AUTOMOTIVE EDGE
COMPUTING CONSORTIUM

# General Principles and Vision

## White Paper

Version 4.0.4 • October 2024

# Contents

# 1. Introduction

## 1.1. Background

To make driving safer, traffic flow smoother, energy consumption more efficient and emissions lower, mobile communication in vehicles is increasing in importance [1][4]. Several emerging services, such as intelligent driving, the creation of maps with real-time data and driving assistance based on cloud computing, require vehicles to be connected to the cloud and networks. This will facilitate the transfer of a large amount of data among vehicles and between vehicles and the cloud. The following forecasts are made for the 2030 timeframe [1] [2][3][4].

1. $250 billion to $400 billion could be created in annual incremental value enabled using vehicle data.
2. The number of connected vehicles will grow to over 600M globally.

In the forecasts, 1~2 terabytes of raw data per car each day will be required to enable continuous product and service improvements. Future automotive services could require data transfer offload even higher.

There will be a need for new network architectures and computing infrastructure to support massive computing resources and topology-aware storage capacity to balance quality and cost. This, however, cannot be achieved without taking further actions, and failure to do so will limit the evolution of future services in the automotive industry. The cellular network is one of the major mobile networks for connected vehicles, and many specifications have been standardized in the 3rd Generation Partnership Project (3GPP). However, the present work within 3GPP has not fully addressed the challenge of automotive big data, and therefore future network deployments and business models will fail to support the future needs of connected vehicles.

The cellular vehicle-to-everything (C-V2X) communication considered in 3GPP, for example, mainly covers latency-sensitive safety applications and may not fully ensure the big data capacity growth between vehicles and the cloud. Toward 5G, massive machine-type communication (MTC), including narrowband (NB)-IoT has been considered by 3GPP, and it is intended to connect a massive number of small low-power sensor devices. Still, the data volumes are considered fairly modest. But adding to this, the current trend of concentrating data processing at central locations will cause huge data transmission traffic, which will lead to unnecessarily long response times and in turn will increase computation time. Assuming 1TB per day per vehicle and 18 million vehicles (12% market share and 25% regional ratio of 600 million vehicles), 18 exabytes of vehicle data will come to the cloud every day. For this reason, establishing a practical platform to serve Vehicle-to-Cloud (V2Cloud) services, both computation and network performance, needs to be taken into account (see Figure 1).

We believe that the current mobile communication network architectures and cloud computing systems are not fully optimized to handle the requirements of connected vehicles effectively. Therefore, it is beneficial to investigate how to redesign the system architecture and reconsider network deployments to better accommodate network traffic.

One possible solution is through topology-aware computing and storage resources. Our aim is to deploy this redesigned system architecture on a global scale, which will require collaboration among worldwide partners and the system architecture to comply with relevant standards.

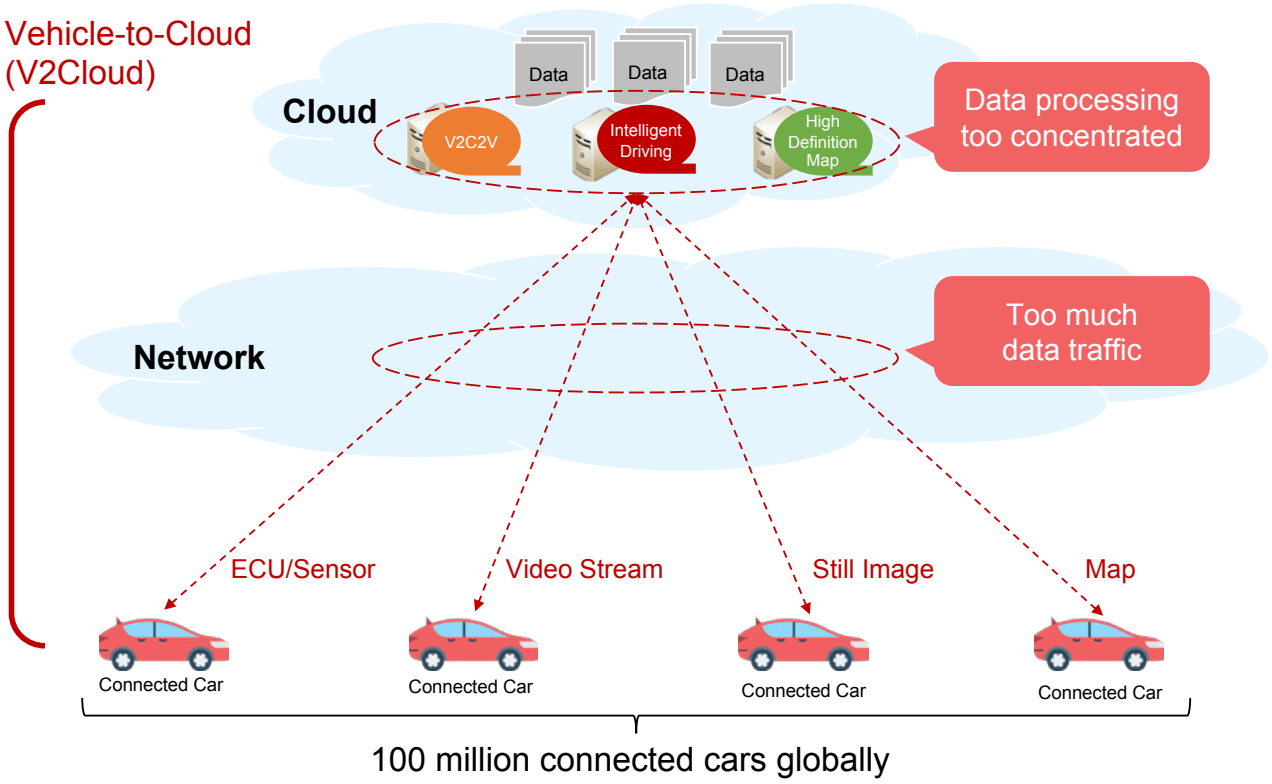


*Figure 1: Problems of existing technologies' deployment*

## 1.2. Objective

Network capacity planning has become a major challenge for mobile network operators due to the soaring costs associated with the exponential increase of data traffic. At the same time, new vertical markets, such as the automotive sector, have an ever-increasing number of devices with high-capacity demands connected to the network Thus, a new communications offering is needed to address these industries' specific business and technical requirements.

As next-generation networks are being standardized, we have a unique opportunity to ensure that future networks are designed and deployed to provide new services in a reasonable fashion to vertical markets such as the automotive industry. At the same time, these networks can bring new customers and generate new revenue for mobile network operators.

This white paper highlights the need for market actors such as communication technology companies and automotive manufacturers to work together to ensure that future systems are designed to address the challenges mentioned above. One technology explored in the new network design will be topology-aware distributed clouds with multi-operator edge computing capabilities.

# 2. Concept

## 2.1. Distributed Computing on Localized Network

To solve the problems of data processing and traffic volume on the existing mobile and cloud systems described above, we introduce “Distributed Computing on Localized Networks” (see Figure 2). In this concept, several localized networks accommodate the connectivity of vehicles in their respective areas of coverage. Computation power is added to these localized networks to enable them to process local data, allowing connected vehicles to obtain responses in a timely fashion.

The concept is characterized by three key aspects:

1. **Localized Network.** A local network that covers a limited number of connected vehicles in a certain area. This splits the huge amount of data traffic into reasonable volumes per area of data traffic between vehicles and the cloud.
2. **Distributed Computing.** Computation resources are geographically distributed within the vicinity of the localized networks’ terminations. This reduces the concentration of computation and shortens the processing time needed to conclude a transaction with a connected vehicle.
3. **Local Data Integration Platform.** Integration of local data by utilizing the combination of the localized network and distributed computation. By narrowing relevant information down to a specific area, data can be rapidly processed to integrate information and notify connected vehicles in real time.

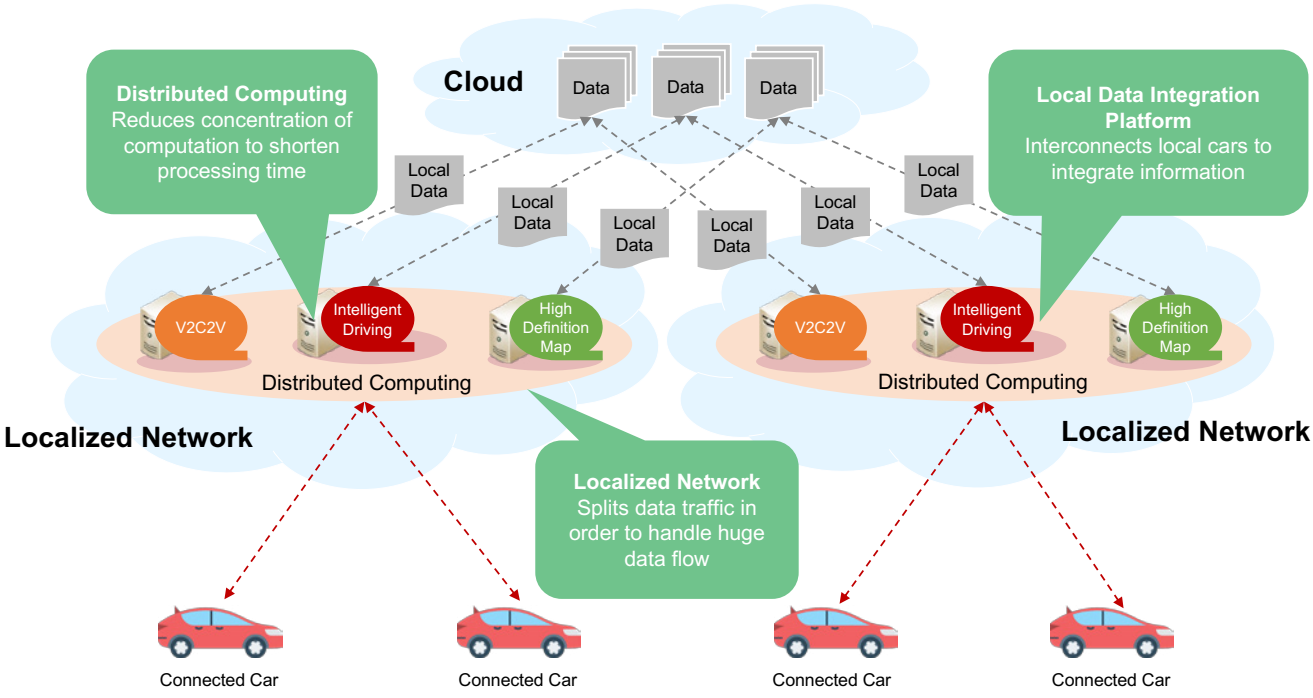


*Figure 2: Distributed computing on localized networks*

## 2.2. Edge Computing for Automotive Use

As mentioned in the previous chapter, the concept of distributed computing on localized networks has three key aspects that need to be implemented. Edge computing is one promising technology, due to its features and advantages, that could be adopted to realize this concept. In automotive use cases, edge computing technology will provide an end-to-end system architecture framework that enables distribution of computation processes over localized networks as depicted in Figure 2.

