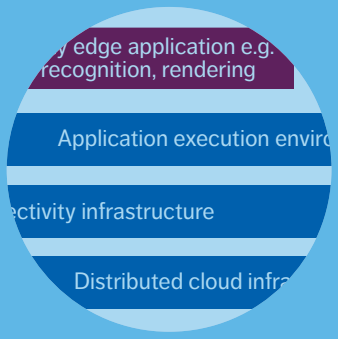
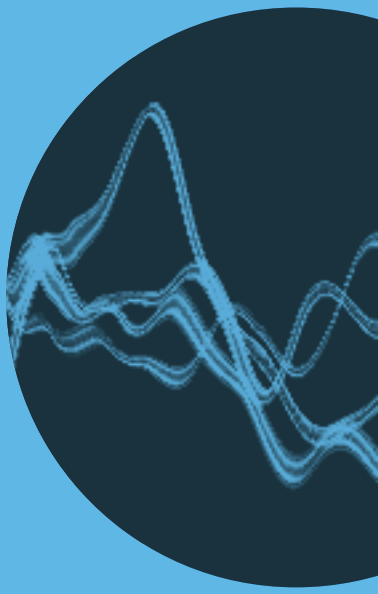
CHARTING THE FUTURE OF INNOVATION | #11·2019

ERICSSON TECHNOLOGY Review





NEXT-GENERATION EDGE-CLOUD ECOSYSTEM

ERICSSON

CREATING THE

Next-generation edge-cloud ecosystem

Edge computing has great potential to help communication service providers improve content delivery, enable extreme low-latency use cases and meet stringent legal requirements on data security and privacy. To succeed, they need to deliver solutions that can host different kinds of platforms and provide a high level of flexibility for application developers.

PÉTER SUSKOVICS, BENEDEK KOVÁCS, STEPHEN TERRILL, PETER WÖRNDLE

As well-established, trusted partners that already provide device connectivity, mobility support, privacy, security and reliability, the telecommunications industry and communication service providers (CSPs) more broadly have a competitive advantage in edge computing. This advantage is compounded by their ability to reach out globally to all edge sites with relative ease.

The main benefit of edge computing is the ability to move workloads from devices into the cloud, where resources are less expensive and it is easier to benefit from economies of scale. At the same time, it is possible to optimize latency and reliability and achieve significant savings in network communication resources by locating certain application components at the edge, close to the devices. To efficiently meet application and service needs for low latency, reliability and isolation, edge clouds are typically located at the boundary between access networks or on-premises for local deployments.

Since its invention a decade ago, edge computing has mainly been used to improve consumer QoE by reducing network latency and potential congestion points to speed up content delivery. It also lowers

operator costs by reducing peering traffic. Now, as a result of the surge in data volume that will come from the massive number of devices enabled by New Radio, the rollout of 5G has made edge computing more important than ever before.

Beyond its abilities to reduce peering traffic and improve user experience in areas such as video, augmented reality, virtual reality, mixed reality and gaming, edge computing also plays a key role in enabling ultra-reliable low-latency communication use cases in industrial manufacturing. It also helps operators meet stringent legal requirements on data security and privacy that are making it increasingly problematic to store data in a global cloud.

Edge-computing applications will have differing requirements depending on which driver has motivated them, and they will be built around different ecosystems that utilize platforms that may be ecosystem-specific. For example, the platforms and application programming interfaces (APIs) for smart manufacturing are different from those required for gaming and other consumer-segment-related use cases, which can be based on web-scale platforms and APIs. A robust edge-computing solution must be able to host platforms of different kinds and provide a high level of flexibility for application developers.

Key factors shaping the edge-cloud ecosystem

On top of being able to meet the requirements of emerging 5G use cases, there are other important factors to consider when designing an edge-computing solution, namely:

- Application design trends, life-cycle management and platform capabilities
- Expectations on management and orchestration
- Edge-computing industry status.

