教育部「5G行動寬頻人才培育跨校教學聯盟計畫」
5G行動網路協定與核網技術聯盟中心

行動邊緣計算
可推廣教材模組

單元-04：行動邊緣計算服務情景 (Service Scenarios)

授課教師：萬欽德

國立高雄科技大學 電腦與通訊工程系

MEC: A Key Technology for 5G

- the emergence of new compute-intensive applications
- the vision of the Internet of Things (IoT)
- ➢ foreseen that the emerging 5G network will face an unprecedented increase in traffic volume and computation demands.
- However, end users mostly have limited storage capacities and finite processing capabilities
- How to run compute-intensive applications on resource-constrained users becomes a natural concern.
- Mobile edge computing (MEC), a key technology in the emerging 5G network, can optimize mobile resources.

MEC: A Key Technology for 5G

Mobile edge computing (MEC)

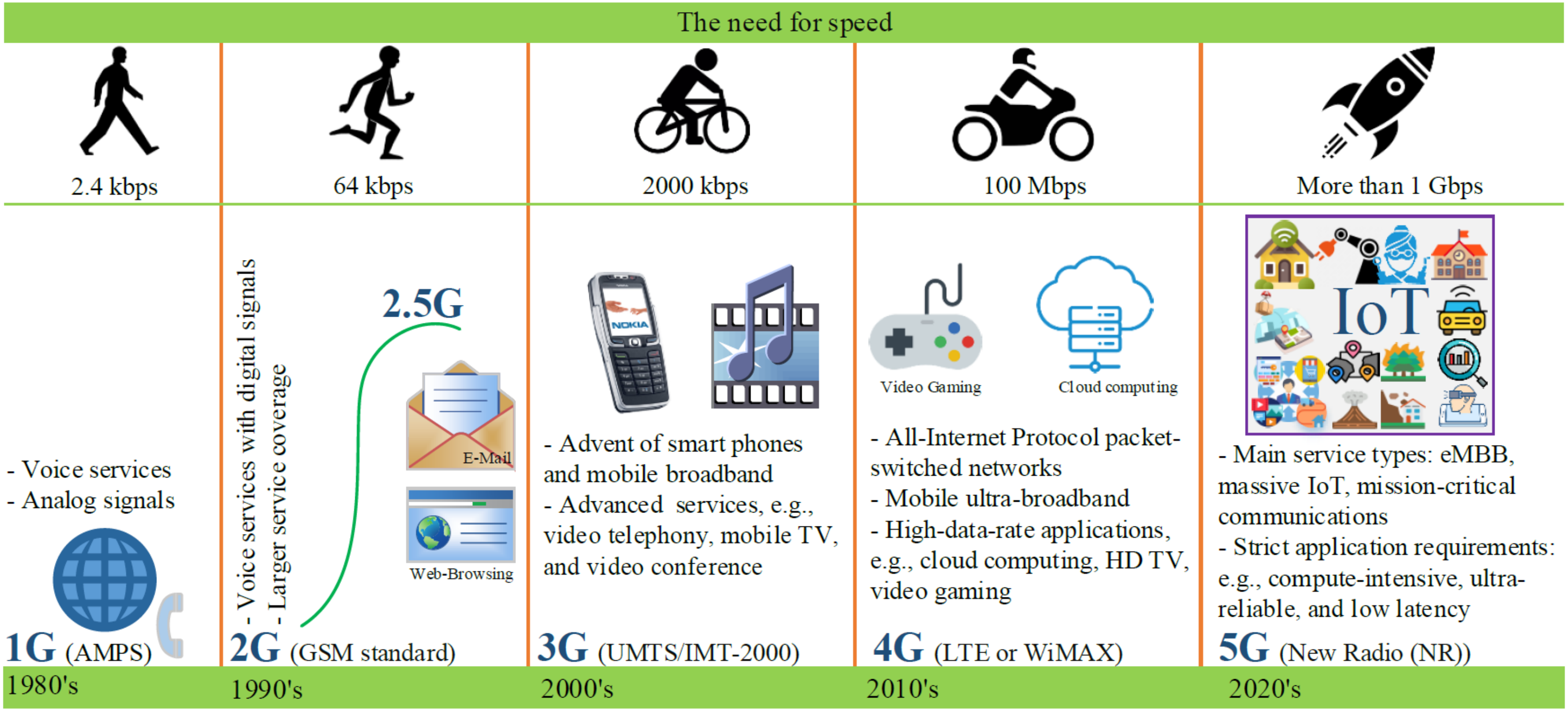
- a key technology in the emerging 5G network
  - optimize mobile resources by hosting compute-intensive applications
  - process large data before sending to the cloud
  - provide the cloud computing capabilities within the radio access network (RAN) in close proximity to mobile users
  - offer context-aware services with the help of RAN information

Three Categories of 5G

- 5G communications can be categorized into three categories:
  - enhanced mobile broadband (eMBB)
  - ultra-reliable low-latency communication (URLLC)
  - massive machine type communications (mMTC)
- Compared with previous generations, 5G will support not only <u>communication</u>, but also <u>computation</u>, <u>control</u>, and <u>content delivery</u> (4C) functions

Generations of Mobile Networks

- **1G**: Mobile voice calls
- **2G**: Mobile voice calls and SMS
- **3G**: Mobile web browsing
- **4G**: Mobile video consumption and higher data speed
- **5G**: Technology to serve consumers and digitalization of industries


The need for speed
2.4 kbps
64 kbps
2000 kbps
100 Mbps
More than 1 Gbps
- Voice services
- Analog signals
1G (AMPS)
1980's
- Voice services with digital signals
- Larger service coverage
2.5G
E-Mail
Web-Browsing
2G (GSM standard)
1990's
NOKIA
- Advent of smart phones and mobile broadband
- Advanced services, e.g., video telephony, mobile TV, and video conference
3G (UMTS/IMT-2000)
2000's
Video Gaming
Cloud computing
- All-Internet Protocol packet-switched networks
- Mobile ultra-broadband
- High-data-rate applications, e.g., cloud computing, HD TV, video gaming
4G (LTE or WiMAX)
2010's
IoT
- Main service types: eMBB, massive IoT, mission-critical communications
- Strict application requirements: e.g., compute-intensive, ultra-reliable, and low latency
5G (New Radio (NR))
2020's

Resources in Demand (1/2)

- Many new applications and use cases are expected with the advent of 5G
  - virtual/augmented reality (VR/AR)
  - autonomous vehicle
  - Tactile Internet
  - IoT scenarios
- Applications are poised to induce a significant surge in demand for communication and computation resources.

