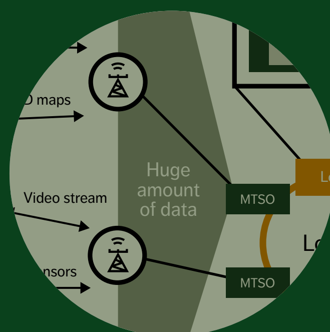
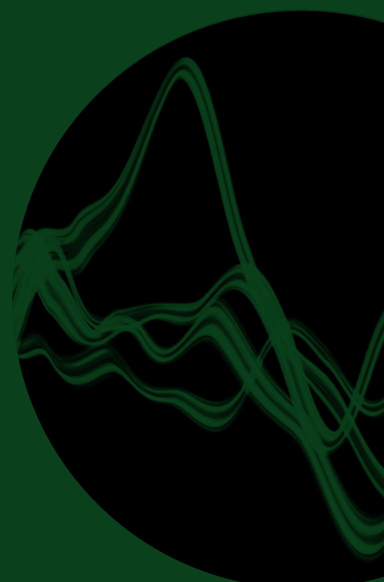
CHARTING THE FUTURE OF INNOVATION | #11·2018

Ericsson Technology Review





DISTRIBUTED CLOUD, AUTOMOTIVE AND INDUSTRY 4.0

ERICSSON

Distributed cloud

A KEY ENABLER OF AUTOMOTIVE AND INDUSTRY 4.0 USE CASES

Emerging use cases in the automotive industry – as well as in manufacturing industries where the first phases of the fourth industrial revolution are taking place – have created a variety of new requirements for networks and clouds. At Ericsson, we believe that distributed cloud is a key technology to support such use cases.

CHRISTER BOBERG, MALGORZATA SVENSSON, BENEDEK KOVÁCS

Both 4G and 5G mobile networks are designed to enable the fourth industrial revolution by providing high bandwidth and low-latency communication on the radio interface for both downlink (DL) and uplink (UL) data. Distributed cloud exploits these features, enabling a distributed execution environment for applications to ensure performance, short latency, high reliability and data locality.

■ Distributed cloud maintains the flexibility of cloud computing while at the same time hiding the complexity of the infrastructure, with application components placed in an optimal location that utilizes the key characteristics of distributed cloud. The automotive sector and many manufacturing industries already have use cases that make them very likely to be early adopters of distributed cloud technology.

Next-generation automotive services and their requirements

Mobile communication in vehicles is increasing in importance as the automotive industry works to make driving safer, smooth the flow of traffic, consume energy more efficiently and lower emissions. Automated and intelligent driving, the creation and distribution of advanced maps with real-time data, and advanced driving assistance using cloud-based analytics of UL video streams are all examples of emerging services that require vehicles to be connected to the cloud. These services also require networks that can facilitate the transfer of a large amount of data between vehicles and the cloud, often with real-time characteristics within a limited time frame while the vehicle is in active operation.

High data volume

Looking at the automotive industry, we often focus on the real-time use cases for safety, as defined by V2X/C-ITS (vehicle to everything/cooperative intelligent transport system), where real-time aspects such as short latency are the most significant requirements. However, the automotive industry's new mobility services also place high demands on network capacity due to the extreme amount of data that must be transported to and from highly mobile devices, often with near-real-time characteristics. Data needs to be transported within a limited time window (~30 min/day), with a varying geographical concentration of vehicles using a multitude of different network technologies and conditions.

The market forecasts that are generally referred to indicate that the global number of connected vehicles will grow to approximately 700 million by 2025 and that the data volume transmitted between vehicles and the cloud will be around 100 petabytes per month. At Ericsson, however, we anticipate that the automotive services of the near future will be much more demanding. We estimate that the data traffic could reach 10 exabytes or more per month by 2025, which is approximately 10,000 times larger than the present volume. Gartner recently raised the expectations further in its latest report (June 2018), estimating the volume to be as high as one terabyte per month per vehicle [1].

