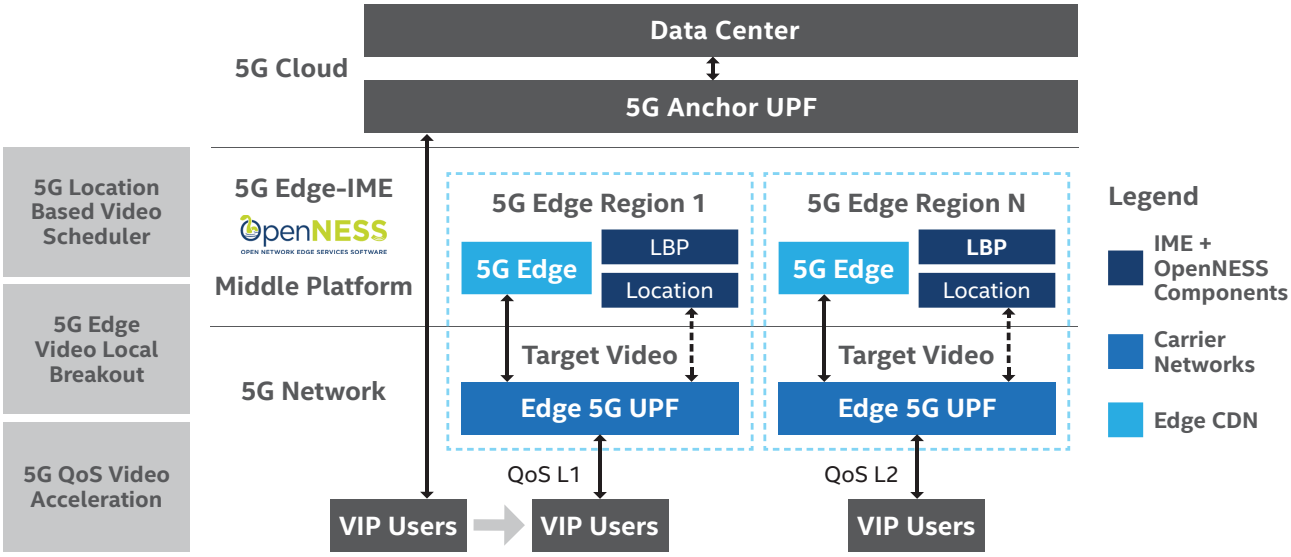

Figure 9. 5G Video Acceleration Logic

[1] The AI Edge: Intelligent Vehicle-Infrastructure Cooperation System(I-VICS), https://wiki.akraino.org/pages/viewpage.action?pageId=6129013

Forward Looking: Continuous Integration of AI, 5G and Edge Innovations

Using both OpenNESS from Intel and IME from Baidu, the companies are enabling a 5G + AI Edge Middle Platform – an open platform, which integrates carriers' MEC capabilities to meet the needs of AI edge computing. At the current stage, we are validating the application of this Middle Platform through use case scenarios such as V2X and edge video acceleration. In parallel, Intel and Baidu continue to collaborate to improve the performance of the Middle Platform. In the near future, we wish to co-develop innovations with the telecommunications community, and also to explore wider collaboration opportunities. We will collaborate with multiple industries to promote further innovation of technology, products and business models in the 5G and AI era, bringing intelligence closer to all industries, and users.

OpenNESS is available to download now at https://github.com/open-ness.

Find out more at https://www.openness.org/.

IME Stack is available to download now at https://github.com/baidu/ote-stack.

Find the solution that is right for your organization. Contact your Intel representative or visit www.intel.com/networktransformation.

Learn More

- Intel VCAC-A
- Intel Xeon Processors
- Intel Movidius™ Vision Processing Units

Authors

Soo Jin Tan
Intel 5G Solution Architect

Andrew Chih Howe KHOR
Intel 5G Infrastructure Marketing Director

Candy He
Intel Global Account Technical Manager

Xiaopeng Tong
Intel Senior Software Engineer

Chen Gang
Baidu Cloud Edge Computing Senior Architecture

Kong Dechao
Baidu Cloud Edge Computing Senior Architecture

Dong Xun
Baidu Cloud Edge Computing Senior System Engineer

He Siyi
Baidu Cloud Edge Computing Senior R&D Engineer