Resources in Demand (2/2)

- To meet the ever-growing demands, various technological concepts have been developed for 5G in terms of
  - radio access
  - network resource management
  - Applications
  - network architectures and scenarios
  - power supply
  - performance improvement

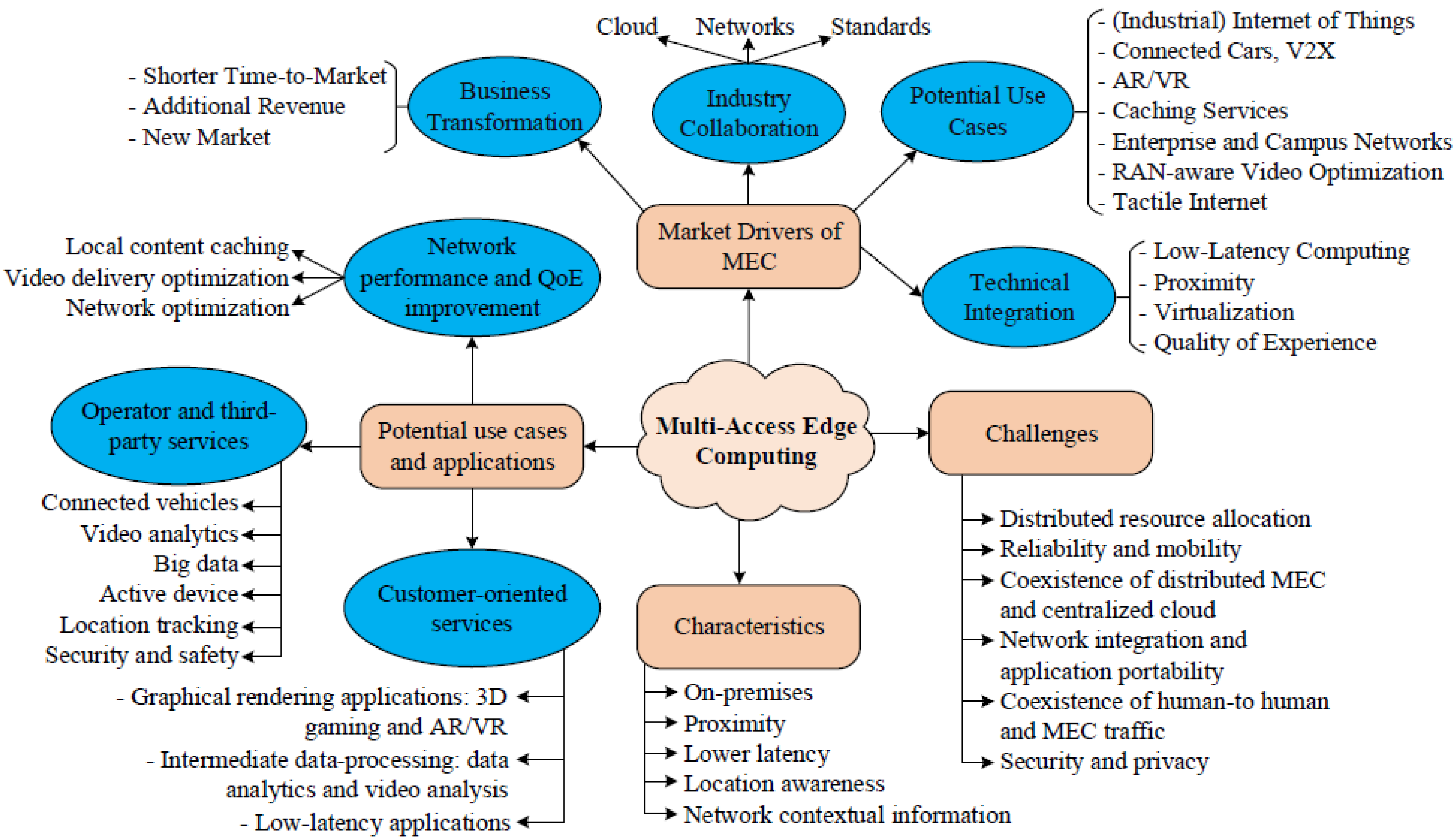
The Need of MEC

- Despite recent advancements in the hardware capability, mobile computing still cannot cope with the demand of many applications that need to generate, process, and store a massive amount of data and require large computing resources.
- One potential solution to these challenges is to transfer computations to centralized clouds (mobile cloud computing, MCC), which can be, however, burdened by many issues, such as network congestion and privacy policies.
- This has driven the development of mobile edge computing (MEC).