Such massive amounts of data will place new demands on the radio network, as the main part is UL data. New business models will be required, as a result of the high cost of handling massive amounts of data. As explained in the AECC (Automotive Edge Computing Consortium) white paper [2], the current mobile communication network architectures and conventional cloud computing systems are not fully optimized to handle all of this data effectively on a global scale. The white paper suggests many possible optimizations to consider – based on the assumption that much of the data could be analyzed and filtered at an early stage to limit the amount of data transferred.

Definition of key terms

- **Distributed cloud** is a cloud execution environment for applications that is distributed across multiple sites, including the required connectivity between them, which is managed as one solution and perceived as such by the applications.
- **Edge computing** refers to the possibility of providing execution resources (compute and storage) with the adequate connectivity (networking) at close proximity to the data sources.
- **The fourth industrial revolution** is considered to be the fourth big step in industry modernization, enabled by cyber-physical systems, digitalization and ubiquitous connectivity provided by 5G and Internet of Things (IoT) technologies. It is also referred to as Industry 4.0.

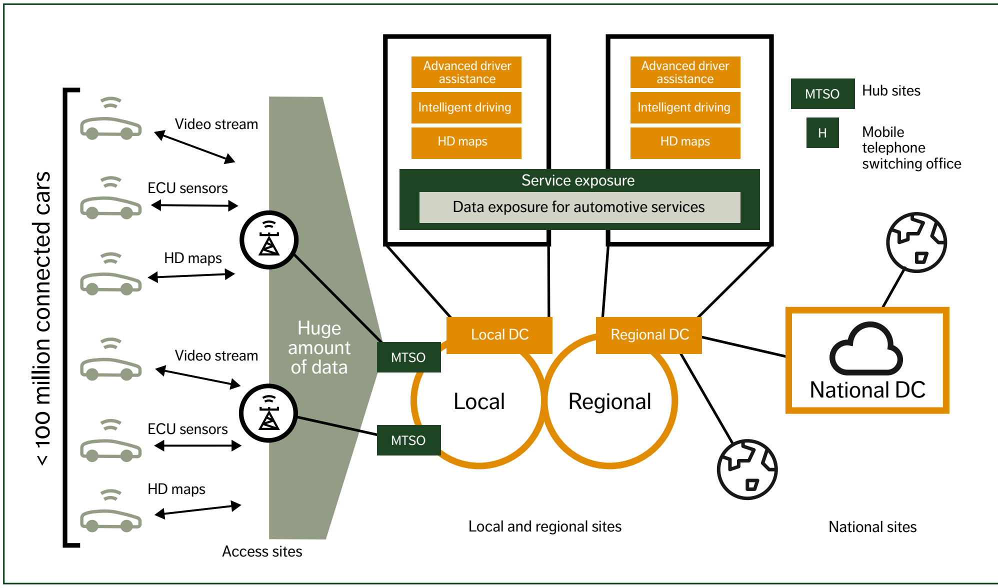


Figure 1 High-volume data automotive services and their characteristics

Topology-aware cloud computing and storage is an example of one such solution that provides what we call a global automotive distributed edge cloud. The limitation on the amount of data that can be effectively transported over the cellular network must not be allowed to affect the service experience negatively, as that would hinder the evolution of new automotive services. It is therefore necessary to increase capacity, availability and coverage as well as finding appropriate mechanisms to limit the amount of data transferred. Orchestrating applications and their different components running in a multitude of different clouds from different vendors is one of the challenges. Vehicles connecting to networks without an existing application edge infrastructure is another.

The placement of application components at edges depends on the behavior of the application and the available infrastructure resources. When dealing with highly mobile devices that connect to a multitude of networks, it must be possible to move execution of the edge application automatically when a more appropriate location for the vehicle is discovered. Some applications require transfer of previously analyzed data and findings to the new location, where a new application component instance will seamlessly take over to serve the moving vehicle.

Distributed computing on a localized network

We have developed the concept of distributed computing on a localized network to solve the problems of data processing and traffic in existing mobile and cloud systems. In this concept, several localized networks accommodate the connectivity of vehicles in their respective areas of coverage. As shown in *Figure 1*, computation power is added to these localized networks, so that they can process data locally. This reduces the total amount of data exchanged between vehicles and clouds while enabling the connected vehicles to obtain faster responses. The concept is characterized by three key aspects: a localized network, edge computing and data exposure.