The edge computing technology used for our concept of distributed computing on localized networks consists of two key components: the network and the computation resources.

The network is designed to split data traffic into several localities that cover reasonable numbers of connected vehicles. The computation resources are hierarchically distributed and layered in a topology-aware fashion to accommodate localized data and to allow large volumes of data to be processed in a timely manner (see Figure 3). In this infrastructure framework, localized data collected via local networks and wide-area data stored in the central cloud are integrated in the edge computing architecture. This provides real-time information necessary for the services of connected vehicles. In the context of edge computing for automotive use, the “edge” means the hierarchically distributed non-central clouds where computation resources are deployed, and edge computing technology can be used to design such a flexible topology-aware cloud infrastructure.

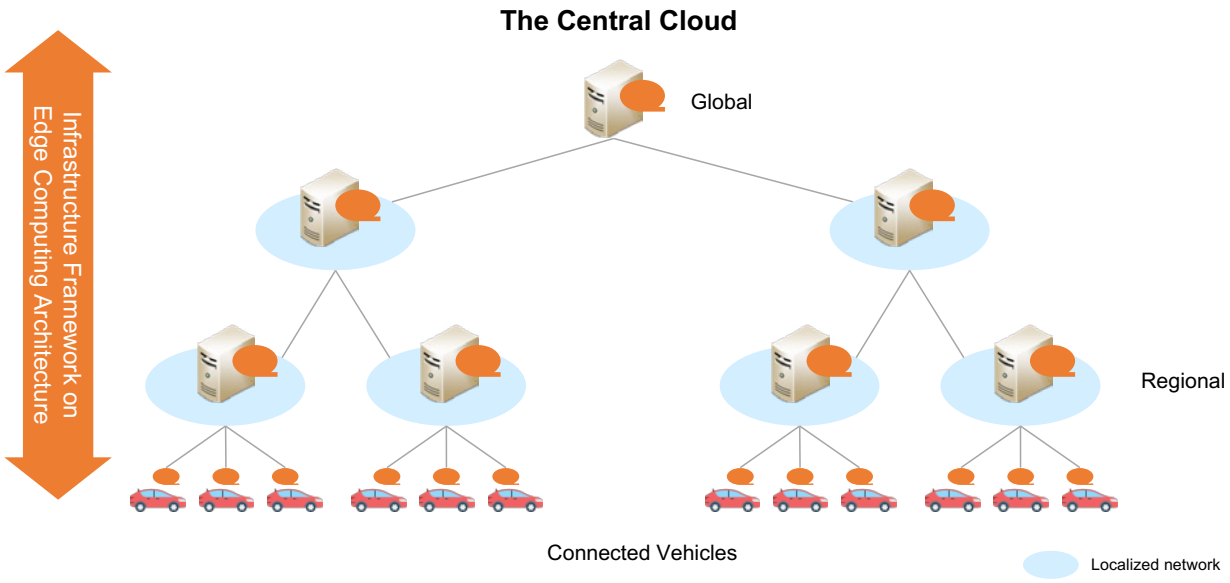


*Figure 3: Edge computing for automotive use*

Edge computing is the key technology to realize the distributed computing on localized networks concept in the automotive industry. Therefore, the Automotive Edge Computing Consortium will focus on increasing capacity to accommodate automotive big data in a reasonable fashion between vehicles and the cloud. We will do this by means of edge computing technology and more efficient design of networks.

The consortium will define the requirements and develop use cases for emerging mobile devices, with a particular focus on the automotive industry, bringing them into standards bodies, industry consortia and solution providers. The consortium will also encourage development of best practices for the distributed and layered computing approach.

# 3. Service Scenarios

Network-based computation will make it possible for automotive services, especially V2Cloud services as shown in Figure 4, to come to life.

These V2Cloud services cover a broad range of functions, from sales and marketing to digital twin and vehicle teleoperation. The enhanced vehicle feature, in particular, is the most promising business area for next-generation connected cars. This service scenario includes, amongst other services, intelligent driving, high-definition map generation and V2Cloud cruising assist. These services will produce huge traffic volumes with varying levels of latency requirements.

Beyond these services, some extended services might also arise, such as telematics, insurance/financial services and traffic control. These extended services will also generate a tremendous amount of data traffic and processing for future infrastructure.

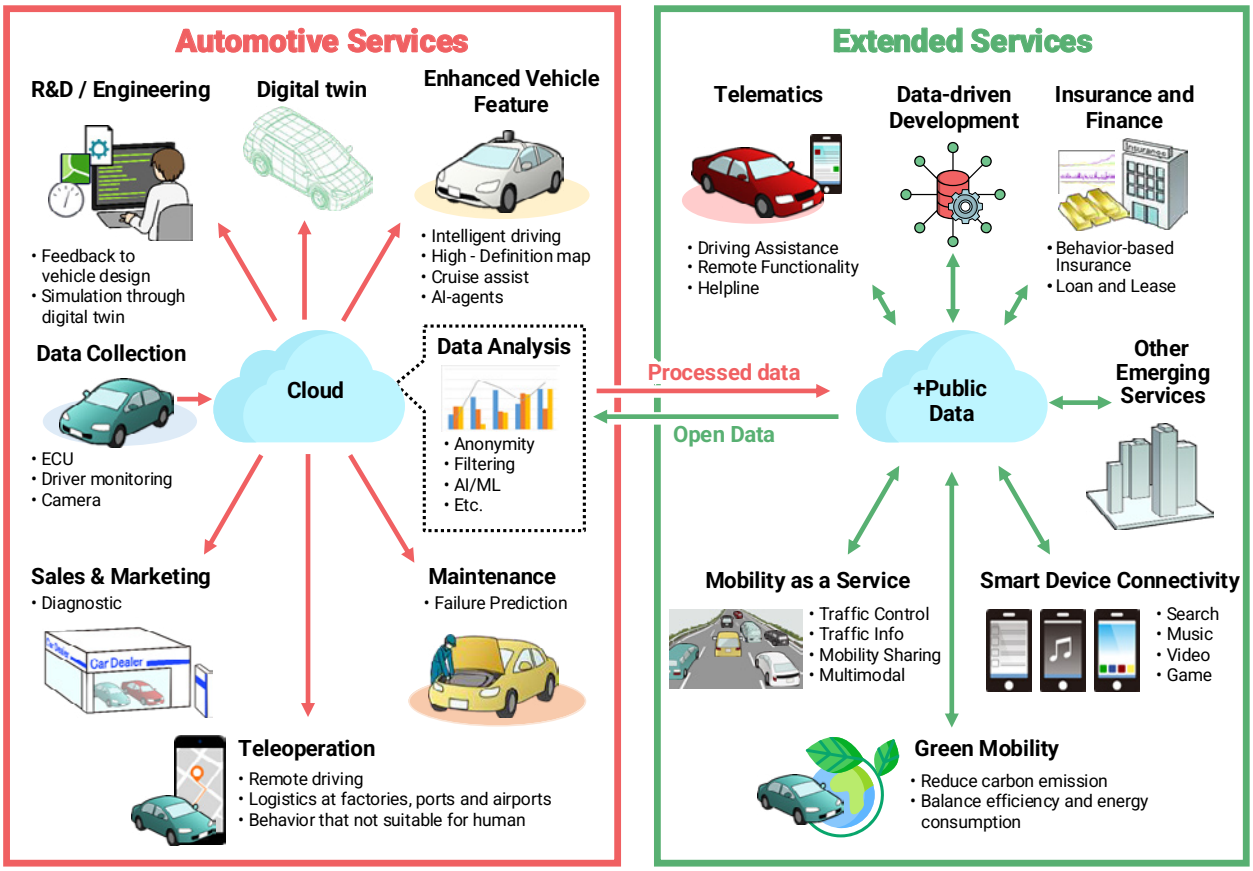


*Figure 4: Emerging automotive V2Cloud services*

The following examples show some typical V2Cloud service scenarios.

## 3.1. Intelligent Driving

Intelligent driving currently means safe and efficient driving support, but over time will incorporate autonomous driving. In this context, vehicles need to exchange vast amounts of various kinds of data with applications in the cloud. Although there are many types of service scenarios related to intelligent driving, we will describe one typical scenario for smart driving support here.

In our intelligent driving scenario, the driver's physical condition is monitored and an evaluation of his or her driving performance is given as the output. In this service scenario, the cloud service needs to collect data such as cruising data, biometric sensor data and control data. These are all gathered from various sources including movement logs from in-vehicle sensors and on-board biometric sensors/cameras.

The data volume is very large, creating a heavy load on both networking and computing resources and/or less optimal use of the network resources. Edge servers can help to pre-process the data on the way to the cloud, and instruct the vehicle on what data to send and how to process the data, decreasing the amount of data sent to the cloud.

The collected data is then sent to the cloud via the access networks for processing. The transfer of this data should be done effectively (preventing any loss of data) and efficiently (justifying the cost-performance ratio).

To support vehicles that are moving, the AECC system should have the ability to transfer the ongoing data session from one edge server to another.

Based on the data collected, the cloud computes the intelligent driving parameter set using advanced machine-learning techniques. It is important to send those parameters to vehicles in a timely manner. (As these parameters' data volume is relatively small, high-data-volume handling is not necessary to disseminate the parameters.)

In an advanced scenario, the cloud could serve multiple vehicles via multiple cellular networks operated by different MNOs.