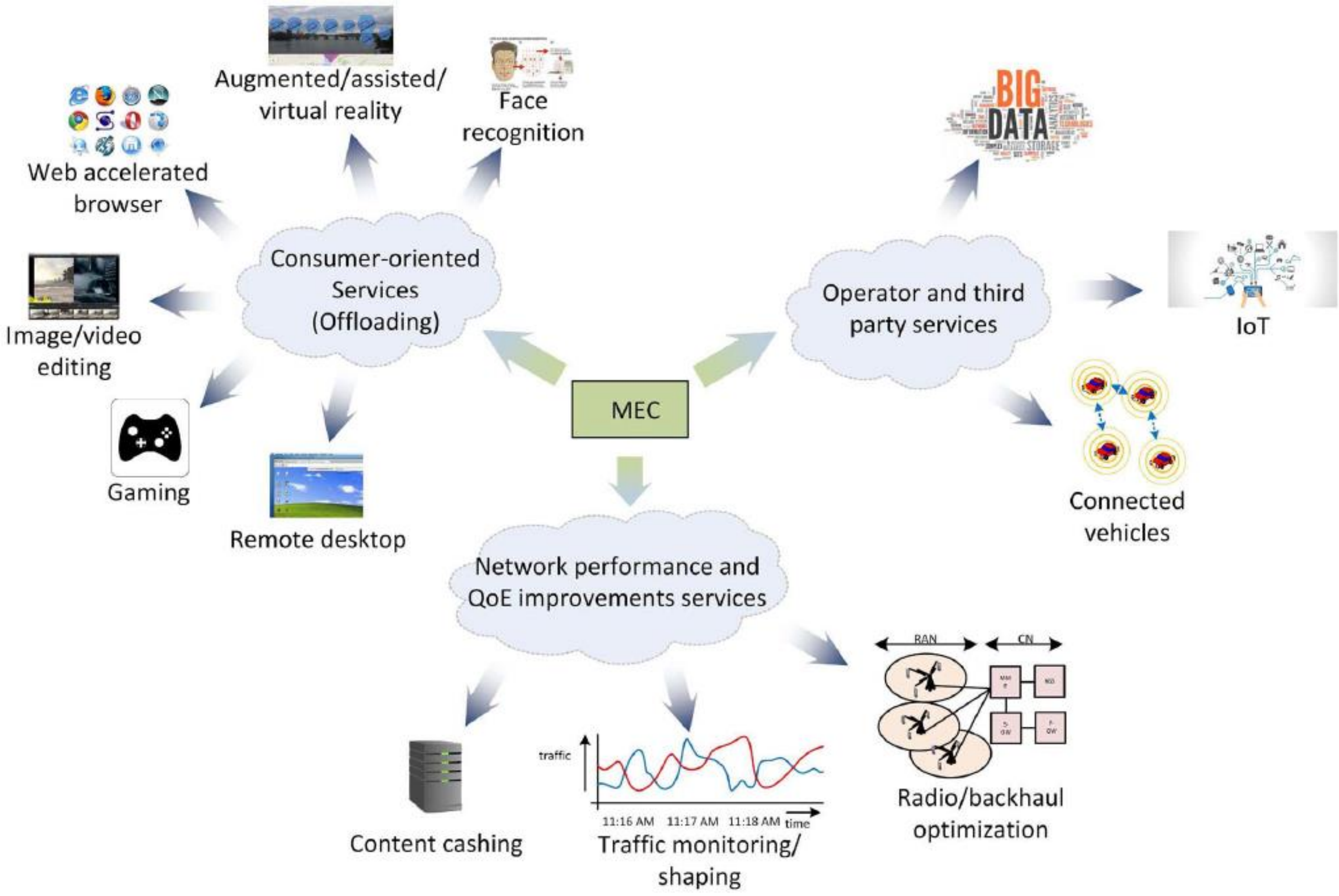
Comparison of MCC and MEC

Technical aspect	MCC	Edge computing
Deployment	Centralized	Distributed
Distance to the UE	High	Low
Latency	High	Low
Jitter	High	Low
Computational power	Ample	Limited
Storage capacity	Ample	Limited

Overview of MEC



Example of Use Cases and Scenarios for the MEC



Computation Offloading with the MEC

- A **decision on the computation offloading** to the MEC with the purpose to determine whether the offloading is profitable for the UE in terms of energy consumption and/or execution delay
- An efficient **allocation of the computing resources** within the MEC if the computation is offloaded in order to minimize execution delay and balance load of both computing resources and communication links
- **Mobility management** for the applications offloaded to the MEC guaranteeing service continuity if the UEs exploiting the MEC roams throughout the network

Key Service Scenarios

- Consumer-Oriented Services
- Operator and Third Party Services
- Network Performance and QoE Improvement Services

Consumer-Oriented Services (1/3)

- Beneficial directly to the end-users
- In general, the users profit from the MEC mainly by means of the computation offloading, which enables running new emerging applications at the UEs
- Example I: a Web accelerated browser, where most of the browsing functions (Web contents evaluation, optimized transmission, etc.) are offloaded to the MEC
- Example II: face/speech recognition or image/video editing are suitable for the MEC as these require large amount of computation and storage