Application design trends, life-cycle management and platform capabilities

Cloud-native design principles have become a common design pattern for modern applications – both for telecom workloads [1] as well as other services. The modular, microservice-based architecture of cloud native applications enables significant efficiency gains and innovation potential when paired with an execution environment and a management system designed to handle cloud-native applications.

Reuse of generic microservice designs across different applications and enhanced platform services allows developers to focus on core aspects of the service with regard to quality and innovation.

Edge computing

Edge computing is a form of cloud computing that pushes the data processing power (compute) out to the edge devices rather than centralizing compute and storage in a single data center. This reduces latency and network contention between the equipment and the user, which increases responsiveness. Efficiency may also improve because only the results of the data processing need to be transported over networks, which consumes far less network bandwidth than traditional cloud computing architectures. The Internet of Things – which uses edge sensors to collect data from geographically dispersed areas – is the most common use case for edge computing.

Hyperscale cloud providers are extending their ecosystem toward the edge, and as part of the Industry 4.0 transformation enterprises are establishing use-case-specific development environments for their edge. The Cloud Native Computing Foundation [2] is gaining traction across all these development ecosystems, enabling portability of applications to private and public clouds.

> IT IS VITAL TO PLACE ONLY THE APPLICATIONS THAT WILL PROVIDE THE MOST BENEFIT AT THE EDGE

The increased amount of individual software modules and the demand to manage them efficiently implies the use of container technology to package and execute those software modules.

Kubernetes has become the platform of choice for container-based, cloud-native applications in both the telecom industry as well as for general-purpose services. Northbound management systems for telecom edge workloads as well as non-telecom edge workloads delegate some life-cycle management functionality to Kubernetes, thus reducing complexity in those management systems.

The Cloud Native Computing Foundation (CNCF) ecosystem has become a focal point for developers aiming to build modern, scalable cloud-native applications and infrastructure. Embracing a certified Kubernetes platform is the best way to become compatible with the CNCF ecosystem and thereby utilize the speed of innovation and variety of applications being developed.

Expectations on management and orchestration

The primary role of management and orchestration is to assure and optimize the application platform, 3GPP-defined connectivity, cloud infrastructure and transport, as well as ensuring the optimal placement of the edge application.

Put in the simplest terms, edge computing is an optimization challenge at scale that consists of several different aspects. The first is supporting the consumer experience by placing appropriate functionality – such as latency-sensitive applications – at the edge. The second aspect is ensuring that the users are connected to these applications. The third aspect is reducing the stress on transport resources and contributing to network efficiency by placing certain types of caching functions at the edge.

While it may seem ideal from a performance perspective to place all applications at the edge, edge resources are limited and prioritizations must be made. From an optimization perspective, it is vital to place only the applications that will provide the most benefit at the edge. Determining the best location for the management functionality – that is, the analytics functionality that can reduce traffic backhaul at the cost of local processing – is a critical aspect of the optimization process. In some cases, local deployment of the management functionality may be necessary to meet service continuity expectations.

A related consideration is the life-cycle management of the edge applications and the edge application platform, which must be efficiently onboarded from a central location, distributed and instantiated to the correct locations. The responsibilities for this can differ depending on the agreement between the edge application platform provider and the CSP. When deploying the edge applications and the edge application platform, appropriate connectivity to both the radio and the broader network must be established.

An edge-computing solution must be able to manage many distributed edge sites that each have their own needs based on local usage patterns. The massive scale that arises from this presents a multidimensional challenge. To overcome it, the

Terms and abbreviations

API – Application Programming Interface | **CNCF** – Cloud Native Computing Foundation | **CSP** – Communication Service Provider | **DNS** – Domain Name System | **IoT** – Internet of Things | **NFV** – Network Functions Virtualization | **ONAP** – Open Networking Automation Platform | **UPF** –User Plane Function | **VNF** – Virtual Network Function | **WAN** –Wide Area Network