## 3.2. High-definition Map

The high-definition map consolidates static and dynamic information (e.g., vehicle position, pedestrians and obstacles) and is essential for autonomous driving. Creating and distributing the map require many data transactions with high capacity as well as efficient processing to keep the information up to date.

This high-definition map must be able to accurately localize dynamic objects including vehicles, which is required for automated driving beyond the traditional route guidance. A large amount of data transfer is especially required to update the map. Data is collected from on-board cameras, radar sensors and laser scanners (LIDAR), transferred and processed in the cloud. Typically, what might be sent to the cloud are deviations ("Map says X, but Camera says Y"). These deviations are sent to the cloud to update the high-definition map.

The completed map information is stored in the central cloud server or the edge server and needs to be distributed to relevant vehicles in a timely manner.

When a vehicle is using an edge server, it can provide a data compression feature to reduce the amount of traffic volume between it and its associated cloud. In other cases, an edge server can provide a data extraction feature to reduce the amount of data transferred between the edge server and the cloud.

If network congestion happens due to a concentration of vehicles, collected data for servers can be rerouted to other servers that have the capacity to process it. Otherwise, the edge server experiencing network congestion can cache the collected data temporarily until the network congestion is resolved.

Because vehicle transportation traffic flow changes from moment to moment, it sometimes creates a traffic jam, which may cause network congestion. For example, a vehicle accident will increase the number of vehicles in the radio cell where the accident occurs, and this in turn propagates to linked cells as well. In such a case, a large number of vehicles will upload quite similar data used for creating a high-definition map. This redundant data has a negative impact on the entire system. To avoid such a situation, communications traffic management schemes should be introduced.

A large amount of redundant information produces huge negative impacts on network and computation infrastructure, as stated above. To avoid this, the edge server needs to identify groups of data generated from the same event, such as a vehicle accident, by integrating them into one associated batch of information. This kind of data integration on the edge server is an effective pre-processing function to reduce computation cost.

## 3.3. V2Cloud Cruise Assist

V2Cloud cruise assist is an example of a use case with a more flexible service evolution model than conventional dedicated short-range communications (DSRC). Here the network mediates vehicle-to-vehicle communications by integrating information obtained from neighboring cars. This mechanism is called the vehicle-to-cloud-to-vehicle service or simply V2C2V. This service scenario is especially effective when used to broadcast information to vehicles that need the same information, utilizing the combination of neighboring vehicles, roadside units and other entities.

Since the vehicles are expected to be in motion, the application must account for vehicles moving from one access point (such as a cell tower) to another. In addition, vehicles will enter and exit the application's "zone of interest" as a vehicle's path changes, joining or leaving a group of vehicles that are participating in the application instance. As a vehicle's journey continues, it must also be expected that a vehicle will exit one localized network and its associated edge server, entering and joining an adjacent localized network.

First, the system needs to collect vehicle data in the same way as with intelligent driving. The target data includes various types of data written in the intelligent driving scenario and high-definition map scenario (e.g., cruising data, biometric sensor data, control data, on-board camera data, radar sensor data and laser scanner (LIDAR) data).

Since the analyzed information can be used by vehicles in specific locations (e.g., a small town, a crossroad, specific area of a highway), an edge server can create local information (e.g., a localized high-definition map) and can update it continuously based on various locally-collected data. The necessary functionality to create and manage the local information can be off-loaded to the edge server from its associated cloud.

The system to mediate data among different vehicles in the neighboring area requires a fast response that meets service timing criteria (i.e.., less than 10 seconds). To realize this scenario, the edge server needs to be carefully selected to fulfill the criteria based on latency, availability, computation load and so on.

To realize the vehicle data mediation scenario, vehicles (and in some cases local roadside units) transmit their cruising data to the cloud to be analyzed. This data might be providing information for driving assistance (such as collision avoidance, cruise control for platooning and signal control). The system to mediate data among different vehicles in the neighboring area also requires ultra-fast computing processing to meet service timing criteria (again, less than 10 seconds).

The information generated will then be distributed to relevant vehicles and roadside facilities in the neighboring area. The vehicle data distribution will require access points between the edge server, the cellular network and other (fixed-line, WLAN) networks, since roadside equipment may not be attached to the edge server.

To distribute information to each vehicle and select a correct route for it, the vehicle position needs to be tracked. When an edge server is used to notify a vehicle in real time, its position in relation to networks (including cellular and other types of network) is important information for identifying the best route to the vehicle from the edge server.

V2Cloud cruise assist is a use case where a service model that is more flexible than most conventional services is required. The V2C2V cruise assist application will be executed on the edge server.

Due to the nature of the application, service guarantees must be met in order to support the latency requirements. Further, during the application's lifecycle, edge server resource utilization levels will change, the application itself may require updates or maintenance, the edge server will require upgrades, etc. With a vehicle traveling at 100km/h, covering a distance of ~27 meters/second, it is vital that the impact to the vehicle be minimized during any operations that affect the function of the application.

## 3.4. Teleoperation

Teleoperation is a service scenario where an operator remotely drives a vehicle or a machine via network.

The core concept of teleoperation is to augment and extend human capabilities beyond geographical barriers. By removing constraints on the location of operators and their devices, teleoperation can provide new opportunities for people facing physical or geographical challenges and thus improve diversity and inclusion. Furthermore, teleoperation can encourage people to engage in value-added work by replacing low-value or hazardous manual tasks with remote operations.

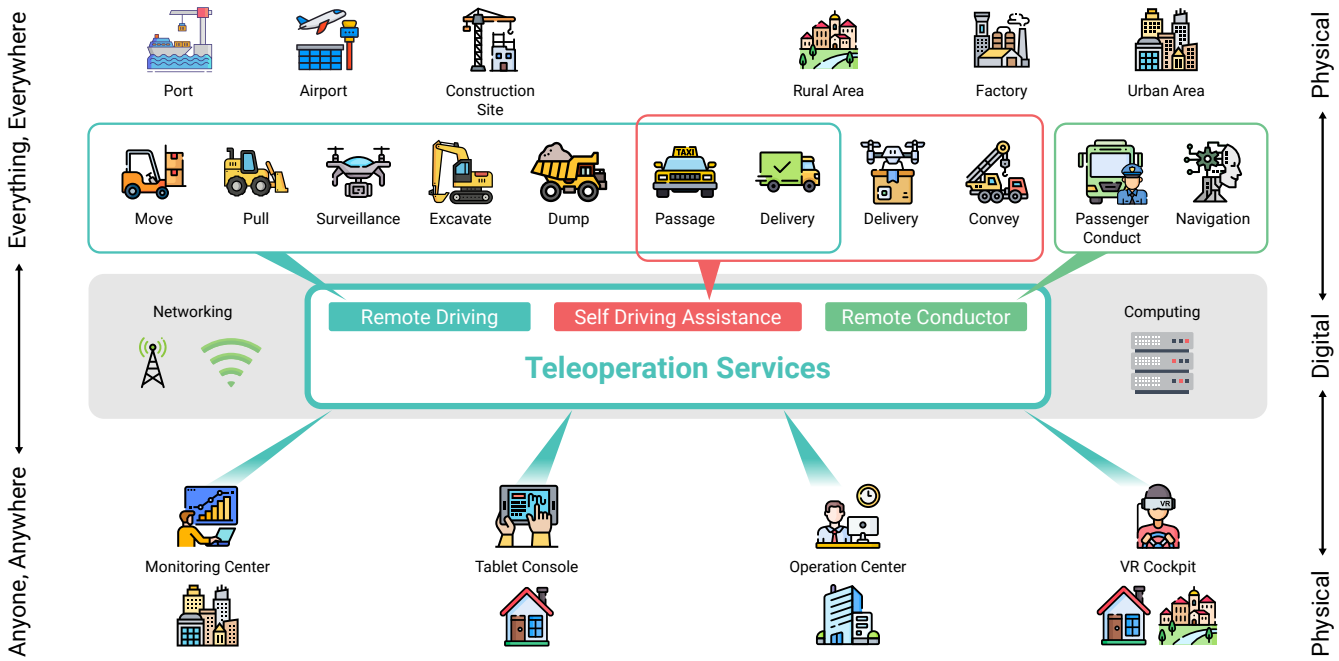


*Figure 5: Teleoperation target*

### 3.4.1. Vehicle Teleoperation

Vehicle teleoperation can improve the availability and efficiency of mobility services through vehicle remote driving and in-car monitoring. Automated driving can also benefit from vehicle teleoperation. When self-driving is unsuitable or unavailable, teleoperation allows human intervention at a remote site.

Vehicle teleoperation is applicable on both private property and public roads. On private property, vehicle teleoperation is mainly used for logistics at factories, ports and airports, so that site operators can use vehicles to replenish materials from storage, convey products to a deposit space or collect waste at a dump site. For example, an automotive manufacturer can move completed cars from a manufacturing facility to a temporary storage site.

Another use case for vehicle teleoperation is remote driving on public roads. Vehicle teleoperation is applicable for uncrewed transport between logistics centers and the last-mile deliveries to a final destination.

In addition, public transport with automated driving can also benefit from vehicle teleoperation. For example, teleoperated conductor services on self-driving vehicles can assure the safety of passengers through remote monitoring and intervention. With vehicle teleoperation, a self-driving vehicle can continue its mobility service by switching to remote driving when an automated driving function is failing or temporarily unavailable.

One example of a vehicle teleoperation system consists of remote vehicles and an operator's terminal. The remote vehicles are capable of being teleoperated: they are equipped with onboard sensors such as cameras to collect sufficient environmental information and receive commands from the operator's terminal through the network to control themselves. The operator's terminal has a cockpit system with displays and speakers for monitoring the vehicle and a steering wheel and pedals for controlling the vehicle. The operator connects to the vehicle that is to be teleoperated and starts the teleoperation.

The vehicle teleoperation system requires a large volume of low-latency and stable data traffic between vehicles and the operator's terminal. If loss of connection or an extra delay and jitter on the teleoperation system occur due to network congestion, poor signal conditions or communication disconnections, there will be interference with teleoperation. For example, corruption of visual information from an on-vehicle camera can disturb driving operations, and a delay of control information can lead the vehicle into unintended conditions.