A localized network is 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 clouds.

> INDUSTRY VERTICALS AND COMMUNICATION SERVICE PROVIDERS ARE DEFINING A SET OF NEW USE CASES FOR 5G

Edge computing refers to the geographical distribution of computation resources within the vicinity of the termination of the localized networks. This reduces the concentration of computation and shortens the processing time needed to conclude a transaction with a connected vehicle.

Data exposure secures integration of the data produced locally by utilizing the combination of the localized network and the 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. The amount of data that needs to be exchanged is kept to a minimum.

Private and local connectivity

As part of the fourth industrial revolution, industry verticals and communication service providers (CSPs) are defining a set of new use cases for 5G [3]. Private deployments and 5G networks provided by CSPs to manufacturing companies, smart cities and other digital industries are on the horizon as well. However, there are two main challenges to mobile network operators' ability to deliver. The first is the tough latency, reliability and security requirements of these new use cases. The second is figuring out how to shield the industries from the complexity of the infrastructure, to enable ease of use when programming and operating networks.

Secure private networks with centralized operations

Security and data privacy are key requirements for industrial networks. In some cases, regulations or company policies stipulate that the data must not leave the enterprise premises. In other cases, some or all of the data must be available at remote locations for purposes such as production analytics or emergency procedures. A typical industrial environment has multiple applications deployed and operated by different third parties. What this means in practice is that the same on-premises, cloud-edge instance that a factory already uses for business support and IT systems would also need to support the connectivity for its robots to interact with each other. As a result, there is a requirement of multi-tenancy for both the devices and the infrastructure.

Tactile internet and augmented reality

Augmented reality (AR) and machine learning (ML) technologies are widely recognized as the main pillars of the digitalization of industries [4], and research suggests that wide deployment of interactive media applications will happen on 5G networks. Many observers envision the worker of tomorrow as someone who is equipped with eye-tracking smart glasses [5] and tactile gloves rather than screwdriver sets [6]. Human-to-machine applications require low latency while demanding high network bandwidth and heavy compute resources. Running them on the device itself would result in high battery consumption and heat dissipation. At the same time, latency requirements do not allow the running of the complete application in large central databases due to the physical limits of light speed in optical fibers.

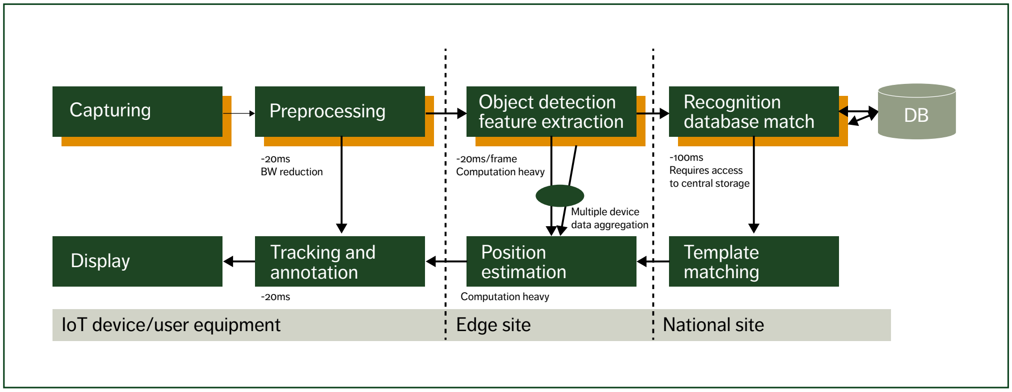


Figure 2 An AR application and its modules optimized for edge computing

A simple AR application and its main components are shown in *Figure 2*. The components of the application could be executed either on the device itself, the edge server or in the central cloud. Deploying application components at the network edge may make it possible to offload the device while maintaining short latency. Edge compute is also optimizing the flow when coordination is required – for example, when using multiple real-time camera feeds to determine the 3D position of objects, also as shown in Figure 2. Furthermore, advanced cloud software as a service – ML, analytics and DBs as a service, for example – may also be provided on the edge site.