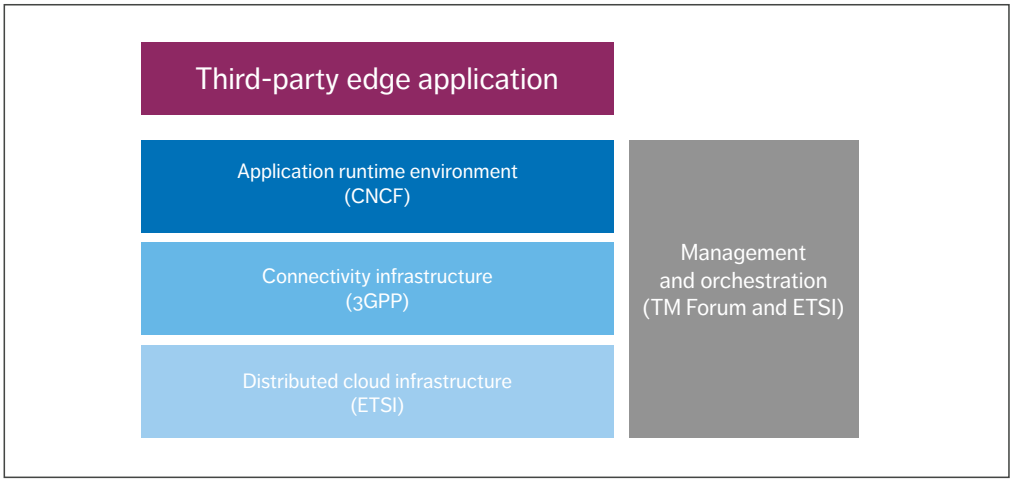


Figure 1 Relevant standardization and open-source forums

different domains of management and orchestration – ranging from hardware to virtualization infrastructure to radio and core network applications, together with edge-application platform orchestration – must all work together in an optimal manner.

Edge-computing industry status

Edge computing is dependent on functionalities in multiple domains. For example, the first step in application deployment is to ensure that runtime is available in the appropriate place, which puts requirements on the orchestration layer and placement capabilities, as well as on business interfaces.

Once the runtime is deployed, anchoring and connectivity are required to configure the necessary local breakout points and steer the traffic to where the edge runtime requires it. Most of these functionalities are not specific to edge computing and have either been addressed by industry standardization or open source. *Figure 1* presents the most relevant standardization and open-source forums for third-party edge applications.

On the networking side, the 3GPP has been addressing edge-computing requirements since release 14, both from the connectivity perspective as well as from a service and exposure perspective. Addressing edge computing under the 3GPP is the only guarantee to secure full compatibility with existing telecommunication network deployments and their future evolution [3].

In the implementation domain, ETSI (the European Telecommunications Standards Institute) Network Functions Virtualization (NFV) [4] defines the infrastructure, orchestration and management, while TM Forum leads the way for the digital transformation of CSPs.

When it comes to runtime and APIs, the fragmentation of the use cases is standing in the way of the vision of one runtime and one type of API. Some developers will use a widely adopted runtime like Kubernetes, especially its versions certified by the CNCF, or embrace web-scale platforms, while some verticals will probably develop, or set requirements on, their own platform and/or APIs. The 5G-ACIA (5G Alliance for Connected Industries and Automation) consortium [5] is one such example. A comparable initiative in the automotive sector is the AECC (Automotive Edge Computing Consortium) [6].

By utilizing standard components and telecommunication infrastructure that is already