To realize teleoperation, the AECC system expects high throughput, low latency and continuous data flow between connected cars and the cloud to provide a stable driving experience. In a vehicle teleoperation scenario, it is assumed that a teleoperated vehicle is equipped with front/rear, left/right, and left/right diagonal forward-facing cameras with 2 MP (megapixel) resolution (i.e., 1920x1080 pixels) and 30 frames per second (FPS) video encoding. It's also important that images from those cameras are sent to the operator through the AECC system. As for the required throughput, the traffic volume for a single camera is 1 to 4 Mbps. Therefore, three pairs of cameras require 6 to 24 Mbps of available bandwidth during the service period.

The latency requirement depends on various factors and cases, such as acceleration/deceleration, steering, obstacle avoidance, etc. For simplicity, here it is assumed that there is a case of sudden braking in a car-following scenario where two vehicles are traveling at 100 km/h with an inter-vehicle distance of 100 meters between them, and the rear one is teleoperated. In order to stop the rear vehicle safely when the front one starts braking suddenly, the rear vehicle must start braking before it overruns the 100-meter distance -- that means it must start braking within 3.6 seconds.

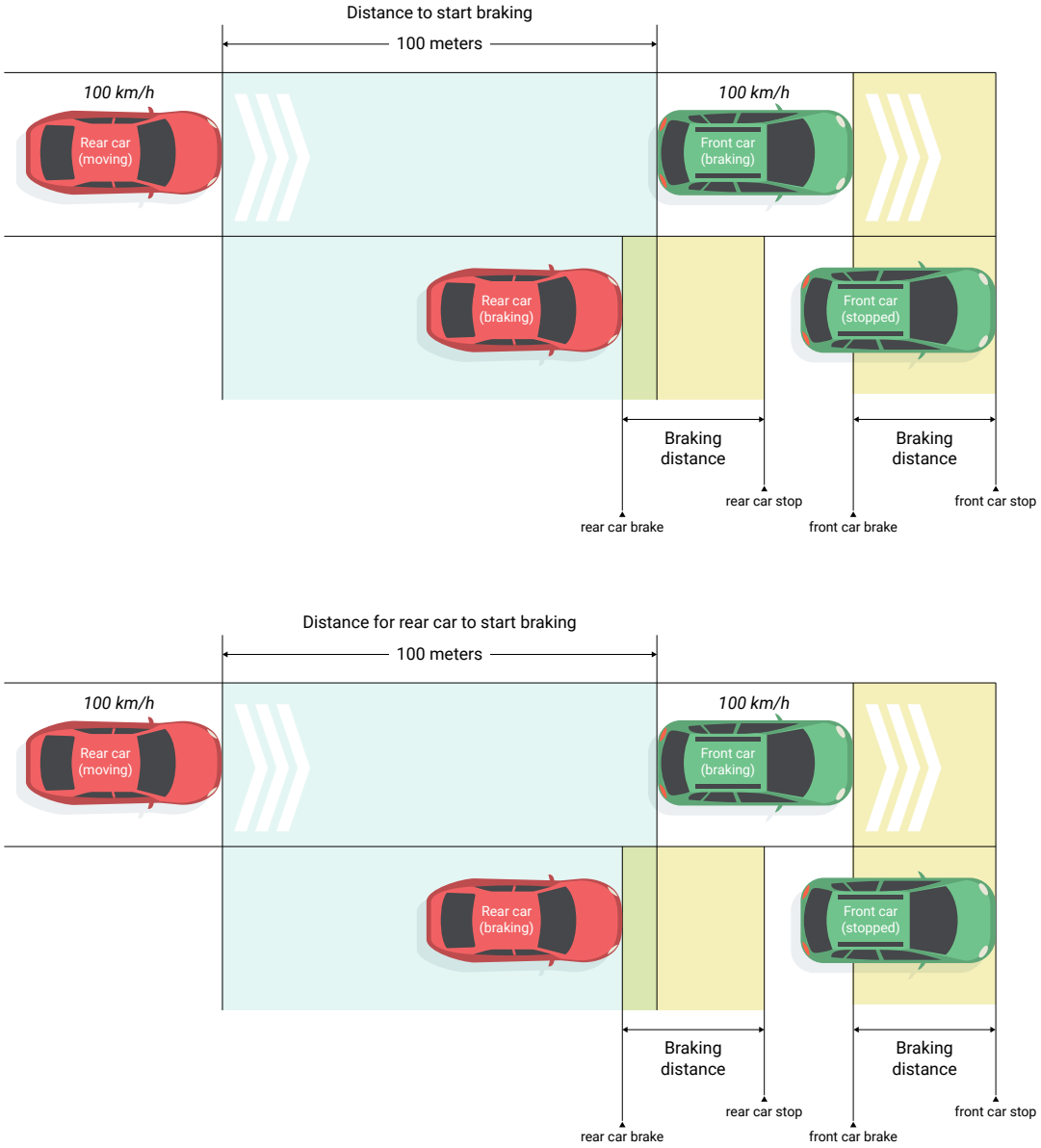


*Figure 6: Sudden braking in car-following scenario*

Because the average human braking reaction time is known to be 1.3 seconds, the latency of the teleoperation system including the AECC system must be within 2.3 seconds. Here, the latency refers to the sum of the duration required for the camera image of the vehicle to be sent to the remote operator (uplink latency) and the duration required for the operator's command to be executed in the vehicle (downlink latency). As shown in the figure below, the round-trip time, excluding the AECC system, can be subdivided as follows.

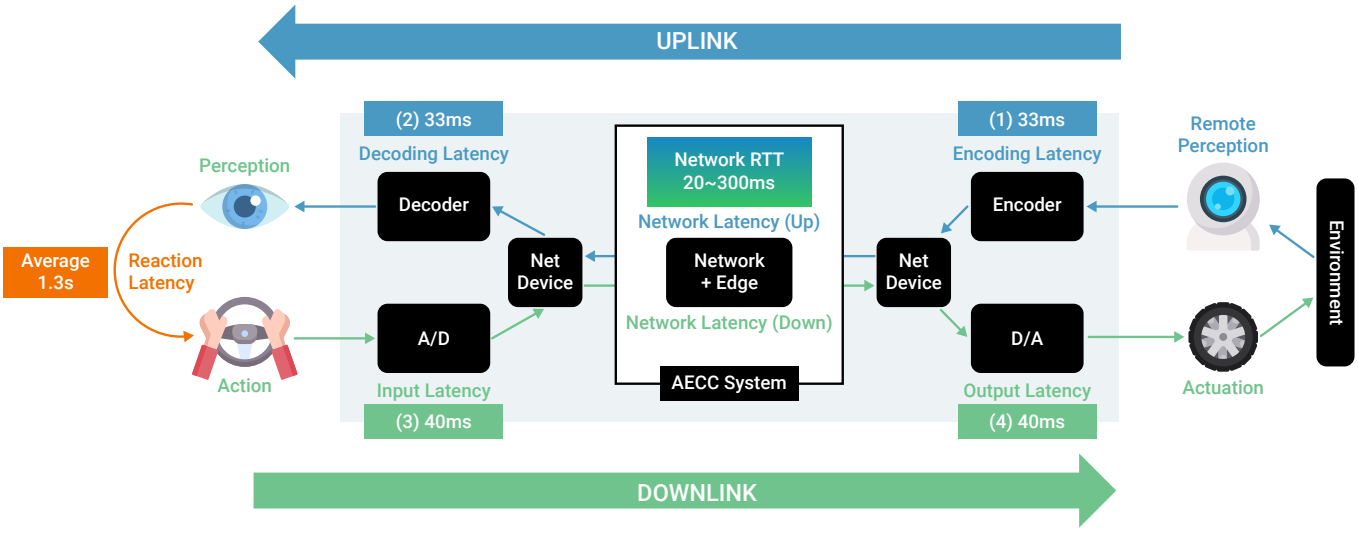


*Figure 7: Latency factors of a vehicle teleoperation system*

For the uplink direction:

1. 33 ms is required for sensing and encoding the video from the vehicle's onboard camera,
2. 33 ms is required for decoding and displaying an image on the cockpit system side

For the downlink direction:

1. 40 ms is required to detect and process the operator's command input, and
2. 40 ms is required to reflect the command on the vehicle side

The total of these latencies, excluding the AECC system, is about 146 ms. Therefore, the latency of the AECC system should be less than 2.154 seconds (2.3 seconds minus 146 ms). Considering the margin of error and other factors, it should be less than approximately 2 seconds.

A wide variety of efforts are applicable to the AECC system: from optimization within a single vehicle to collaboration between vehicles, edge servers and the cloud. Vehicles can utilize multiple networks from different network operators to enhance communication availability and quality. Edge servers can intervene in traffic and optimize data flow for better performance. The cloud system can cooperate with network operators to estimate and predict network availability to control network usage on each vehicle.

## 3.5. Voice-interactive AI Agents

Voice user interface (VUI) applications that engage in dialogue with AI agents are beneficial for enhancing the user experience of drivers who cannot afford to take their hands off the wheel while driving. As a result, many OEMs offer various in-vehicle VUI applications for many tasks, such as phone calls, music playback, navigation and temperature control.

However, as the functionality of in-vehicle systems becomes more complex, it is becoming more difficult for drivers to intuitively use VUI applications. Current VUIs tend to focus on accomplishing tasks with pre-defined voice commands, and it is not easy for drivers to remember all the voice commands for every VUI application. If more flexible voice dialogues were possible, drivers could utilize the features of the in-vehicle systems through natural conversations and spend more time engaging in purposeful experiences while driving.

Recently, there has been a growing interest in voice-interactive AI agents that utilize large language models (LLMs). LLMs are rapidly being developed worldwide. Conversational AI agents and chatbots are among the most widely used use cases for LLMs. Enabling voice dialogues with various types of AI agents, such as simple chat, education, tourist information and vehicle-related consultations, or conversations with digital copies of oneself or family members, is expected to meet the diverse needs of drivers. Moreover, multimodal LLM technology that handles various types of data, such as images and videos, in addition to text, is also rapidly advancing. In-vehicle AI agents are also expected to have various uses of multimodal LLM, such as inputting in-vehicle camera data to provide traffic condition explanations and guide sightseeing.