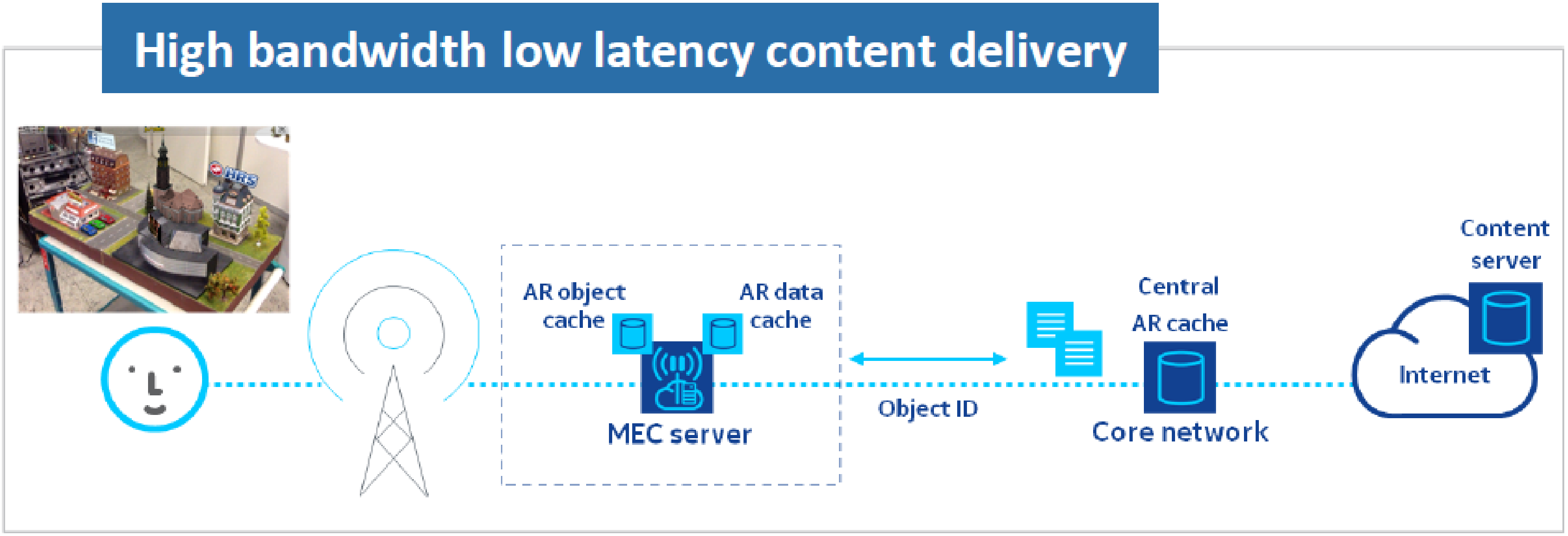
Consumer-Oriented Services (2/3)

- Computation offloading to the MEC can be exploited by the applications based on augmented, assisted or virtual reality.
- These applications derive additional information about users' neighborhood by performing an analysis of their surroundings (e.g., visiting tourists may find points of interest in his/her proximity).
- This requires fast responses, and significant amount of computing resources are not available at the UE.
- A real MEC testbed that the reduction of latency up to 88% and energy consumption of the UE up to 93% can be accomplished by the computation offloading to the MEC [Dolezal-2016]

Consumer-Oriented Services (3/3)

- Users running low latency applications, such as online gaming or remote desktop, may profit from the MEC in proximity.
- In this case a new instance of a specific application is initiated at an appropriate mobile edge host to reduce the latency and resources requirements of the application at the UE.

Consumer-Oriented Services: Augmented Reality (1/2)



Consumer-Oriented Services: Augmented Reality (2/2)

- The MEC application analyzes the output from a device's camera and the precise location
- Objects viewed on the device camera are overlaid with local augmented reality content
- Enables unique experience of a visitor to a museum or other (indoors or outdoors) points of interest
- Ensures low latency and high rate of data processing

Operator and Third Party Services (1/3)

- the services from which operators and third parties can benefit

Example: a gathering of a huge amount of data from the users or sensors

- Such data is first pre-processed and analyzed at the MEC
- The pre-processed data is then sent to distant central servers for further analysis
- This could be exploited for safety and security purposes, such as monitoring of an area (e.g., car park monitoring)

Operator and Third Party Services (2/3)

Example II: MEC for IoT (Internet of Thing) purposes

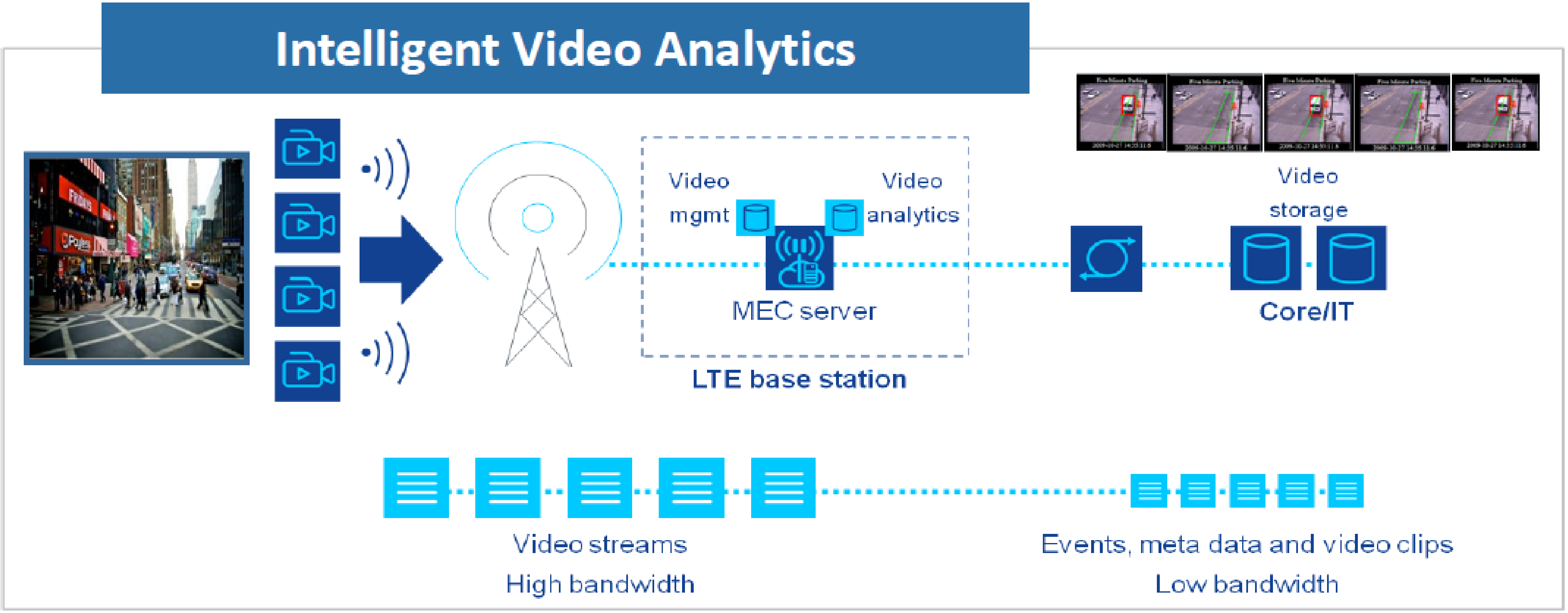
- IoT devices are connected through radio technologies (e.g., 3G, LTE, WiFi, etc.) using diverse communication protocols
- Low latency aggregation point is needed to handle various protocols, distribution of messages and for processing
- This can be enabled by the MEC acting as an IoT gateway
  - to aggregate and deliver IoT services into highly distributed mobile base stations
  - to enable applications responding in real time

