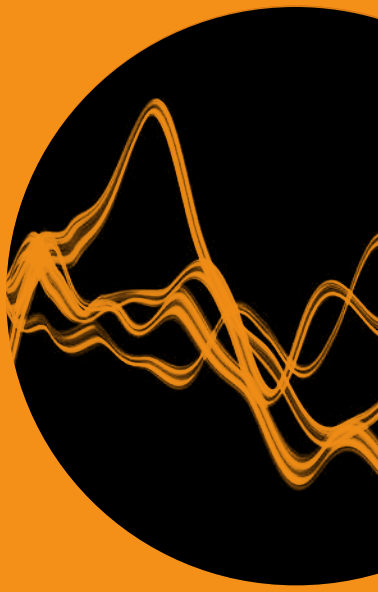
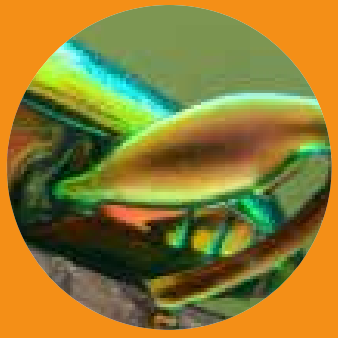
CHARTING THE FUTURE OF INNOVATION | #13·2021

ERICSSON TECHNOLOGY Review



5G SERVICE AUTOMATION: KEY ENABLERS



ERICSSON

Service exposure and automated life-cycle management:

THE KEY ENABLERS FOR 5G SERVICES

Service exposure and automated life-cycle management enable communication service providers to offer a variety of innovative services to enterprises and application developers, while simultaneously establishing new revenue streams through relationships with hyperscale cloud providers.

MALGORZATA SVENSSON, BENEDEK KOVÁCS, ELISABETH MUELLER, MASSIMILIANO MAGGIARI, RÓBERT SZABÓ

The digitalization of society and the growing popularity of 5G-enabled use cases are creating new business opportunities for communication service providers (CSPs) to utilize 3GPP and cloud-based technologies in the enterprise domain.

■ CSPs that want to pursue new business opportunities in the enterprise domain must be able to frame their service offerings to fit the individual needs and desired use cases of enterprises and their partners. Three capabilities are required to achieve this. Firstly, CSPs need to create an expanded service portfolio that combines their connectivity offerings with the cloud and edge platform offerings of hyperscale cloud providers (HCPs). Secondly, they need a network that can serve as a programmable platform with the ability to expose services to developers. Thirdly, they need the ability to onboard new applications into the network with optimized runtime support for traffic routing toward applications.

Being able to offer connectivity services in combination with HCPs' cloud and edge platform offerings will make enterprise applications available to users close to their office locations in an efficient and scalable way, and in compliance with security requirements and local regulations. A growing number of CSPs have already started establishing relationships with HCPs to deploy applications and network functions in HCP environments. Expanding these relationships to offer the HCP's environment in combined offerings for enterprises, partners and application developers is a logical next step. Service-level-agreement (SLA) driven automation is a critical system capability to enable such combined offerings.