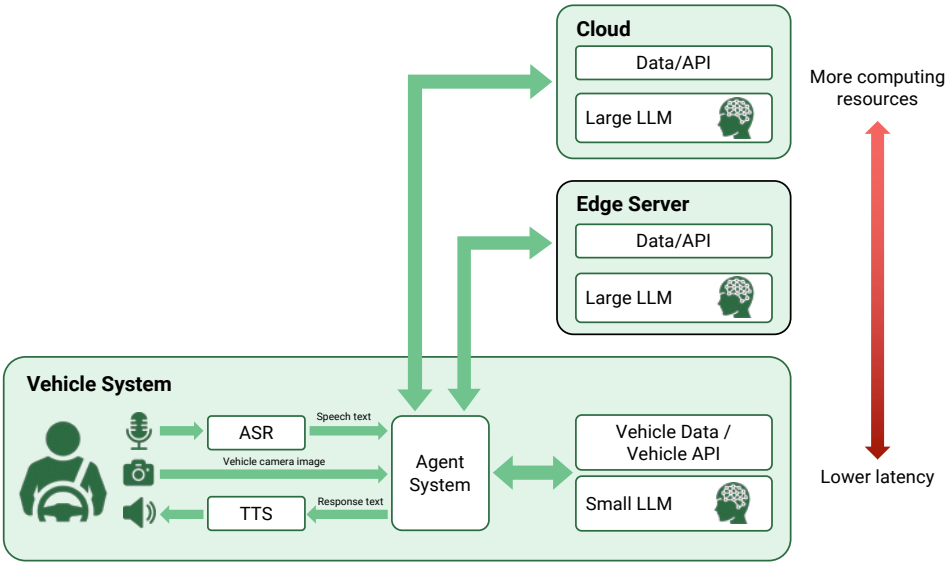


*Figure 8: Voice-interactive AI agent system utilizing multiple LLMs*

Figure 8 shows an overview of a voice-interactive AI agent system employing LLMs. The driver's spoken voice inputs, captured by a microphone, are sequentially converted to text using automatic speech recognition (ASR) and then fed into the agent system in the vehicle system. The agent system then creates prompts by interacting with databases and APIs as needed and inputs them into one of the LLMs. When the agent system receives a response from the LLM, it reflects the results to the databases and APIs if necessary, generates a response message for the driver and inputs it into a text-to-speech (TTS) engine to create audio data. This audio data is then played back through speakers as the response message.

LLMs tend to provide more accurate responses as the number of parameters in the model increases. However, this also means that the required computational resources and processing time increase. Therefore, it is necessary to appropriately choose where to execute the LLM based on the service requirements of the AI agent being used. Running an LLM on an in-vehicle system has the advantage of operability regardless of the network connection status. However, since in-vehicle systems have limited computational resources, it is necessary to use relatively lightweight LLMs. Even lightweight models of LLMs usually require tens of GBs or more in data size, so updating the model on an in-vehicle system involves significant data traffic.

Using LLMs on the cloud or edge servers allows for leveraging more substantial computational resources to employ larger models compared to in-vehicle systems. It also makes it easier to integrate with external cloud services and APIs. However, it has the disadvantage of requiring a network connection and experiencing increased response times due to communication delays.

The execution patterns for LLMs -- whether on an in-vehicle system, edge server or cloud -- have their individual pros and cons, which complement each other. Therefore, choosing the appropriate execution pattern for an LLM according to the service requirements of each AI agent and its configuration is essential for the efficient realization of an end-to-end LLM infrastructure.

## 3.6. Digital Twins

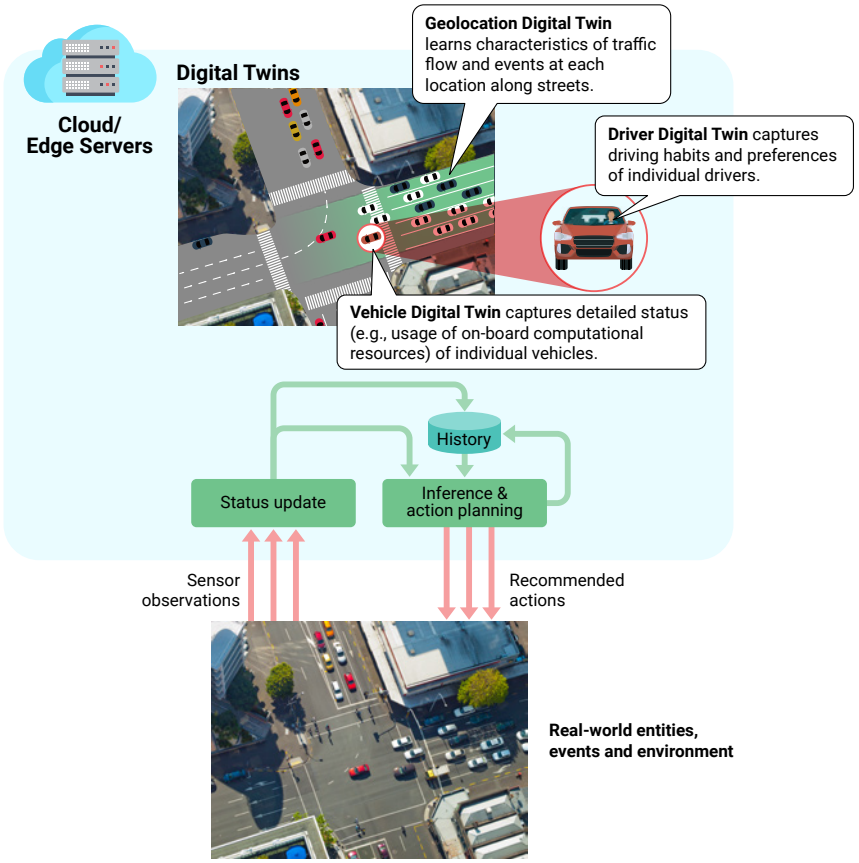


*Figure 9: Overview of Digital Twin*

Digital twins are high-fidelity digital representations of real-world entities, events and environments that are relevant to AECC mobility services (generally referred to herein as *observed entities*). Figure 9 shows the overview of the digital twin concept in the AECC context. The bottom part of the figure illustrates the real-world roadway environment. It consists of a myriad of observed entities, such as vehicles, drivers, pedestrians, bicyclists, road infrastructure, parking spots, potholes, etc. In some use cases, road traffic events (e.g., traffic jam, lane closure, etc.) or environmental factors affecting the road traffic (e.g., weather conditions) can also be considered as part of the observed entities. A digital twin is defined by a set of parameters, which represent the characteristics, state and behavior of their physical counterparts. Cloud and edge servers collect data from vehicles, road infrastructure and/or other information sources to keep the digital twins synchronized with the corresponding observed entities. Digital twins may also save the historical state of observed entities to maintain snapshots of the past.

One important aspect of a digital twin is its ability to perform inference based on the known status of observed entities to unveil the status of unobserved parts of the cities (e.g., roadways that are not monitored by any sensors).

It may also be possible to predict the future status of the physical environment by simulating interactions between digital twins, and even identify optimal actions the observed entities shall take in order to bring the real-world environment to a more desirable state. The results of optimization can then be fed back to the corresponding observed entities for them to take appropriate actions. The history of inference results and recommended actions may also be stored in cloud/edge servers for later analysis and model refinement.

Different types of digital twins would enable different varieties of services. Geolocation digital twins maintain road traffic flow and events at each location along streets (e.g., traffic jams, lane closure, roadworks, parking space occupancy, traffic light phases, etc.), which can be utilized to recommend routes, speeds and lanes that best facilitate smooth travel. Digital twins of drivers can learn their driving habits and preferences to deliver personalized adaptive cruising experience. Digital twins of vehicles may capture the status of their on-board computational resources, which can be used to coordinate collaborative computing among multiple vehicles over edge networks.

Depending on use cases, the status of observed entities may be updated at various levels of timeliness and resolution. To efficiently schedule computational tasks in the collaborative computing use case, digital twins of vehicles shall precisely capture the amount of computational resources available on each participating vehicle. At the same time, such microscopic status of individual vehicles may not be of importance in some other use cases like route guidance. In this case, geolocation digital twins may just keep track of aggregated statistics on vehicle traffic flow (e.g., average travel speed) along streets.

Digital twins impose new technical requirements, which can be summarized as follows:

- **Data Fidelity:** Digital twins must collect measurements of observed entities to update the corresponding digital entities in a timely manner. As digital twins aim at mirroring the entire AECC ecosystem lying across wide geographical areas, the measurements should be collected from millions of connected vehicles, road infrastructure and many other information sources. Some of the measurements may be incomplete, corrupted or erroneous due to sensor noise, occlusion, environmental factors and processing/communication failures. Aggregation and curation of data at edge servers therefore play a vital role to ensure timely synchronization of observed entities and their digital counterparts, while keeping the trustworthiness of the data constituting digital twins.
- **Just-in-time Inference and Action Planning:** Timely inference and action planning are also of critical importance for digital twins. It is especially true when providing a service to vehicles on the move, as roadway environment around the vehicles may change in a short period of time. Too much latency in inference and action planning would easily make the recommended actions stale before being delivered back to vehicles, causing the risk of misleading drivers. The latency reduction, though, should not sacrifice the accuracy of inference as it would otherwise degrade the quality of action plans recommended.

# 3.7. Extended Services

## 3.7.1. Mobility as a Service

Many route navigation services rely on mobility data from vehicles to provide real-time navigation. The data gathered can be used by third parties to offer new services, one example being traffic flow control by road authorities. These kinds of services are the building blocks of Mobility as a Service, which will bring improvement to mobility experiences. As these services evolve, there will soon be new emerging services beyond the current ones, such as mobility-sharing and multimodal navigation.

Mobility-sharing is a service that includes ride-sharing, car-sharing and even parking-lot-sharing. Multimodal navigation services provide end-to-end route guidance that uses various modes of transportation and also provides mobility sharing services information. Mobility-sharing services will involve various types of information being shared in a timely manner between asset owners, service providers and end users; accordingly, these types of services should be built on top of intelligent driving, high-definition maps and cruise assist.

### 3.7.2. Finance and Insurance

Auto insurers are adopting the usage-based-insurance model by monitoring driving habits, including driving behavior, how often people drive and the times of day during which they drive. By doing so, insurers will be able to better assess the customer's risk level, which will lead to a more reasonable cost for the insurers. In a future world where real-time information can be provided to users, real-time dynamic insurance premiums will be a possible product.