Our distributed cloud solution

Ericsson has developed a distributed cloud solution that provides the required capabilities to support the use cases of the fourth industrial revolution, including private and localized networks. Our solution satisfies the specific security requirements needed to digitalize industrial operations, with automotive being one of the key use cases. Ericsson's distributed cloud solution provides edge computing and meets end-to-end network requirements as well as offering management, orchestration and exposure for the network and cloud resources together.

As shown in *Figure 3*, we define the distributed cloud as a cloud execution environment that is geographically distributed across multiple sites, including the required connectivity in between, managed as one entity and perceived as such by applications. The key characteristic of our distributed cloud is abstraction of cloud infrastructure resources, where the complexity of resource allocation is hidden to a user or application. Our distributed cloud solution is based on software-defined networking, Network Functions Virtualization (NFV) and 3GPP edge computing technologies to enable multi-access and multi-cloud capabilities and unlock networks to provide an open platform for application innovations. In the management dimension, distributed cloud offers automated deployment in heterogeneous clouds. This could be provided by multiple CSPs, where workload placement is policy driven and based on various externalized criteria.

To enable monetization and application innovation, distributed cloud capabilities are exposed on marketplaces provided by Ericsson, third parties and CSPs. The distributed cloud capabilities can be offered according to various business and operational models. One example of a possible scenario is for a CSP to offer connectivity and a cloud execution environment to enterprises as a service. In this case, a CSP manages the computation and connectivity resources, but these are located at the enterprise premises. The application characteristics determine the placement of applications at various geolocations. In the case of AR/VR and image recognition applications used by technicians to fix a broken power station, for example, it would be most effective to place them close to the broken power station.

Edge computing

Our distributed cloud solution enables edge computing, which many applications require. We define edge computing as the ability to provide execution resources (specifically compute and storage) with adequate connectivity at close proximity to the data sources.

OUR DISTRIBUTED CLOUD SOLUTION ENABLES EDGE COMPUTING, WHICH MANY APPLICATIONS REQUIRE

In the automotive use case, the network is designed to split data traffic into several locations 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. In this infrastructure framework, localized data collected via local and wide area networks is stored in the central cloud and integrated