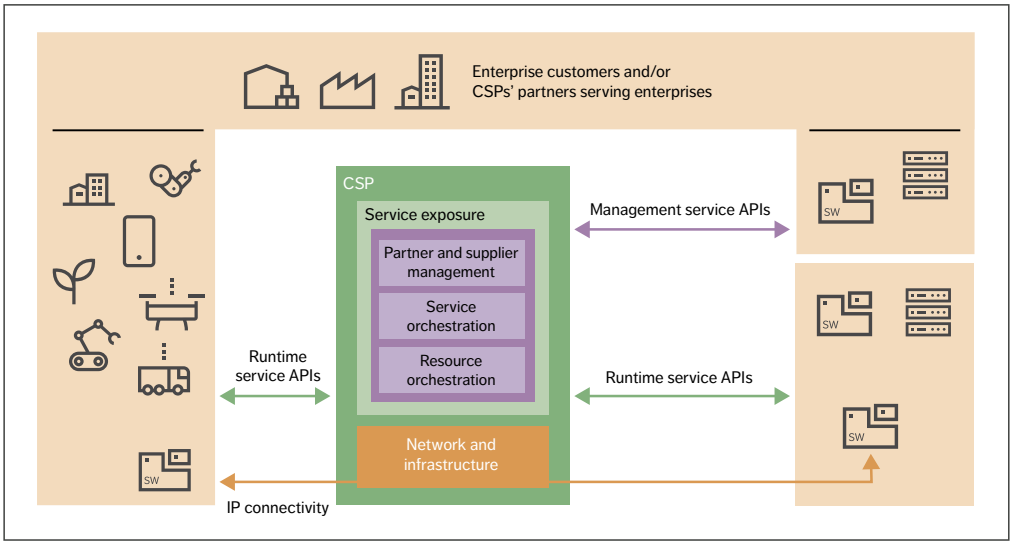


Figure 1 Service exposure and automated life-cycle management

Exposing the network as a programmable connectivity platform enables enterprises and their partners to build applications that can observe the network, influence it, and optimize the traffic flow to enable use cases like unmanned aerial vehicles that demand well-defined response times. As a consequence, developer communities require easy access to exposed service application programming interfaces (APIs).

In the later stages, once applications have been built, they need to be onboarded and integrated into a CSP's environment, so that they can seamlessly be made available on the new services platform, bringing connectivity and applications together. Mechanisms for monetization of combined offerings and service exposure need to be put in place, where the latter is powered by centralized API hubs that monitor API usage and provide input to the various business models.

Service exposure and automated life-cycle management

The term "service exposure and automated life-cycle management" refers to a set of capabilities that targets increased efficiency in CSPs' operational processes through a higher level of automation and scalability in service orchestration architectures. As shown in *Figure 1*, service exposure and automated life-cycle management combine various service management and runtime connectivity services that

Terms and abbreviations

API – Application Programming Interface | **CSP** – Communication Service Provider | **DNS** – Domain Name System | **E2E** – End-to-End | **HCP** – Hyperscale Cloud Provider | **I-WLP** – Intelligent workload placement | **NF** – Network Function | **OTT** – Over-the-Top | **SLA** – Service Level Agreement | **SW** – Software | **UE** – User Equipment | **URSP** – User Equipment Route Selection Policy

are targeted to enterprises, partners and application developers. APIs and orchestration are used to enable the simple and efficient creation and launch of new services and provide value-added functionality in runtime on top of the basic connectivity.

Service exposure provides service and network abstraction layers on top of network APIs, available for usage and integration by enterprise customers and partners to support various business models, where the abstraction layer hides the network complexity.

Partner and supplier management (the top purple box in Figure 1) helps CSPs manage agreements with all types of business partners, including HCPs. These agreements regulate service requirements and the infrastructure capacity required for own or resell usage.

Service orchestration (the middle purple box in Figure 1), together with inventory and assurance, are used for service design, creation and activation, as well as to ensure that the SLA is always fulfilled. These capabilities include interaction with service management on the partner side to onboard services provided by partners and intelligent workload placement that enables data-driven service and network design. The intelligent workload placement fulfills requirements for geographical availability with the distributed target environments and must be integrated with inventory, topology, and service assurance to ensure that decisions are based on the real state of the network. Service orchestration must be tightly integrated to service order management processes to automate the service life-cycle management.

Resource orchestration (the bottom purple box in Figure 1) provides an abstraction layer to various technology domains that result from different business agreements, such as those between CSPs and HCPs. Resource orchestration offers efficient, flexible and automated life-cycle management of CSP software, realizing the network functions (NFs) as well as the consumer and enterprise applications. Initial deployments, service allocation, updates and upgrades of the software can be orchestrated in target environments owned by the CSP or through partner and/or supplier agreements. Multi-cloud capacity management is supported to enable service and resource orchestration.

> “PARTNER AND SUPPLIER MANAGEMENT HELPS CSPs MANAGE AGREEMENTS WITH ALL TYPES OF BUSINESS PARTNERS”

Service exposure (the light green box in Figure 1) enables management service APIs (represented by the purple arrow) to manage connectivity offerings and runtime service APIs (represented by the green arrows) to influence network behavior to meet the desired service characteristics. The management service APIs enable the simple and efficient creation and launch of new services, offering traffic steering for applications at edges. The service APIs are well defined and compliant with the relevant standards, including those of the 3GPP and TM Forum, to ensure interoperability and ease of use.