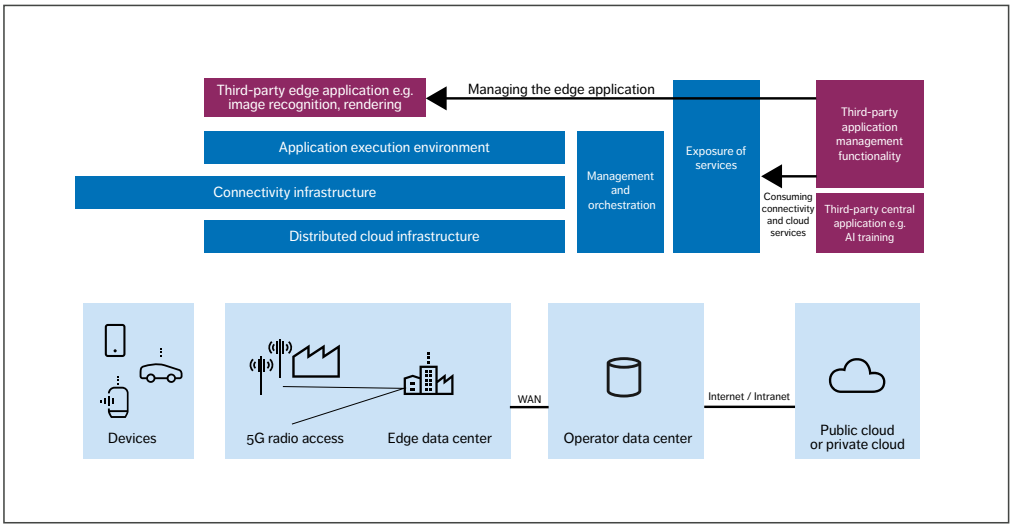


Figure 2 High-level architecture of an edge-computing solution for a typical application

in place, a CSP will be prepared to host any type of third-party application or application platform.

Our high-level solution proposal

Based on our understanding of the key factors shaping the edge-cloud ecosystem, we have defined three main principles that underpin our approach to edge computing:

- Reuse industrialized and proven capabilities whenever possible.
- Ensure backward compatibility.
- Capitalize on existing ecosystems.

The first principle is a reminder that many of the functionalities needed to enable edge computing are not specific to edge computing. They have been used and improved over time, and they should be reused where appropriate. Further, the first principle discourages the adoption of highly specialized solutions early in the process, in light of the current market fragmentation and the uncertainties about the winning use cases in this segment.

The second principle highlights the importance of ensuring that it is possible to deploy existing applications that would benefit from edge deployment without requiring a rewrite on both the device and backend sides.

The third principle pushes us to make the transfer of applications from a central cloud to the edge as transparently as possible to the developers. This means there should be no changes to the life-cycle management of the applications, and existing platforms (along with any specialized ones) should continue to be used for application management and to provide the services the developers need.

With these principles to guide us, we propose a solution with the capabilities to onboard edge applications and edge application platforms into a CSP environment, which can be distributed to the edge data center, central data center or public cloud. *Figure 2* depicts the high-level architecture. The dark-blue boxes represent the main components of our solution and the purple ones indicate third-party applications.

We designed this solution to meet four key criteria:

1. The solution must be able to host different kinds of platforms for different application types.
2. To harmonize with existing developer communities, the execution environment must be CNCF certified (when it is provided by the CSP).
3. To address scaling and mobility issues, the orchestration and management solution of the runtime environment must be aligned with similar functionalities of the network.
4. The solution must both be compatible with 4G and 5G standards and avoid introducing a new layer of complexity (only simple and necessary APIs should be provided).

The solution is based on the distributed cloud infrastructure for virtual network functions (VNFs) and the ETSI NFV orchestration and management functionalities. The same orchestration and management functions are used for the connectivity infrastructure, distributed cloud infrastructure, wide area network (transport) orchestration and the orchestration and management of the application execution environment. This also ensures that there is a user plane function (UPF) available close to the application runtime at the right scaling level that the session management function can select.

To enable transparent connectivity between the edge application and the device, the connectivity infrastructure in our solution is 3GPP compatible. As a result, no edge-solution-specific enhancements are needed in the device.

The exposure functionality provides the main APIs to the third-party developers, of which there are two main types. The first set of APIs is for the business relation with the operator, to enable the onboarding and management of the runtime environment itself and to configure and monitor the connectivity through aggregated APIs built on top of the 3GPP's service capability exposure function, network exposure function and operations support systems/business support systems APIs.