Operator and Third Party Services (3/3)

Example III: MEC exploited for ITS

- to extend the connected car cloud into the mobile network
- roadside applications running directly at the MEC can receive local messages directly from applications in the vehicles and roadside sensors, analyze them and broadcast warnings (e.g., an accident) to nearby vehicles with very low latency.
- The exploitation of the MEC for car-to-car and car-to-infrastructure communications was demonstrated by Nokia and its partners in LTE network (2016)

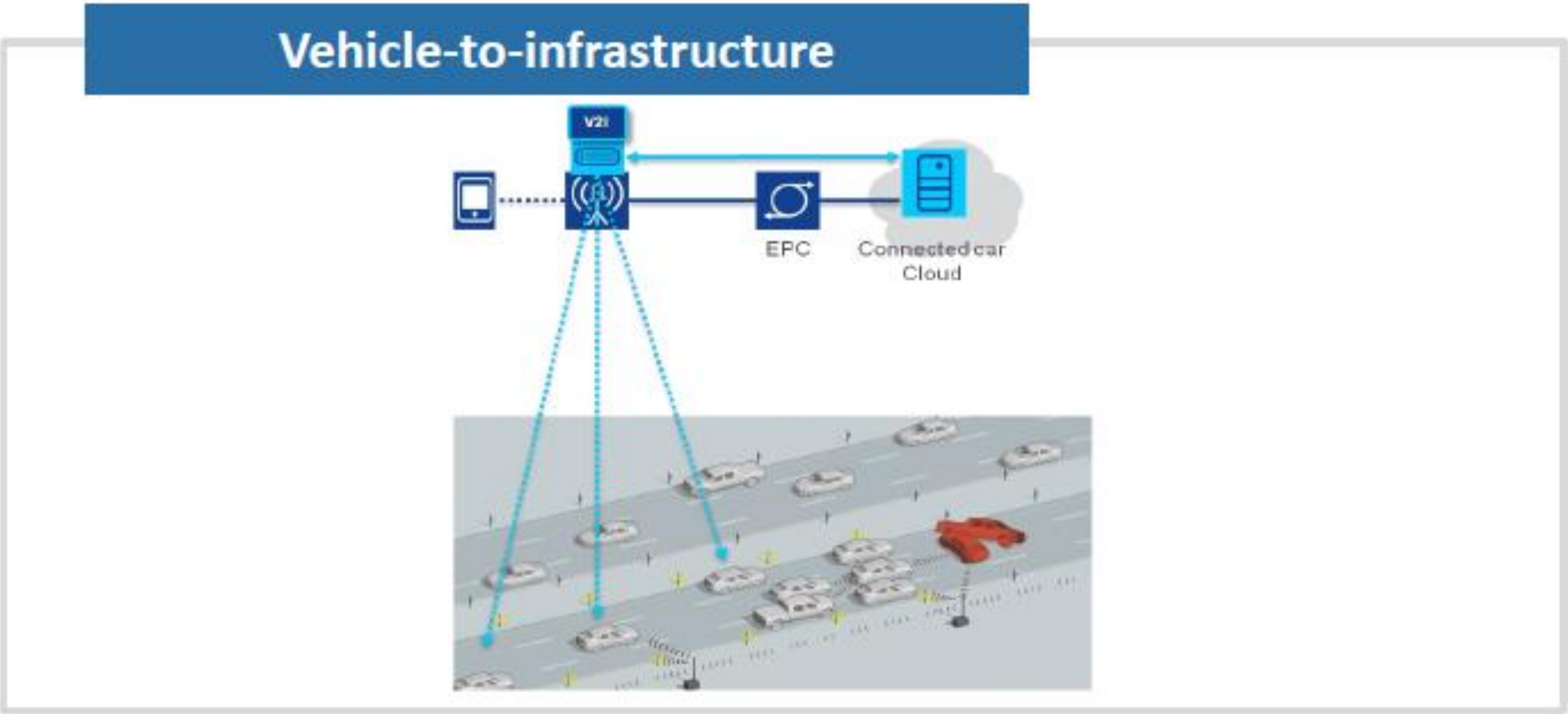
Operator and Third Party Services: Video Analytics (1/2)



Operator and Third Party Services: Video Analytics (2/2)

- Distributed live video streams analytics at the mobile edge
- Events are triggered automatically (e.g. movement, missing objects, crowd, etc.); enables fast detection and action triggering
- Optimizes backhaul and transport capacity
- Applicable to public safety, smart cities

Operator and Third Party Services: Connected Vehicles (1/2)



Operator and Third Party Services: Connected Vehicles (2/2)

- Existing cloud services are extended into the highly distributed mobile base station environment, leveraging the existing LTE connectivity.
- The MEC application operates as a roadside unit for vehicle-to-infrastructure (V2I).
- Road hazards can be recognized and warnings can be sent to nearby cars with extremely low latency.
- Enables a nearby car to receive data in a matter of milliseconds, and the driver to react instantly.