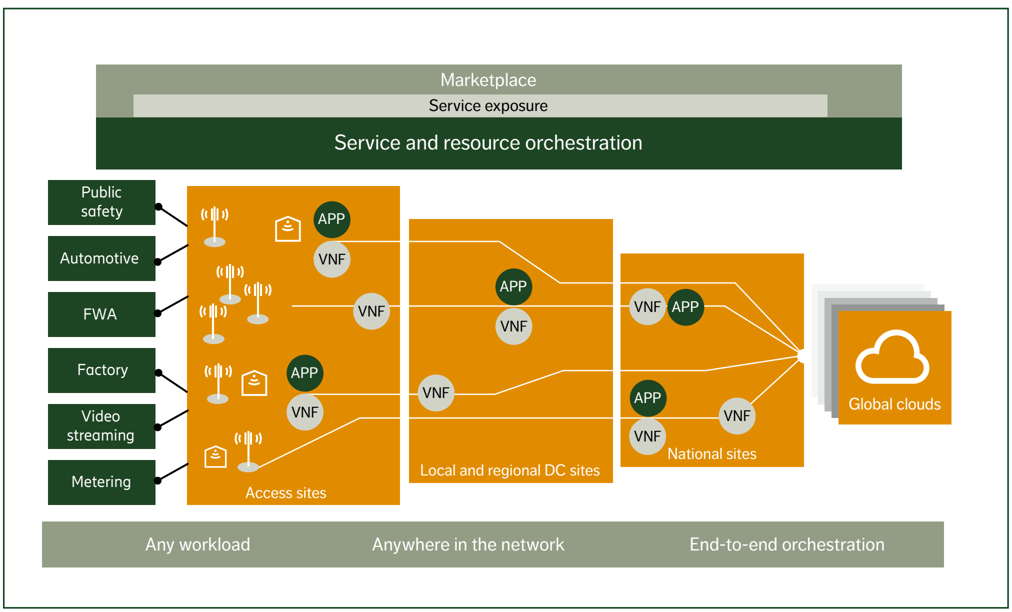


Figure 3 Distributed cloud architecture

> GLOBAL INDUSTRIES SUCH AS AUTOMOTIVE REQUIRE SOLUTIONS THAT WORK SEAMLESSLY FROM LOCAL TO GLOBAL SCALE

on the edge computing architecture to provide real-time information necessary for services of connected vehicles.

The exact locations of the micro and small data centers may be dependent on the CSP's network topology and the requirements of the use cases – this applies to central office sites, base stations and new DCs built on industrial sites. This infrastructure should be flexible, so that it is possible to start with a few sites and grow by adding new sites as required.

Management and orchestration

The distributed cloud relies on efficient management and orchestration capabilities that enable automated application deployment in heterogeneous clouds supplied by multiple actors. Figure 3 illustrates how the service and resource orchestration spans across distributed and technologically heterogeneous clouds. It enables service creation and instantiation in cloud environments provided by multiple partners and suppliers. Discovery, onboarding and auto-enrollment of edges are other important capabilities of distributed cloud management.

When deploying an application or a virtual network function (VNF), the placement decisions can be based on multiple criteria, where latency, geolocation, throughput and cost are a few examples. These criteria can be defined either by an application developer and/or a distributed cloud infrastructure provider, serving as input to the placement algorithm. Once a target cloud has been selected, the workload placement continues in any of the subordinated clouds.

In the automotive applications example, the placement decision could be made based on the geolocation of the moving car, availability of the computation resources and ability to meet regulatory requirements at the edges serving the moving car. Tactile internet and AR applications that are very sensitive to network latency while demanding high bandwidth and high computing power will be deployed at the edges that can fulfill the requirements.

The service orchestration manages the distributed cloud resources as well as the efficient distribution and replication of the applications that utilize the distributed cloud computation and connectivity resources. The service and resource management capabilities are also deployed in a distributed fashion to enable efficient management. For example, the scaling or data functions will be deployed close to the application they supervise.

Service exposure

The applications deployed in the distributed cloud will present their capabilities through the service exposure. With multi-dimensional exposure, each of the layers in the distributed cloud stack will expose its capabilities. The cloud infrastructure layer and the connectivity layer will expose their respective capabilities through the application programming interface(s) (API(s)), which will then be used by application developers of the industries making use of the mobile connectivity. By setting developer needs in focus, the exposed API(s) will be abstracted so that they are easy to use.

Evolution toward the global multi-operator distributed cloud

Global industries such as automotive require solutions that work seamlessly from local to global scale. In light of this, the evolution toward the global multi-operator distributed cloud is no trivial matter.

To be part of the globally distributed cloud, the edge clouds that CSPs provide at access and local sites must support a stringent set of functions and APIs. This implies that CSPs must join forces to create a federated model. Doing so will require significant effort, with the first step being to reach an agreement on the standard mechanisms to use.

The second step will be to gain industry acceptance for the mechanisms, before finally being able to implement the solutions and establish the business models.

One way to evolve the cloud edges that CSPs currently supply is to provide an environment above the current infrastructure that is homogeneous from a consumption perspective but discoverable through APIs and orchestrated in the same way as the CSPs' infrastructure. This would provide an intermediate step, where CSPs without an edge cloud infrastructure could become a part of the global scale distributed cloud. Following this approach, an industry actor could connect to any CSP access network as opposed to being limited to certain CSPs. While these networks will have the same functional scope, they will not be able to provide full edge characteristics. This will also serve as a catalyst for other CSPs to join the global scale distributed cloud. Otherwise, they will not become preferred suppliers.