The other set of APIs can be exposed to third-party developers for the deployment and management of the applications themselves. We propose that, for this type of API, a CNCF-certified Kubernetes distribution should be offered in a way similar to how it is provided on web-scale clouds today. This approach harmonizes with the trends and provides developers with greater flexibility.

Runtime environment

To provide a broad baseline for the adoption of applications at the edge, our solution provides customizable Kubernetes distribution in addition to the ability to onboard arbitrary third-party runtime environments.

One of the main benefits of Kubernetes in many different use cases is its modularity. The plugins available in its runtime environment allow a high degree of customization to fit a specific type of workload. We know, however, that industrial applications often rely on dedicated runtime environments that provide tailor-made characteristics, which means that the edge will generally consist of several different runtime environments. As a result, we believe that efficient management of a multitude of different runtime environments is one of the most important capabilities of the edge-computing solution.

Networking and connectivity aspects

Networking requirements in edge deployments are mainly about facilitating connectivity between attached devices and central services (traditional networking), attached devices and edge applications, and edge applications and central services.

The demands on connectivity typically vary between different types of edge applications – both with regard to the type of connectivity as well as the required characteristics. The execution environment, infrastructure, UPFs and management systems must provide the required connectivity services flexibly and efficiently.

Kubernetes provides a variety of container network interfaces to manage connectivity both between microservices and to external endpoints.

Further connectivity features for north-south traffic are enabled by ingress and egress controllers.

The underlying infrastructure is expected to provide basic layer-2 and layer-3 connectivity to support the Kubernetes networking layer. This significantly reduces the management complexity for the underlying infrastructure and bypasses the need to integrate the Kubernetes layer into any lower layer infrastructure management system.

There are several technologies in 4G and 5G to provide local breakout functionality. The packet core VNFs (such as the UPF) can provide local breakout capabilities for the traffic to be routed to the applications in the edge locations. Distributed Anchor Point is a generic solution available today that successfully addresses many use cases and requires no further standardization.

Looking further ahead, Session Breakout is a Domain Name System (DNS)-based solution for dynamic breakout that still needs industry alignment. It is expected to solve issues in many use cases (including enterprise breakout).

Session Breakout can provide optimal traffic-routing according to a Service Level Agreement, for example. 3GPP standardization will be needed to address DNS/IP and exposure use.

Multiple Sessions is a target solution that requires further industry alignment and support in devices (iOS and Android, for example). Based on service-peering principles, it would map applications to specific sessions on the user equipment side, thereby meeting the needs of all use cases along with operators' expectations for network slicing.

Network management, orchestration and assurance aspects

Distributed cloud and edge capabilities require the support of several layers in the network: the transport layer, the virtualization infrastructure layer, the access and core connectivity layer and the edge application layer. *Figure 3* shows how these four layers fit together in the context of consumer devices (on the left) and distributed sites (at the bottom). The edge-application platform