Data gathered from both the vehicle, such as cruising data, and the driver's condition is processed and is delivered back to the users in the form of insurance premium information in real time. Drivers will be encouraged to drive more safely at all times, as this will lead to their eligibility for lower premiums.

Distributed computing on localized networks is expected to be useful for this service, as there will be a huge amount of data from several sources that must be processed quickly to be able to provide users with insurance premium information in real time.

### 3.7.3. In-vehicle Experience Homogenization

Vehicle systems typically have a lifespan of 10 to 15 years, while cell phones and tablet computers have a lifespan of five years or fewer. While the hardware present in both vehicle systems and cell phones (usually) does not change within the lifespan of the device, software will typically be updated on a more frequent basis.

In-vehicle software can, in some cases, be upgraded, but this may be constrained by the available computation capabilities within the vehicle or by the ability to perform the upgrade itself. By comparison, software-based services that are cloud-hosted have increased agility due to the deployment model, with some services being upgraded on a monthly or even daily basis.

Computation and data storage systems within a vehicle system could be upgraded at points in the vehicle's lifespan, but there is a range of challenges associated with such upgrades being conducted post-sale that will need to be addressed. These challenges are outside of the scope of this paper.

### 3.7.4. Green Mobility

The natural environment is the most important foundation on which everything is based, and there is an increasing need to prevent global warming and air pollution so as to achieve a sustainable society.

The trend toward electrification of automobiles is accelerating, and the development of charging stations to support the use of electric vehicles (EVs) and the supply of renewable energy are important issues. In addition, EVs are in the midst of technological innovation, which requires the collection of usage data and software updates at a high frequency, which in turn requires data distribution. This requires optimizing the three flows of mobility (road traffic), energy and data, and investing in them efficiently. In other words, it is believed that the digital optimization of mobility and energy, that is, green mobility, is required.

Figure 10 shows a conceptual diagram of green mobility, a three-layered structure in which the digital portion optimizes between mobility and energy. In the energy layer, there is the power grid for power generation and

high-voltage transmission, and the micro-grid for local distribution. The digital layer in the middle consists of the infrastructure (central office and data center) and the service platform that provides various functions such as digital traceability, mobility management and energy management. The mobility layer illustrates service scenarios based on the lifecycle of an electric vehicle, such as manufacturing, selling, using, reselling and recycling the vehicle.

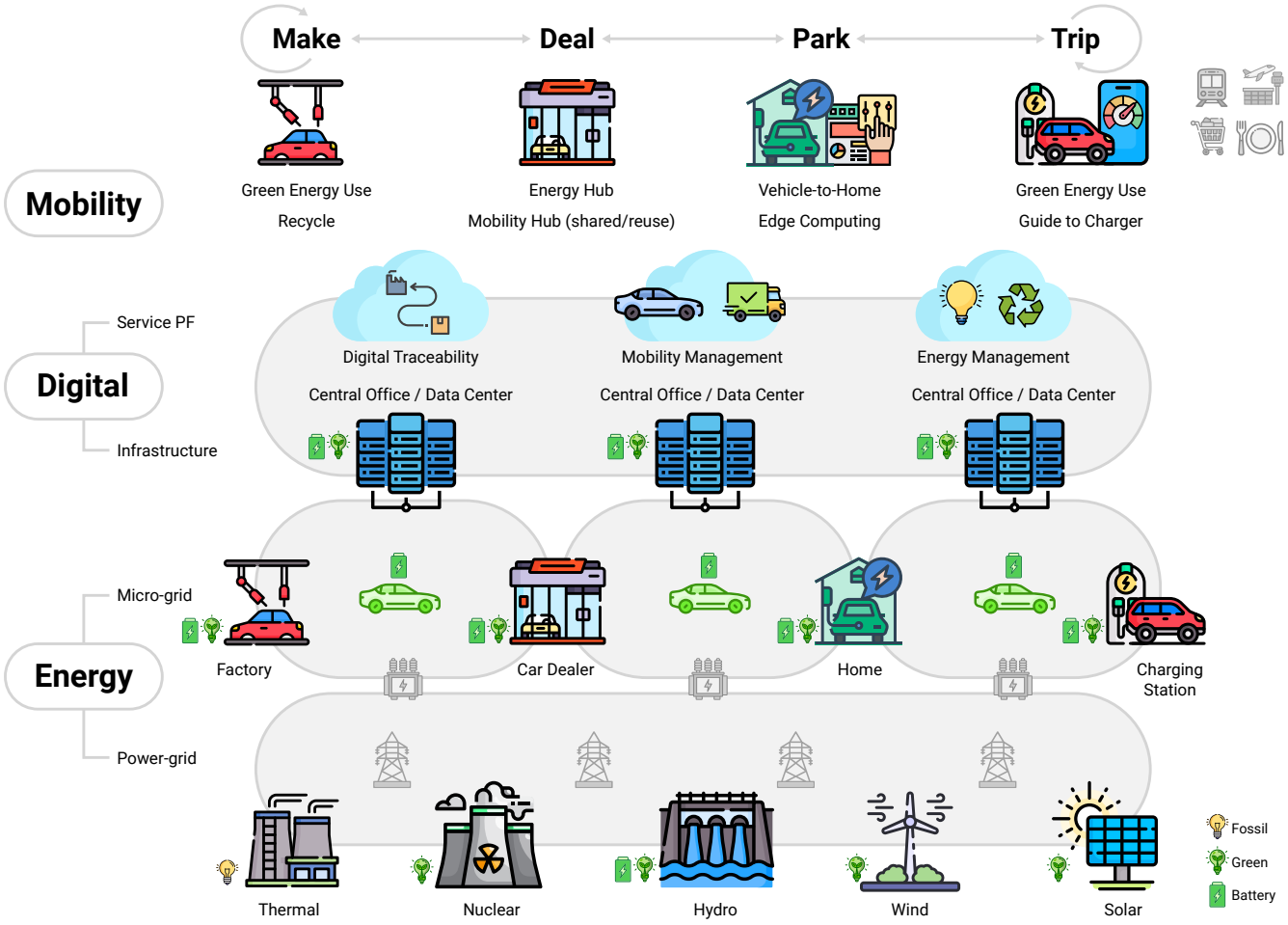


*Figure 10: Conceptual diagram of green mobility*

Here we will consider one significant service: charging guidance. The guidance should include a range of information, such as:

1. How to avoid running out of power
2. Identifying a charging spot where recharging can happen as quickly as possible
3. Identifying a spot that is not busy and has sufficient power load to spare
4. Identifying a spot where of the energy source is renewable. How to make a reservation for the proper time slot and how to encourage people to vacate their spots as soon as they have finished charging

To achieve these goals, it is necessary to use both mobility management and energy management functions of the digital layer in conjunction with each other. It will also be important to understand circumstances in real time,

to calculate an optimal or preferable solution in practical time, and to make suggestions to drivers to guide their behavior.

In this service scenario, we do not proudly list the distribution of vast amounts of data as a requirement, as in the video streaming example in Chapter 4, Service Requirements. However, the realistic business model of supplying energy to cars imposes very severe economic constraints, and efforts must be made to optimize costs as much as possible in the distribution and consumption of data, i.e., in the communication and computation infrastructures.

Since these are activities of local production for local consumption, this is an area where edge computing at charging stations can make a significant contribution. Digital traceability also plays an important role for various choices in the circular economy, such as the use of renewable energy in manufacturing and transportation, vehicle maintenance, reusing (reselling) and recycling. The evolution towards Green Mobility is one key area for the automotive industry and a focus for AECC going forward.

### 3.7.5. Data-driven Development Platform

Information services provided to users while driving are commonly referred to as telematics or infotainment, and they require a response in real time, 24/7, to user requests. The infrastructure that provides such services will be referred to here as the service delivery platform.

Mobility is expected to serve as a good conversational partner while driving and to support safe driving, allowing users to achieve what they want more freely and safely. The advancement of artificial intelligence is being considered for this purpose. To train the automotive software, including artificial intelligence, it is necessary to collect and analyze data obtained while the vehicle is in operation, and to process a large amount of data in accordance with the development schedule. The evolved software is then delivered to the car to activate its functions. The platform that holds software that will be trained by the collected data is called a data-driven development platform.”

The service delivery platform is required to receive requests at the timing desired by the end user, to respond without making the user wait and to achieve zero downtime of service. In contrast, the data-driven development platform needs to process a large amount of data in accordance with the schedule planned by developers and to do so cost-effectively. Therefore, when constructing the data-driven development platform using the current vehicle communication mechanism of the service delivery platform, it is necessary to design with careful attention to the different requirements of each.