Partner and supplier management

On top of managing the SLAs between CSPs, HCPs and other digital partners, the partner and supplier management stack also serves as the connection between CSPs’ and partners’ catalogs, making it possible to compose offerings targeted at enterprise customers and application providers. The partner and supplier management stack controls how multi-tenancy is used by both service and resource orchestration, keeping track of which services and resources are used by whom.

CSPs can use these new capabilities to pursue new business opportunities through building relationships with HCPs. HCPs can become the CSPs’ partners, visible to enterprise customers and application providers in combined exposed services offered by the CSP.

Partnerships with HCPs enable CSPs to broaden their end-to-end (E2E) connectivity capabilities for enterprises with application platform offerings and

thereby facilitate other kinds of business relationships as well. Relationships between CSPs and industry verticals is just one example.

CSPs may also choose to extend their relationships with the HCPs, so that the HCPs become their suppliers. In the supplier role, the HCP provides a full-service offering comprised of critical infrastructure along with the cloud and edge stack for consumer applications and telecommunication service deployment.

When CSPs have their own cloud platform offerings, they typically create them by leveraging the services in their partner catalogs that are provided by HCPs. In the case of telecommunication services, applications and NFs are deployed in the cloud, and the cloud resources for service deployment are shielded by services owned and supplied by an HCP. This means that telecommunication services must refer to services defined in the CSP's partner catalog.

Services supplied by HCPs are subject to SLAs between CSPs and HCPs and must be monitored accurately. It is the HCP's responsibility to provide the service assurance for these services in compliance with the service level objectives agreed between the CSP and the HCP. The supplier SLA sets boundaries for any customer-facing SLAs agreed between a CSP and its enterprise customers when delivering combined offerings.

The service-ordering process is driven by connectivity and application platform offering definitions stored in the service catalog. The ordered combined services are decomposed into various domain-level services, and the instantiation of these services is triggered toward service orchestration.

New monetization opportunities arise from flexible business-to-business-to-anything business models, which can consider multiple parties (customers and partners) during charging and billing business processes.

Service and resource orchestration

Automated service life-cycle management aims to simplify the CSP's operational processes and increase efficiency. Service and resource orchestration plays an important role in enabling the simple and efficient creation and launch of new services, as well as providing value-added functionality in runtime on top of the basic connectivity.

“ NEW MONETIZATION OPPORTUNITIES ARISE FROM FLEXIBLE BUSINESS-TO-BUSINESS-TO-ANYTHING BUSINESS MODELS ”

The service orchestration stack includes capabilities to manage services across multiple technologies and business domains in complex operator environments. It also makes it possible to onboard partner services and associated deployment rules in an automated way. These partner services are used as building blocks in the service creation process of services offered by the CSP to enterprise customers. It is also possible to orchestrate various services including network connectivity, enterprise connectivity services and applications, along with orchestration of their deployments in HCP-provided clouds.

The introduction of HCP cloud infrastructure in the telecommunication world is happening gradually and at different speeds, depending on the business objectives of individual CSPs. To address the market and technology situations, full, partner and limited edge [1] have been identified as the go-to-market scenarios. In the full edge scenario, the CSP operates the infrastructure for telco workloads, third-party and over-the-top (OTT) workloads. The infrastructure provider can be any private infrastructure provider.

In the partner edge scenario, the HCP operates the infrastructure for third-party and OTT workloads, but it is deployed in the CSP or enterprise premises. The infrastructure provided for third-party and OTT workloads is used exclusively for workloads controlled by the CSP. The infrastructure for telco workloads can be operated by the CSP or the HCP, deployed in the CSP or in the enterprise