Embracing industry initiatives and standardizations

We believe that the evolution toward the global multi-operator distributed cloud is dependent on a few key actions. First, we must take action to address the fact that the current mobile communication network architectures and conventional cloud computing systems are not designed, orchestrated or exposed in a way that can handle the industries' requirements effectively. We must scrutinize the system architectures and investigate network deployments and preferred profiling to better accommodate the outlined requirements. The architecture evolution will be driven by the relevant standardizations such as 3GPP and ETSI (European Telecommunications Standards Institute) NFV.

Secondly, we believe that it is critical to drive industry alignment by getting reference implementations of edge cloud software. This is why Ericsson has joined the industry collaboration project OPNFV (Open Platform for NFV) and ONAP (Open Network Automation Platform) [7], which provides the management capabilities of distributed cloud.

Finally, we believe that participating in ecosystems that provide the opportunity for interactions between the industries and vendors is critical to the evolution. This is particularly true for ecosystems that formulate requirements and ways of working, define use cases, agree on a common, easy-to-use reference implementation, and drive alignment in standardization bodies based on those implementations. Examples of such ecosystems are the AECC and 5GAA (the 5G Automotive Association) for automotive and 5G-ACIA (the 5G Alliance for Connected Industries and Automation) [8], Industry 4.0 and the IIoT (Industrial Internet of Things) for the fourth industrial revolution.

> “IT IS CRITICAL TO DRIVE INDUSTRY ALIGNMENT BY GETTING REFERENCE IMPLEMENTATIONS OF EDGE CLOUD SOFTWARE”

Conclusion

Distributed cloud is a cornerstone of the intelligent networks that will play a key enabling role in the fourth industrial revolution. A robust distributed cloud solution requires efficient and intelligent management and orchestration capabilities that span heterogeneous clouds supplied by multiple actors. Service exposure will enable monetization and application innovation through integration with the marketplaces and/or integration with the industries' IT systems.

The evolution toward globally distributed cloud requires action to align the industry both through traditional standardizations as well as active participation in open-source projects aimed at providing reference implementations. Ecosystems such as the AECC play an important role by examining the high-volume data use cases for the automotive industry.

References

1. **TelecomTV, Gartner says 5G networks have a paramount role in autonomous vehicle connectivity, June 21, 2018, available at:** *https://www.telecomtv.com/content/tracker/gartner-says-5g-networks-have-a-paramount-role-in-autonomous-vehicle-connectivity-31356/*
2. **AECC White Paper, available at:** *https://www.ericsson.com/res/docs/2014/consumerlab/liberation-from-location-ericsson-consumerlab.pdf*
3. **Government Technology, Making 5G a Reality Means Building Partnerships — Not Just Networks, June 5, 2018, Descant, S, available at:** *http://www.govtech.com/network/Making-5G-a-Reality-Means-Building-Partnerships--Not-Just-Networks.html*
4. **Wired, Eye tracking is coming to VR sooner than you think. What now?, March 23, 2018, Rubin, P, available at:** *https://www.wired.com/story/eye-tracking-vr/*
5. **Think Act, Digital factories – The renaissance of the U.S. automotive industry, Berger, R, available at:** *https://www.rolandberger.com/en/Publications/pub_digital_factories.html*
6. **Ericsson, Technology Trends 2018, Five technology trends augmenting the connected society, 2018, Ekudden, E, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2018/technology-trends-2018*
7. **Ericsson Technology Review, Open, intelligent and model-driven: evolving OSS, February 7, 2018, Agarwal, M; Svensson, M; Terrill, S; Wallin, J, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2018/open-intelligent-and-model-driven-evolving-oss*
8. **5G-ACIA, 5G for Connected Industries and Automation, April 11, 2018, available at:** *https://www.5g-acia.org/publications/5g-for-connected-industries-and-automation-white-paper/*