Network Performance and QoE Improvement Services (1/3)

- To optimize network performance and/or improve QoE

Example I: To enable coordination between radio and backhaul networks

- if the capacity of either backhaul or radio link is degraded, the overall network performance is negatively influenced since the other part of the network (either radio or backhaul) is not aware of the degradation.
  - an analytic application exploiting the MEC can provide real-time information on traffic requirements of the radio/backhaul network
  - an optimization application, running on the MEC, reshapes the traffic per application or re-routes traffic as required.

Network Performance and QoE Improvement Services (2/3)

Example II: to alleviate congested backhaul links

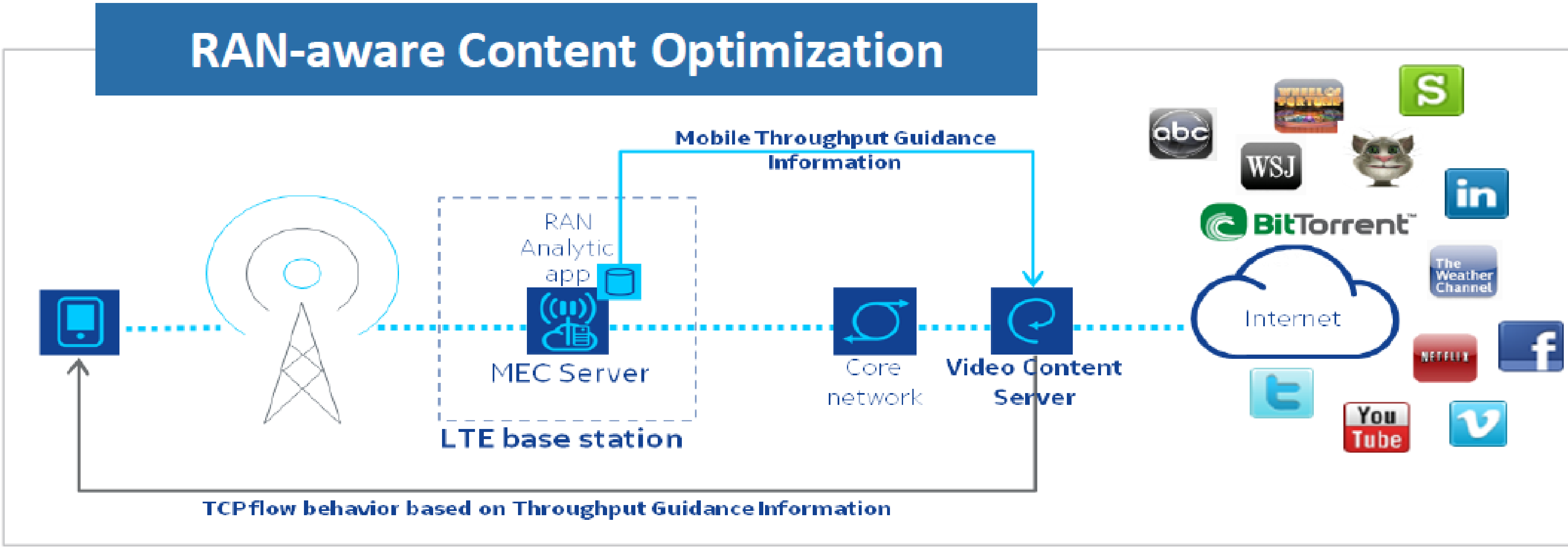
- to improve performance of the network by local content caching at the mobile edge
  - the MEC application can store the most popular content used in its geographical area
  - If the content is requested by the users, it does not have to be transferred over the backhaul network

Network Performance and QoE Improvement Services (3/3)

Example III: MEC for radio network optimization

- gathering related information from the UEs and processing these at the edge will result in more efficient scheduling.
- the MEC can be used for mobile video delivery optimization using throughput guidance for TCP (Transmission Control Protocols)
- The TCP has an inherent difficulty to adapt to rapidly varying condition on radio channel resulting in an inefficient use of the resources.
  - To deal with this problem, the analytic MEC application can provide a real-time indication on an estimated throughput to a backend video server in order to match the application-level coding to the estimated throughput.

Network Performance and QoE Improvement Services: Intelligent Video Acceleration (1/2)



Network Performance and QoE Improvement Services: Intelligent Video Acceleration (2/2)

- A Radio Analytics application provides the video server with an indication on the throughput estimated to be available at the radio downlink interface
- The information can be used to assist TCP congestion control decisions and also to ensure that the application-level coding matches the estimated capacity at the radio downlink.
- Enables improved video quality and throughput

MEC Scenarios and Services

- The basic characteristic of the edge computing paradigm is that the infrastructure is located closer to the end user.
- The scale of site distribution is high and that the edge nodes are connected by WAN network connections.
- Examining the scenarios in additional depth is helpful for evaluating current capabilities that map to the use case, as well as highlighting weaknesses and opportunities for improvement.

MEC Scenarios

- Retail/finance/remote location “cloud in a box”
- Mobile connectivity
- Network-as-a-Service (NaaS)
- Universal Customer Premises Equipment (uCPE)
- Satellite enabled communication (SATCOM)

Retail/finance/remote location "cloud in a box"

- Edge computing infrastructure that supports a suite of applications customized to the specific company or industry vertical.
- Often used by the enterprise, edge computing infrastructure, ultimately coupled together into distributed infrastructure, to
  - reduce the hardware footprint
  - standardize deployments at many sites
  - deliver greater flexibility to replace applications located at the edge (and to have the same application running uniformly in all nodes irrespective of HW)
  - boost resiliency
  - address concerns about intermittent WAN connections.
- Caching content or providing compute, storage, and networking for self-contained applications are obvious uses for edge computing in settings with limited connectivity.