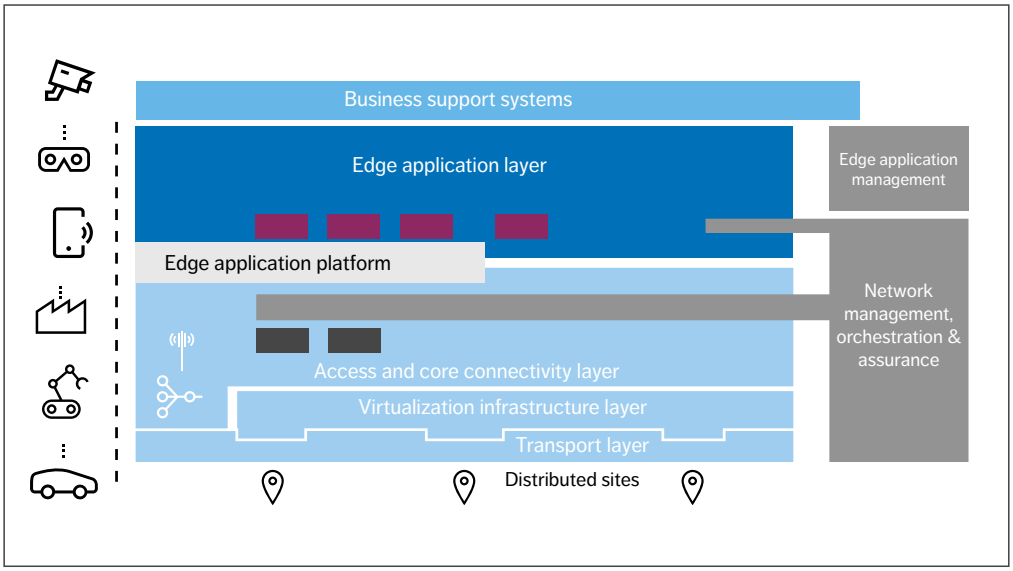


Figure 3 Edge deployment and orchestration

> IN THE FUTURE, ORCHESTRATION AND CONFIGURATION CAPABILITIES MAY ALSO BE ABLE TO PERFORM LOCAL HEALING ACTIONS

(the runtime environment) sits between the third and fourth layers, supported by network management, orchestration and assurance.

Several orchestration and management aspects must be considered, particularly with respect to edge applications (the purple boxes in Figure 3), the edge-application platform, VNFs (shown as dark gray boxes in Figure 3), virtualization infrastructure and the distribution of management functionality.

There are two general approaches to handling edge applications. The first is to treat them like an operator's VNF. Third-party edge applications that will be executed on the edge-application platform require a different approach. These applications will be centrally onboarded, then distributed, and life-cycle managed by the edge platform. When instantiating, the CSP's orchestration and management will create the connectivity to the consumer device over the radio network as well as the connectivity to the internet. The application management and overall assurance can be performed by the edge-application provider or by the CSP.

The edge-application platform(s) can be onboarded and managed by the CSP like a VNF, in the same way that any VNF is onboarded and operated on any virtual infrastructure. The CSP will expose capabilities to instantiate the edge application platform instances when and where required.

VNFs (including cloud-native VNFs) are onboarded, designed and life-cycle managed in a similar way to central data center deployments, with two additional considerations: transport between sites and appropriate distribution of VNFs to the edge to optimize user experience. These capabilities already exist in products based on ETSI Management and Orchestration (MANO) [7] and/or the Open Network Automation Platform (ONAP) [8].

Finally, in a distributed environment, it can be useful to distribute certain management functionalities such as analytics and artificial intelligence functions that can perform local analysis and deliver processed insights, rather than backhauling unnecessary data to a central server. In the future, orchestration and configuration capabilities may also be able to perform local healing actions to support either efficient edge operations or edge-service continuity, even when communication to the edge site has been lost.

Conclusion

Edge computing will play a vital role in enabling a wide range of 5G use cases and helping service providers meet stringent legal requirements on data security and privacy. Beyond its abilities to reduce peering traffic and improve user experience in areas such as video, augmented reality, virtual reality, mixed reality and gaming, edge-computing is also needed to enable ultra-reliable low-latency communication use cases in industrial manufacturing and a variety of other sectors. To meet these diverse needs, communication service providers must be able to deliver edge-computing solutions that can host different kinds of platforms and provide a high level of flexibility for application developers.

It is our view that successful development of an edge-computing solution requires a solid understanding of the use cases, associated deployment options and application-developer communities. It is of critical importance that the solution is able to onboard third-party applications and/or application environments, utilizing methods defined by operations support systems standardization bodies such as TM Forum. Rather than building a new application ecosystem and platform, we strongly recommend reusing industrialized and proven capabilities, utilizing the momentum created with CNCF, and ensuring backward compatibility.