Further reading

- **Ericsson Consumer & IndustryLab, 5G business value: A case study on real-time control in manufacturing, April 2018, available at:** *https://www.ericsson.com/assets/local/reports/5g_for_industries_report_blisk_27062018.pdf*
- **Ericsson, Turn on 5G: Ericsson completes 5G Platform for operators, February 8, 2018, available at:** *https://www.ericsson.com/en/press-releases/2018/2/turn-on-5g-ericsson-completes-5g-platform-for-operators*
- **Ericsson, Going beyond edge computing with distributed cloud, available at:** *https://www.ericsson.com/digital-services/trending/distributed-cloud*
- **Ericsson/KTH Royal Institute of Technology, Resource monitoring in a Network Embedded Cloud: An extension to OSPF-TE, available at:** *https://www.ericsson.com/assets/local/publications/conference-papers/03-cloud-ospf-camera-ready.pdf*
- **M2 Optics Inc., Calculating Optical Fiber Latency, January 9, 2012, Miller, K, available at:** *http://www.m2optics.com/blog/bid/70587/Calculating-Optical-Fiber-Latency*
- **Application function placement optimization in a mobile distributed cloud environment, Anna Peale, Péter Kiss, Charles Ferrari, Benedek Kovács, László Szilágyi, Melinda Tóth, in Studia Informatica - Issue no. 2/2018, pp37-52, available at:** *http://www.studia.ubbcluj.ro/arhiva/abstract_en.php?editie=INFORMATICA&nr=2&an=2018&id_art=15974*

THE AUTHORS

The authors would like to thank the following people for their contributions to this article: Carlos Bravo, Ala Nazari, Stefan Runeson, Ola Hubertsson, Thorsten Lohmar and Tomas Nylander.

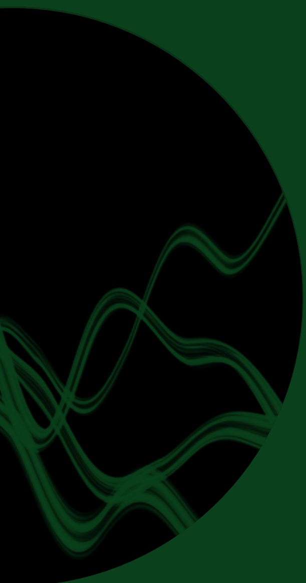
Malgorzata Svensson
◆ is an expert in operations support systems. She joined Ericsson in 1996 and has worked in various areas within research and development. For the past 10 years, her work has focused on architecture evolution. Svensson has broad experience in business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an M.Sc. in technology from the Silesian University of Technology in Gliwice, Poland.

Christer Boberg
◆ serves as a director at Ericsson's CTO office, responsible for IoT technology strategies aimed at solving networking challenges for the industry on a global scale. He initially joined Ericsson in 1983 and has in his career within and outside Ericsson focused on software and system design as a developer, architect and technical expert. In recent years, Boberg's work has centered on IoT and cloud technologies with a special focus on the automotive industry. As part of this work, he drives the AECC consortium together with industry leading companies.

Benedek Kovács
◆ joined Ericsson in 2005 as a software developer and tester, and later worked as a system engineer. He was the innovation manager of the Budapest R&D site 2011-13, where his primary role was to establish an innovative organizational culture and launch internal startups based on worthy ideas. Kovács went on to serve as the characteristics, performance management and reliability specialist in the development of the 4G VoLTE solution. Today he is working on 5G networks and distributed cloud, as well as coordinating global engineering projects. He holds an M.Sc. in information engineering and Ph.D. in mathematics from the Budapest University of Technology and Economics in Hungary.

Terms and abbreviations

AECC – Automotive Edge Computing Consortium | **API** – Application Programming Interface | **APP** – Application | **AR** – Augmented Reality | **BW** – Bandwidth | **CSP** – Communication Service Provider | **DB** – Database | **DC** – Data Center | **ECU** – Engine Control Unit | **ETSI** – European Telecommunications Standards Institute | **FWA** – Fixed Wireless Access | **IoT** – Internet of Things | **ML** – Machine Learning | **MS** – Millisecond | **MTSO** – Mobile Telephone Switching Office | **NFV** – Network Functions Virtualization | **UL** – Uplink | **VNF** – Virtual Network Function | **VR** – Virtual Reality | **V2X/C-ITS** – Vehicle-to-everything/ Cooperative Intelligent Transport System



ISSN 0014-0171
284 23-3323 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000