Mobile connectivity (1/2)

- Mobile/wireless networks are likely to be a common environmental element for cloud edge computing
- Mobile networks will remain characterized by limited and unpredictable bandwidth, at least until 5G becomes widely available.
- The applications will all benefit from edge computing's ability to move workloads closer to the end user.

Mobile connectivity (2/2)

- Applications will all rely on the mobile network to greater or lesser degrees
  - augmented reality for remote repair and telemedicine
  - IoT devices for capturing utility (water, gas, electric, facilities management) data
  - Inventory
  - supply chain and transportation solutions
  - smart cities
  - smart roads
  - remote security applications

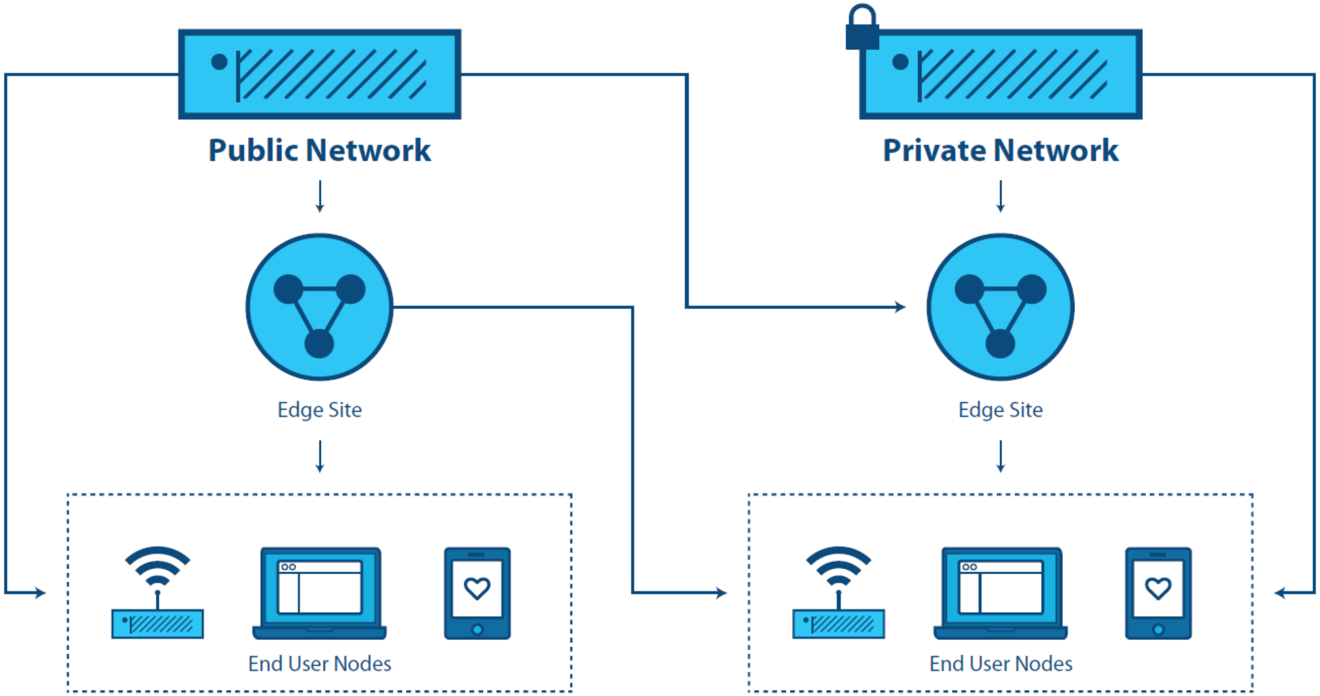
Network-as-a-Service (NaaS) (1/3)

- For the need to deliver an identical network service application experience in radically different environments, the NaaS use case requires both a small footprint of its distributed platform at the edges, and strong centralized management tools that cross over unreliable or limited WAN network connections in support of the services out on the edge.
- The main characteristics of this scenario:
  - small hardware footprint
  - moving (changing network connections) and constantly changing workloads
  - hybrid locations of data and applications

Network-as-a-Service (NaaS) (2/3)

- This case needs infrastructure to support micro nodes—small doses of compute in non-traditional packages.
- NaaS will require support for thousands or tens of thousands of nodes at the edge and must support mesh and/or hierarchical architectures as well as on demand sites that might spin up as they are needed and shutdown when they are done.
- APIs and GUIs will have to change to reflect that large numbers of compute nodes will have different locations instead of being present in the same data center.

Network-as-a-Service (NaaS) (3/3)