References

1. **Ericsson Technology Review, Cloud-native application design in the telecom domain, June 5, 2019, Saavedra Persson, H; Kassaei, H, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/cloud-native-application-design-in-the-telecom-domain*
2. **Cloud Native Computing Foundation (CNCF), available at:** *https://www.cncf.io*
3. **3GPP, 3GPP SA6 accelerates work on new verticals!, June 7, 2019, Chitturi, S, available at:** *https://www.3gpp.org/news-events/2045-sa6_verticals*
4. **ETSI, Network Functions Virtualisation (NFV), available at:** *https://www.etsi.org/technologies/nfv*
5. **5G Alliance for Connected Industries and Automation (5G ACIA), available at:** *https://www.5g-acia.org/*
6. **Automotive Edge Computing Consortium (AECC), available at:** *https://aecc.org/*
7. **ETSI, Open Source MANO, available at:** *https://www.etsi.org/technologies/nfv/open-source-mano*
8. **Open Network Automation Platform, available at:** *https://www.onap.org/*

Further reading

- **Going beyond edge computing, available at:** *https://www.ericsson.com/en/digital-services/trending/distributed-cloud*
- **Cloud native applications, available at:** *https://www.ericsson.com/en/digital-services/trending/cloud-native*
- **How to orchestrate your journey to Cloud Native, available at:** *https://www.ericsson.com/en/blog/2019/5/how-to-orchestrate-your-journey-to-cloud-native*
- **Is cloud native design really needed in telecom?, available at:** *https://www.ericsson.com/en/blog/2019/1/are-cloud-native-design-really-needed-in-telecom*

THE AUTHORS

Péter Suskovics

◆ joined Ericsson in 2007 as a software developer and participated in several product development groups through contributor and leader roles. The main technology areas were IP, operations and maintenance, NFV, performance management, 5G and the Internet of Things (IoT).
As a strong proponent of open source, Suskovics now works as a system architect in the field of cloud, 5G and the IoT in Business Area Digital Services with a major focus on technology and innovation projects.
He holds an M.Sc. in information engineering (2008) and completed his Ph.D. in network optimization (2011) at the Budapest University of Technology and Economics, Hungary.

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 start-ups 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 works on 5G networks and distributed cloud, as well as coordinating global engineering projects.
Kovács holds an M.Sc. in information engineering and a Ph.D. in mathematics from the Budapest University of Technology and Economics.

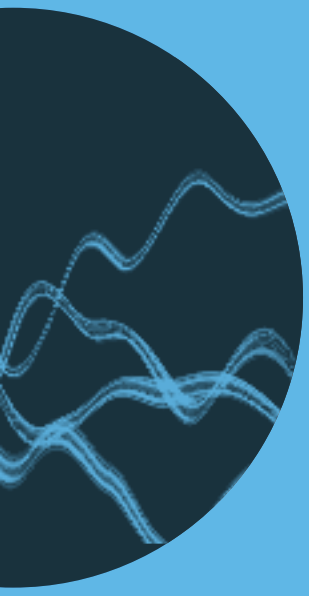
Stephen Terrill

◆ is a senior expert in automation and management, with more than 20 years of experience working with telecommunications architecture, implementation and industry engagement. His work has included both architecture definition and posts within standardization organizations such as ETSI, the 3GPP, ITU-T (ITU Telecommunication Standardization Sector) and IETF (Internet Engineering Task Force).
In recent years, his work has focused on the automation and evolution of operations support systems, and he has been engaged in open source on ONAP's Technical Steering Committee and as ONAP architecture chair.
Terrill holds an M.Sc., a B.E. (Hons.) and a B.Sc. from the University of Melbourne, Australia.

Peter Wörndle

◆ is a technology expert in the area of NFV with responsibility for NFV technology evolution, technology strategy and architecture, as well as cloud-native and edge technologies. Since joining Ericsson in 2007, he has held different positions in R&D and IT, working mainly with cloud and virtualization in R&D, IT operations and

standardization. Wörndle holds an M.Eng. in electrical engineering and communication from RWTH Technical University in Aachen, Germany, and currently serves as the vice-chair of the ETSI NFV Technical Steering Committee.



ISSN 0014-0171
284 23-3335 | Uen

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