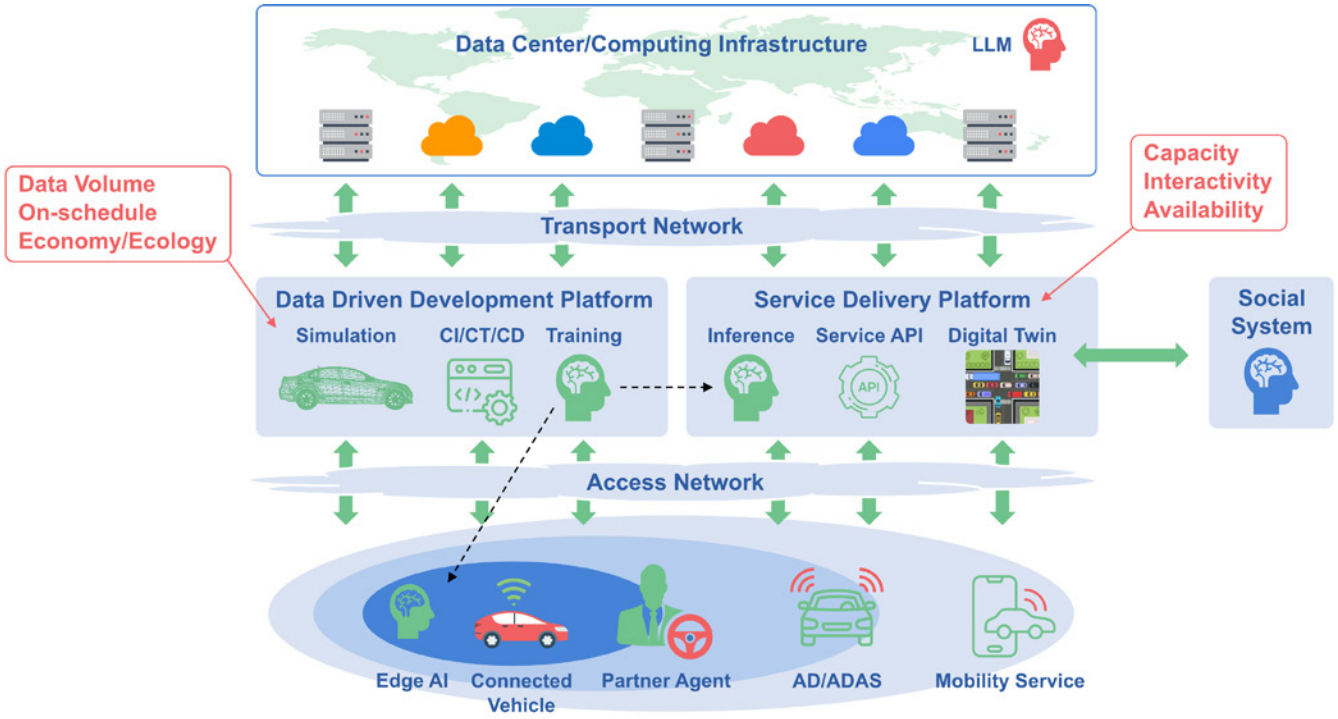


*Figure 11: The Introduction of the Data-driven Development platform*

Figure 11 illustrates the introduction of the data-driven development platform. The following should be taken into account due to the differences in requirements from the service delivery platform:

- Access Networks
  - Service delivery platform: because it needs to respond immediately to user requests anytime and anywhere, cellular communication is generally used. For future important services, extra fees could be charged for multiple cellular communications, satellite communication and priority communication.
  - Data-driven development platform: since it is sufficient to follow the product development schedule, a large amount of data could be exchanged with the vehicle cost-effectively by using more opportunistic access networks, such as home Wi-Fi.
- Data Center/Computing Infrastructure
  - Service delivery platform: Because it needs to accept requests at any user's timing, it is necessary to accommodate a number of transactions with room to spare. Also, to achieve zero downtime of service, servers need to be prepared for long power outages with generators, etc. It is difficult to prepare such servers in-house, and it is becoming common to use public cloud services.
  - Data-driven development platform: data can be collected in a planned and levelled manner according to the schedule defined by developers. However, it is necessary to design the locations of data and computing carefully in order to be compliant with regional data regulation and increase the use of local renewable energy. Considering such factors, it is necessary to effectively combine various regional computing resources.

The issues discussed above are just a fraction of those that need to be considered. It is conceivable that edge nodes, between endpoints such as cars and computing resources, could provide flexibility, diversity and scalability while optimizing. The AECC will continue to develop the optimal blueprint suitable for the generative AI era in software-defined vehicles.

# 3.8. The Challenge with Bringing Cloud-based Services to Vehicles

Cloud-based services will use a combination of the computing capabilities within the vehicle system and software executing outside of the vehicle system, hosted on a remote computing platform. Here the expected benefits include:

- Emerging services could be provided to vehicle systems even if there is insufficient in-vehicle computing capacity, due to the vehicle's age.
- The integration of computing could enhance the overall cost-effectiveness of the entire operation.
- Increased agility and efficiency of deployment could make the maintenance of the system easier.
- Bringing information together from multiple sources and performing analytics on that information may result in improved services.

The challenge we address is: how can the agility of cloud-based software services be brought to vehicle systems, particularly when the computing capabilities in the vehicle system will vary depending on the age and model of the vehicle? There are two basic strategies that can be used to mitigate the issue.

## 3.8.1. Software Updates

Software updates can be handled as part of the maintenance process performed by vehicle service centers, where updates are delivered through a designated communication interface. Some updates, especially those for infotainment features such as navigation systems and maps, may be delivered through media such as DVD-ROMs. Alternatively, updates can be delivered "over-the-air" (OTA) via Wi-Fi or cellular networks while the vehicle system is in normal operation. It should be noted that not all vehicle manufacturers support OTA updates and there may be restrictions as to the type of software that can be updated using OTA. For example, OTA updates may be applied to infotainment features but not to safety-critical capabilities. Moreover, OTA updates may be applied during parking via wireless network.

## 3.8.2. Edge Computing

For edge computing, edge servers should be included between vehicles and the cloud in order to reduce communication and computational load, to reduce the data volumes being transported for processing and, where necessary, to reduce latency.

**Application distribution and deployment**

Software updates will need to be distributed to the edge servers within the AECC system. Both the distribution process and the upgrading of running applications will need to be carefully orchestrated so that disruption to services used by the vehicle systems is minimized. This may also require coordination with updates to software running within the vehicle systems.

**Efficient use of server resources**

The application load on a specific server is expected to fluctuate over time and may suddenly spike as a result of incidents such as traffic jams and accidents. It is not practical to allocate resources on servers in order to handle

the maximum expected load of each application. It would be more efficient to be able to dynamically scale the resources assigned to a particular application when the load on the application of the server increases (or is predicted to increase). Further, it would be beneficial to be able to scale a particular application across available servers when required.

In order to realize this scenario, it is necessary to be able to take advantage of resource virtualization, dynamic allocation and optimization among the set of edge servers and center servers.

# 4. Service Requirements

Given the service scenarios described in the previous chapter, service requirements will include the following parameters.

- **Data Generation and Traffic Rates.** This refers to the amount of data generated inside vehicles and the amount of data transmitted between vehicles and the cloud. Vehicles are moving data sources that generate massive volumes of data, which results in heavy uplink traffic. This moving data source is characterized by its high mobility and not-always-on connectivity, which is quite the opposite of the present service requirements for smartphone and internet usage. This is the main requirement for determining the appropriate system architecture for handling the required data processing for services described in this document.
- **Response Time.** This refers to the response time between a vehicle and the cloud, including deviation with regard to latency. These requirements are critical for some of the service scenarios, including vehicle control based on real-time information (such as positions of other vehicles and pedestrians).
- **Availability-Cost Tradeoff.** Some services need less network availability; as a result, cost-effectiveness can be prioritized. Other services, on the other hand, require full cloud service availability regardless of the cost. These considerations call for more diverse network options to balance availability and cost.
- **Data Security and Privacy.** Some of the expected service scenarios include highly confidential data that must be authenticatable to maintain privacy and security. This mandates that the distributed network honor such requirements, with solutions that can give the appropriate security level while keeping service reliability.
- **Data Locality and Data Sovereignty.** The service needs to align with the rules and regulations regarding data locality and data sovereignty where the data is collected and processed. Compliance requirements for data hosting differ among countries. Depending on these requirements, data locality might differ between services and locations.
- **Service for Multiple Vehicle Systems via Multiple Cellular Networks.** In a future scenario, the cloud could be operated by the Mobility Service Provider (MSP) and could serve multiple vehicle systems via multiple cellular networks operated by different MNOs.

Table 1 shows the necessary requirements per service scenario.

| System Requirements* | Major Data Source | Data Generation in Vehicle | Target Data Traffic Rate | Response Time | | Required Availability | |
|---|---|---|---|---|---|---|---|
| | | | | Uplink | Downlink | Uplink | Downlink |
| **V2Cloud Cruise Assist** | Video stream | ~ 1215 EB Per month[1] | 1-10 EB per month in total** | < 10 second | < 10 second | Continuous | Continuous |
| **High-definition Map Generation & Distribution** | Still image (road surface image) | ~ 375 EB Per month[2] | | < 1 week | < 1 week | Occasional | Occasional |
| **Intelligent Driving** | ECU data | ~ 22.5 EB Per month[3] | | < 1 week | < 10 minutes | Occasional | Continuous |
| **Teleoperation** | Video stream | ~ 583 EB Per month[4] | | < 1 second | < 1 second[5] | Continuous | Continuous |
| **Digital twin** | Still image, ECU data | ~ 369 EB Per month[6] | 1 ~ 10 EB/ month, concurrent access from ~ 40 million vehicles[7], ~ 93Mbps/ vehicle[8] (single image upload), ~ 273Mbps/ vehicle[9] (multiple image upload) | < 10 seconds ~ < 1 week | < 10 seconds ~ < 1 week | Continuous/ Occasional | Continuous/ Occasional |

*Table 1 System Requirements*

* The numbers in V2Cloud cruise assist, high-definition map, intelligent driving, and digital twin are total values for 100 million connected cars, and those in vehicle teleoperation are for 2.5 million connected cars, respectively.

** Cost constraint might limit this number.

As indicated in the above table, some of the predicted performance requirements will be difficult for the current communication infrastructure to manage. Note that it will count more data from laser scanners, known as LIDAR, for outside situational awareness. Therefore, it is important to discover any missing links in the technology and

1 [Preliminary assumption] Video stream: 10Mpixel * 3Byte (Color) * 1/4 (Lossless JPEG) * 30FPS, average travel time: 30 min/day

2 [Preliminary assumption] Still image: 10Mpixel * 3Byte (Color) *1/4 (Lossless JPEG) at every 2 meters, average travel distance: 1,000km/month

3 [Preliminary assumption] Automotive Ethernet: 100Mbps *1/3 (effective), average travel time: 30 min/day

4 [Preliminary assumption] Video stream: 2MP *3Byte (Color) * 1/4 (Lossless JPEG) * 30FPS * 6 cameras * travel time 8h/day * 30 * 2.5 million connected cars

5 The response time in the uplink direction represents the duration for a single image from the onboard camera to be displayed in the cockpit system of the remote operator. The response time in the downlink direction represents the duration from the moment when the remote operator gives driving instructions to the cockpit system until they are executed on the vehicle.

6 [Preliminary assumption] Still images: 10Mpixel * 3byte (color) * 1/4 (lossless JPEG). Up to four images per second. Automotive ethernet: 100Mbps * 1/3 (effective). Average travel time: one hour/day.

7 [Preliminary assumptions] A vehicle makes four trips a day, and 10% of the trips are taken in the busiest hour of the day.