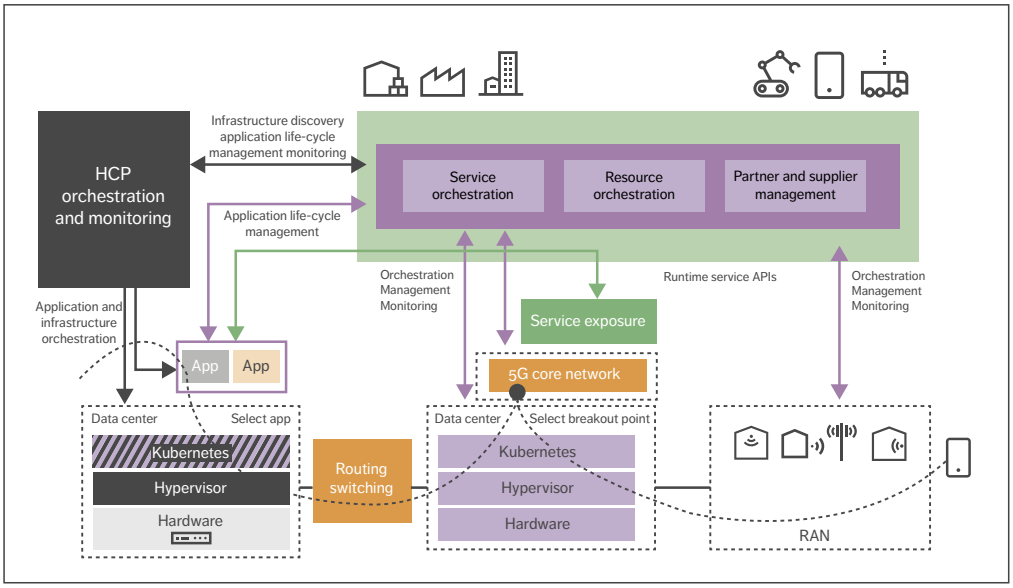


Figure 2 Operations support systems architecture for dual stack

premises and for the exclusive use of the CSP.

In the limited edge scenario, the HCP operates the infrastructure for third-party and OTT workloads. It can be deployed in the CSP or enterprise premises. The HCP will use and control the infrastructure provided for third-party and OTT workloads. The infrastructure for telco workloads can be operated by the CSP, as in the full edge deployment scenario, or operated by the HCP, deployed in the CSP premises or in the enterprise premises and for the exclusive use of the CSP.

These scenarios are not isolated from each other and we can already see the full, partner and limited edge variants happening at the same time, which has led to the dual-stack architecture depicted in *Figure 2*. Telecommunication workloads are deployed in a private telco cloud, while enterprise and consumer applications are deployed in HCP edge zones. In the initial phase, where both the full and partner edge go-to-market scenarios apply, the E2E services, the applications' connectivity awareness, and the operational processes automation across the two cloud stacks are essential to fulfill the service requirements.

At a later stage (or in the short term for more advanced cases) the integrated-stack architecture shown in *Figure 3* will be introduced, in which telecommunication applications are deployed together with enterprise and consumer applications in the cloud infrastructure supplied by the HCP for the exclusive use of the CSP.

In both the dual and the integrated-stack architecture scenarios, the CSP's service and resource orchestration stack performs the life-cycle management of the telecommunication applications and combined services through direct access to the requested Kubernetes clusters.

The main objective of the service and resource orchestration functionality is to minimize the impact of public and/or private cloud deployments on telecommunication services and applications. It is essential to provide the same operational experience to CSPs and enable the commercialization of the services for enterprises irrespective of the variations

of underlying cloud architecture resulting from the various business models and go-to-market tactics. In the full edge scenario, the service and resource orchestration stack handles the life-cycle management of the cloud infrastructure, including Kubernetes clusters. In the partner and the limited edge scenarios, on the other hand, the orchestration stack interworks with the HCP orchestration tools to enable the life-cycle management and monitoring of the cloud infrastructure.

Dual-stack architecture

Service and resource orchestration supports business models in which CSPs own the telco cloud infrastructure and the enterprise cloud infrastructure is provided by HCP suppliers. These business models result in the dual-stack architecture shown in Figure 2, where the CSPs use their own infrastructure (represented by the light purple boxes under the 5G core network in Figure 2) to deploy and life-cycle-manage NFs, and enterprise applications are hosted in the infrastructure provided by the HCP (represented by the purple-black striped box in Figure 2). The two cloud infrastructures are deployed side by side, with E2E automated processes realized by the orchestration stack.