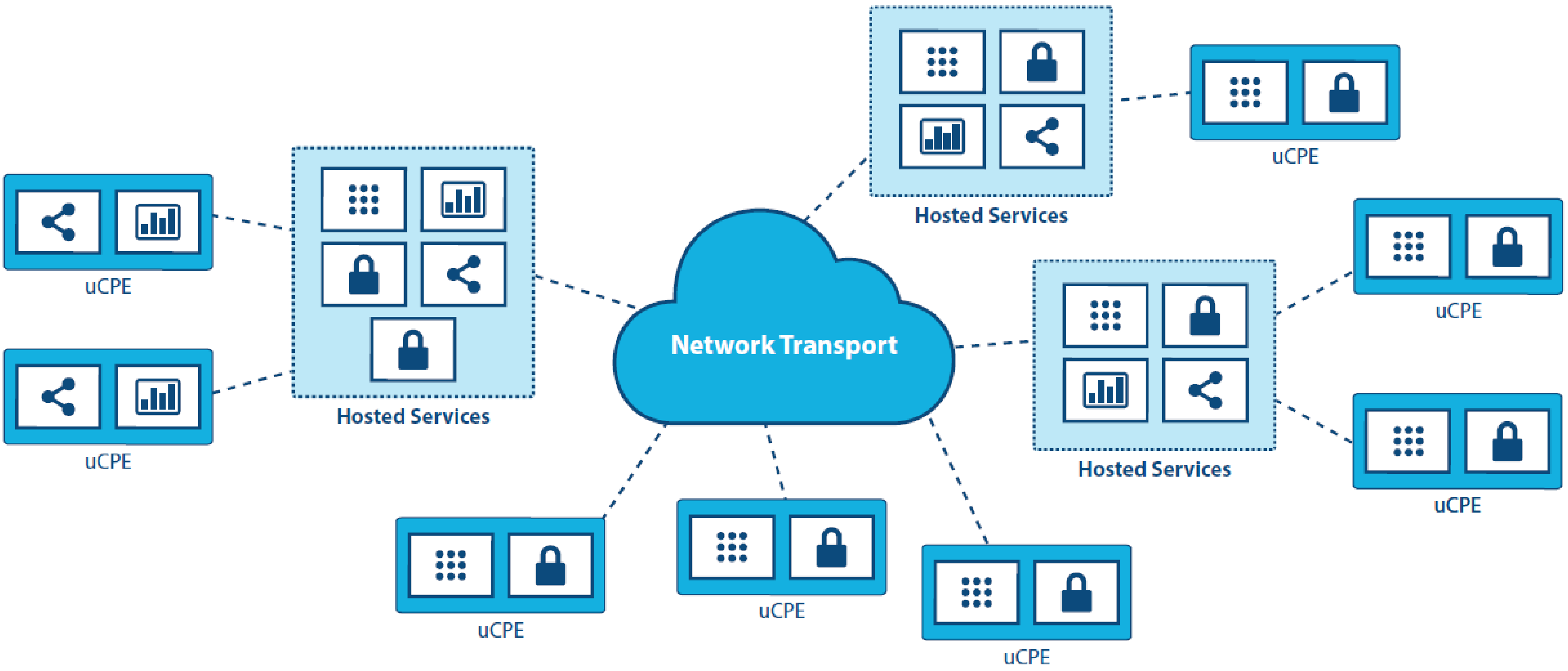
Universal Customer Premises Equipment (uCPE) (1/3)

- This scenario demands support for appliance-sized hardware footprints and is characterized by limited network connections with generally stable workloads requiring high availability.
- The scenario is applicable to NFV applications where different sites might need a different set of service chained applications, or sites with a different set of required applications that still need to work in concert.
- Mesh or hierarchical architectures would need to be supported with localized capacity and the need to store and forward data processing due to intermittent network connections.

Universal Customer Premises Equipment (uCPE) (2/3)

- It needs to support hybrid locations of data and applications across nodes, and scaling existing uCPE deployments will be an emerging requirement.
- Self-healing and self-administration combined with the ability to remotely administer the node are musts.

Universal Customer Premises Equipment (uCPE) (3/3)



Satellite enabled communication (SATCOM)

- This scenario is characterized by numerous capable terminal devices, often distributed to the most remote and harsh conditions.
- It makes sense to utilize the distributed platforms for hosting services, especially considering the extremely high latency, limited bandwidth and the cost of over-the-satellite communications.
- Specific examples of such use cases might include vessels (from fishing boats to tanker ships), aircrafts, oil rigs, mining operations or military grade infrastructure.

Challenges of Edge Computing Platform

- Foreseeable needs for MEC
  - A virtual-machine/container/bare-metal manager in charge of managing machine/container lifecycle (configuration, scheduling, deployment, suspend/resume, and shutdown).
  - An image manager in charge of template files (a.k.a. virtual-machine/container images).
  - A network manager in charge of providing connectivity to the infrastructure: virtual networks and external access for users.
  - A storage manager, providing storage services to edge applications.
  - Administrative tools, providing user interfaces to operate and use the dispersed infrastructure.

More Challenges on MEC

- Addressing storage latency over WAN connections.
- Reinforced security at the edge—monitoring the physical and application integrity of each site, with the ability to autonomously enable corrective actions when necessary.
- Monitoring resource utilization across all nodes simultaneously.
- Orchestration tools that manage and coordinate many edge sites and workloads, potentially leading toward a peering control plane or “self-organizing edge.”

More Challenges on MEC (cont'd)

- Orchestration of a federation of edge platforms (or cloud-of-clouds) has to be explored and introduced to the IaaS core services.
  - Automated edge commission/decommission operations, including initial software deployment and upgrades of the resource management system's components.
  - Automated data and workload relocations—load balancing across geographically distributed hardware.
- Some form of synchronization of abstract state propagation should be needed at the "core" of the infrastructure to cope with discontinuous network links.

More Challenges on MEC (cont'd)

- New ways to deal with network partitioning issues due to limited connectivity—coping with short disconnections and long disconnections alike.
- Tools to manage edge application life cycles, including:
  - The definition of advanced placement constraints in order to cope with latency requirements of application components.
  - The provisioning/scheduling of applications in order to satisfy placement requirements (initial placement).
  - Data and workload relocations according to internal/external events (mobility use-cases, failures, performance considerations, and so forth).

More Challenges on MEC (cont'd)

- Integration location awareness: Not all edge deployments will require the same application at the same moment. Location and demand awareness are a likely need.
- Discrete hardware with limited resources and limited ability to expand at the remote site needs to be taken into consideration when designing both the overall architecture at the macro level and the administrative tools. The concept of being able to grab remote resources on demand from other sites, either neighbors over a mesh network or from core elements in a hierarchical network, means that fluctuations in local demand can be met without inefficiency in hardware deployments.