8 [Preliminary assumption] Single image upload: a vehicle may upload still images and ECU data concurrently. An image with the size of 10Mpixel * 3byte (color) * 1/4 (lossless JPEG) may be uploaded at the maximum frequency of one transaction per second. The ECU data is captured from Automotive ethernet at the typical data rate of 100Mbps * 1/3 (effective).

9 [Preliminary assumption] Multiple image upload: a vehicle may upload still images and ECU data concurrently. Four images with the size of 10Mpixel * 3byte (color) * 1/4 (lossless JPEG) each may be uploaded at the maximum frequency of one transaction per second. The ECU data is captured from Automotive ethernet at the typical data rate of 100Mbps * 1/3 (effective).

to find out how the technology is being deployed in order to realize the envisioned service scenarios, by analyzing the gap between the desired requirements and the existing technology and deployments.

# 5. Next Steps

This consortium will investigate cutting-edge technologies to fulfill the system requirements described in the previous chapter. These technologies should include flexible topology-aware distributed clouds with multi-operator edge computing capabilities, appropriate AI-enabling technologies, improved radio access technologies and other needed technologies. We aim to reveal the best practices in combining these potential technologies to create a provisional reference architecture for next-generation connected vehicles (see Figure 12).

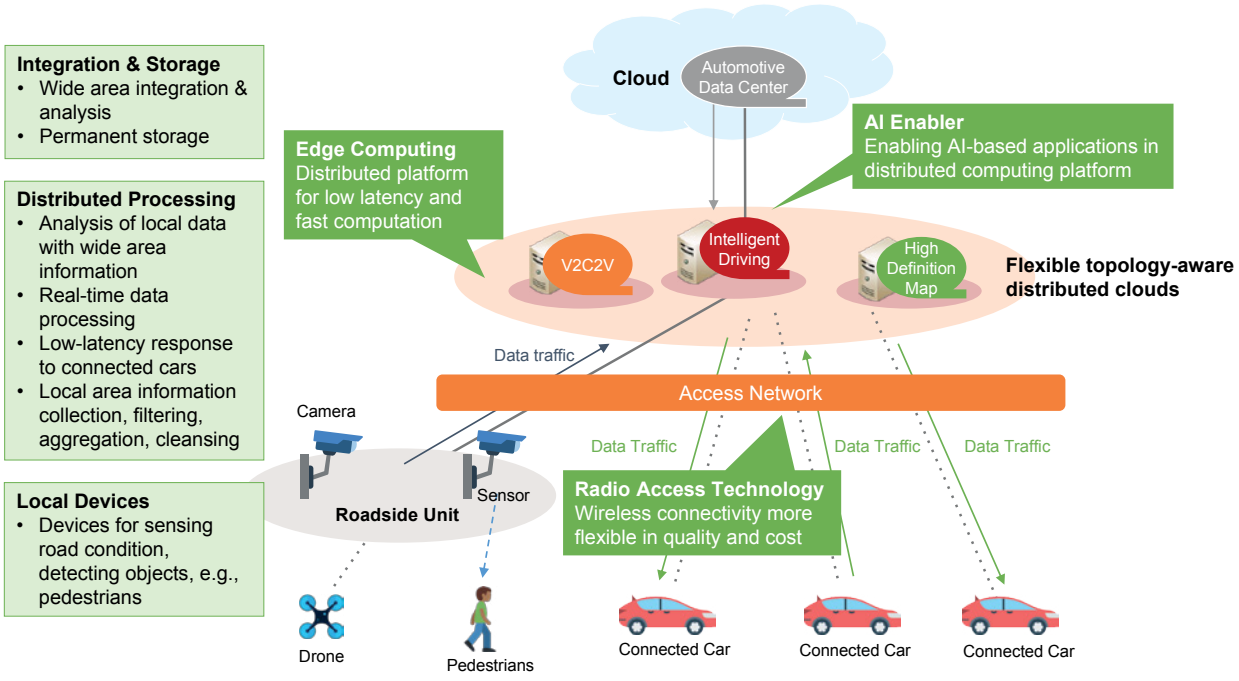


*Figure 12: Potential technologies*

- **Edge Computing.** Here, the computation resources are moved away from the central data center to be distributed further out in the networks, which means the hierarchically distributed non-central clouds where computation resources are deployed. As mentioned, edge computing technology is defined as distributed and even layered computing technology with localized networks, which involves challenges in both computing and networking. Our focus will be to ensure that the network infrastructure can be utilized to improve the characteristics of the indicated services, including the realization of real-time application response through a low-latency network environment and distributed computing.
- **AI Enabler.** Artificial intelligence technologies such as machine learning will implement required intelligence capabilities to support autonomous driving, cruising assist, creation of high-definition map information, etc., which requires big data and highly intelligent analysis. Our focus will be on technologies enabling such AI-driven services in distributed computing with localized networks.

- **Radio Access Technologies.** Wireless technologies will be used to connect a vehicle to distributed computing platforms with more flexibility in quality and cost. This includes not only cellular technologies but also local radio access such as Wi-Fi and low-power wide-area (LPWA).
- **ID Management for Multiple Vehicle Systems via Multiple Cellular Networks.** (To be described in the next version of this paper.)
- **Mobility Support.** (To be described in the next version of this paper.)
- **Data Transfer Preference.** (To be described in the next version of this paper.)

This investigation will help us in determining the necessary architecture and deployment to realize a distributed cloud for automotive use, based on the expected requirements for each service scenario and the technology concepts stated in this document.

The consortium will produce a strategic roadmap to introduce these new technologies to the existing infrastructure in order to realize our future vision. The roadmap will cover various aspects, including technology deployment as well as appropriate business schemes, charging models and multi-operator situations.

# 6. Summary and Conclusions

Network-based computation will make it possible for the next generation of automotive services to come into being. The expected service scenarios include intelligent driving, high-definition map generation, V2Cloud cruising assist and more. Autonomous vehicle services, which will require huge traffic volumes and low latency, will require flexible topology-aware distributed clouds with multi-operator edge computing capabilities.

In the concept of distributed computing on localized networks, several localized networks accommodate the connectivity of vehicles in their respective areas of coverage. Computation power is added to these localized networks so that they are able to process local data, enabling connected vehicles to obtain responses in a timely fashion. To realize the flexible topology-aware distributed clouds, edge computing is a key technology. For automotive use cases, edge computing technology will provide an end-to-end system architecture framework used to distribute computation processes from centralized networks to localized networks.

The Automotive Edge Computing Consortium will focus on increasing capacity to accommodate automotive big data in a reasonable fashion between vehicles and the cloud by means of edge computing technology and more efficient design of networks. The consortium will define requirements and develop use cases for emerging mobile devices, with a particular focus on the automotive industry, bringing them into standards bodies, industry consortia and solution providers. The consortium will also encourage development of best practices for the distributed and layered computing approach.

# 7. Terms and Definitions

| Term | Definition |
|---|---|
| **Cloud** | A logical server that hosts services to store, manage, and process data and which is composed of a set of remote servers accessed via the internet |
| **Central Cloud** | A central hardware or software platform provided by an Mobility Service Provider that supports mobility services |
| **Connected Vehicle** | Network attached vehicle that shares data with other network attached devices and servers |
| **Cruising Data** | Vehicle data about its movement |
| **Data Locality** | Where and how data should be stored and processed in the cloud space |
| **Data Sovereignty** | The handling procedures for data in accordance with the local jurisdiction's requirements |
| **Distributed Computing** | Computing that divides a problem into many tasks that can be served by many computers |
| **Edge Computing** | A type of distributed computing system where applications, memory and processing power are allocated to other computers in order to provide desired service levels |
| **Flexible Topology Aware Distributed Cloud** | A cloud solution that executes applications in a topology and geographically aware fashion, which means that the topology can be determined based on application requirements and the capability of the cloud instances to execute the application and handle its related data, according to the required cost and quality balance |
| **High-definition Map** | A topology representation with a high degree of precision and resolution<br><br>Note: High Definition Map is composed from a variety of source but a primarily intended for consumption by machine (hardware/software) systems rather than human beings. |
| **Intelligent Driving** | A service that augments an Advanced Driver Assistance System (or an Automated Driving System) with strategic decisions based on predictions of conditions along route alternatives that are gathered using vehicle connectivity to external sources |
| **Local Data Integration Platform** | The platform that integrates data on the localized network and the distributed computation |
| **Localized Network** | A local network that covers a limited number of connected vehicles in a certain area |
| **Multi-operator** | A resource (e.g. a network or computing platform) by multiple operators |
| **Telematics** | The technology of sending, receiving and storing information using telecommunication devices to control remote devices or to provide a service |
| **V2Cloud** | Communication between a vehicle and applications or services hosted on a cloud |

# 8. References

[1] GSMA: "The Connected Vehicle Opportunity," January 2021 (https://www.gsma.com/iot/wp-content/uploads/2021/01/Infographic-The-Connected-Vehicle-Opportunity.pdf)

[2] PwC: "Digital Auto Report 2023," 2023 (https://www.strategyand.pwc.com/de/en/industries/automotive/digital-auto-report.html)

[3] SBD, "Connected Car Forecast" (https://insight.sbdautomotive.com/rs/164-IYW366/images/536%20Report%20Preview%20-%20Connected%20Car%20Forecast.pdf)

[4] GSMA Intelligence, February 2021 (https://www.gsmaintelligence.com/data/)

[5] McKinsey & Company: "Unlocking the full life-cycle value from connected-car data," 2021 (https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/unlocking-the-full-life-cycle-value-from-connected-car-data)

# Appendix 1: Document Versions

| Version | Change | Date |
|---|---|---|
| 3.0.0 | Initial version | 2020-Jan-30 |
| 4.0.1 | Add §3.4.4 "Green Mobility" | 2022-Sep-22 |
| 4.0.2 | Add §3.4.5 "Teleoperation"<br>Add teleoperation into terminology<br>Update Table 1 with teleoperation parameters and footnotes | 2023-March-06 |
| 4.0.3 | Added §3.5 Digital Twins | 2024-January-08 |
| 4.0.4 | Update section 1.1 Background<br>Add §3.5 Voice-Interactive AI Agents<br>Add §3.7.5 Data-driven Development Platform | 2024-October |