The dual-stack architecture supports the life-cycle management and monitoring of NFs as well as the life-cycle management of enterprise applications that are part of the CSP-managed catalog. It also enables the life-cycle management of CSP-owned cloud infrastructure. The architecture assumes that the integration with the HCP's orchestration and assurance tools and APIs is done through the CSP's service exposure, which is used to delegate application deployment in HCP-based managed cloud infrastructure (such as container as a service and platform as a service). The CSP's service exposure is used to collect metrics to monitor enterprise applications and it is the CSP orchestration stack that provides an E2E service control and configuration point.

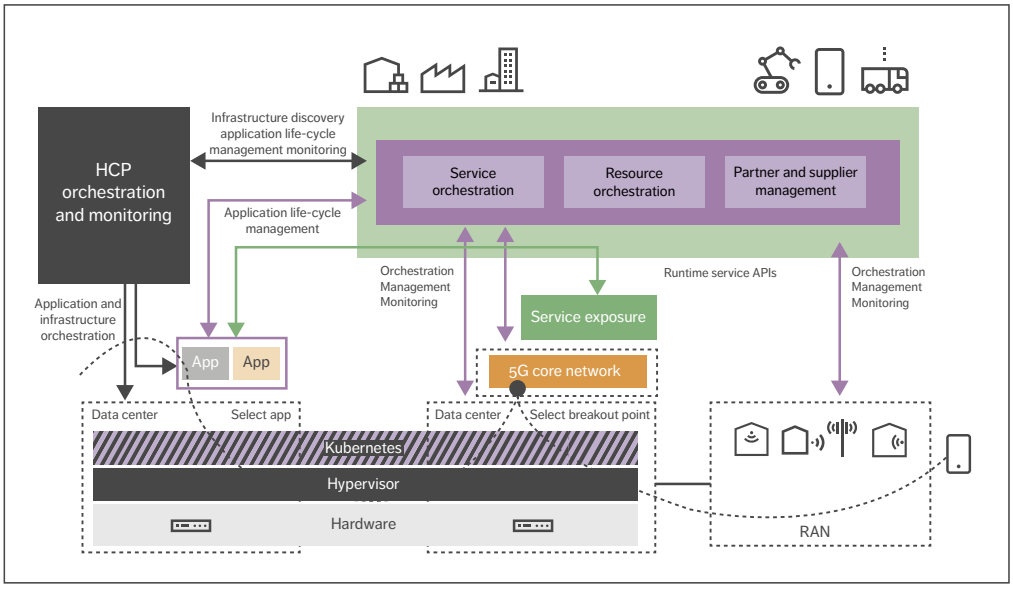


Figure 3 Operations support systems architecture for integrated stack

> THE MULTI-CLOUD ABSTRACTION LAYER HARMONIZES HCP-SPECIFIC BEHAVIORS AND UNIFIES CSP OPERATIONAL EXPERIENCE TO AVOID VENDOR LOCK-IN

The runtime service APIs (represented by the green arrow in Figure 2) can be used by applications to influence network behavior as URSP rules, network slicing policy rules, QoS and traffic steering.

Integrated-stack architecture

In scenarios where the NFs run in the HCP's cloud, service and resource orchestration provides the capabilities to orchestrate and manage the integrated cloud stack. Based on agreements with CSPs, HCP-supplied infrastructure (represented by the purple-black striped box in Figure 3) is used to run NFs along with enterprise and partner applications. The architecture supports management capabilities for the NFs and applications if they are part of the CSP-managed catalog. The management capabilities comprise orchestration, life-cycle management, and monitoring, as well as the integration with the HCP's orchestration and monitoring through service exposure.

Both the dual-stack and integrated-stack architectures utilize the same set of APIs provided by HCPs – that is, infrastructure automation, infrastructure monitoring and Kubernetes APIs. These APIs are consumed by the CSP operations support systems stack. Both architectures also support capabilities such as third-party infrastructure and service discovery, and topology and inventory management, enabling data-driven intelligent service design in mixed cloud environments.

Service and resource orchestration also offers service configuration to allow application traffic breakout points at desired network edges (represented by the purple arrows in Figure 3). Furthermore, it supports the configuration of user equipment route selection policies (URSPs) to support edge application deployments, as well as the enablement of service exposure APIs for application providers (represented by the green arrows in Figure 3).

The multi-cloud abstraction layer harmonizes HCP-specific behaviors and unifies CSP operational experience to avoid vendor lock-in. The layer includes the HCP-agnostic information model and the south-bound adapters for HCP APIs.

Advanced multi-tenancy provides the level of isolation required to manage the HCP's resource allocations. This is essential to enable visibility and troubleshooting of the HCP's availability zones, as well as the onboarded enterprises' and partners' users across the network.

Intelligent workload placement and data-driven network design

Optimal placement of applications and dynamic traffic routing are key differentiators that automated service life-cycle management provides. The key enabler for the optimal placement is topology-aware orchestration, where the knowledge about how to connect 5G connectivity to enterprise and consumer applications resides. Enhanced topology discovery and unified topology models are the most significant capabilities of the orchestration.

An intelligent workload placement (I-WLP) service is characterized by four key features. First and foremost, it must be driven by the CSP's business intents. Secondly, it must respect the onboarding status and availability, SLA objectives, and resource and hardware availability of the components. Thirdly, it must be able to work under continuously changing circumstances, including onboarding changes, software upgrades, new resources and capabilities, changes to the business intents or CSP operational policies, and changes to the costs and availabilities of HCP cloud locations. And lastly, it needs to have knowledge of the

operational state of the distributed system – that is, which service instances are running where with what dependencies. Inventories can provide running, planned and pending reservation states of services, while monitoring systems can provide actual and historical resource usage and SLA metrics.

From the algorithm perspective, I-WLP is a multi-constraint, multi-objective graph allocation problem, which cannot be solved in polynomial time. Therefore, we have created methods based on greedy heuristics, which are very flexible in terms of supporting service constraints such as latency bounds, QoS classes, vendor preferences, locations and anti-affinities. They can also consider combinations of different operational policies relating to cost, utilization, performance and resilience optimizations, for example.

An inventory of all resources, services and capabilities is central to I-WLP. Both the network and the cloud are structured into several layers, each of which represents services, resources and capabilities to the layers above. I-WLP incorporates models for every layer of the CSP's offering. Each layer is life-cycle-managed in the stack, which means that if a higher layer requires reconfiguration of a lower layer the reconfiguration steps are included in a technology-agnostic homing and assignment recommendation. For example, if a service requires additional bandwidth between two datacenters then a VPN connection between the two datacenters can be reconfigured to support those needs.

I-WLP is also invoked by the service orchestrator during service instance design, resulting in a homing and assignment recommendation that is processed by a workflow engine with various southbound adapters.

I-WLP integrates HCP domains into CSP resource and capability pools. It can design and assign network services and applications for co-location, such as dual-stack deployments. Since I-WLP automates the homing and assignment process, the information that would be required for manual homing and assignment is unnecessary. This makes it possible to keep the exposure, APIs and interaction with consumers at a higher level. Intent-based service orchestration, where homing and assignment is derived from service-level requirements and constraints, is an excellent example of this.

Looking ahead, we are investigating how to extend intent-based operation [2] to the full stack and how to provide I-WLP for the network compute fabric [3].

Service exposure

Key components such as partner and supplier management, service and resource orchestration, network and infrastructure leverage exposure capabilities to provide services to customers, partners and application developers. Service exposure provides a uniform way to handle the service APIs (represented by the green and purple arrows in Figures 1, 2 and 3) and user access rights, as well as enabling the delivery of network services in combined offerings to enterprises, partners and application developers.

Additionally, the exposed services are used to influence network connectivity characteristics such as QoS parameters, charging conditions, security settings, network-slicing selection policy rules and traffic routing to selected edge sites to support deployments at the specific location of latency-sensitive applications. In the latter case, we encourage the use of the network slicing signaling to user equipment (UE) based on the URSP paradigm, which introduces rule-based mechanism to separate the Protocol Data Unit (PDU) session traffic toward different edge or non-edge sites.

Industry alignment on the required UE exposure capabilities is important to enable the more

> SERVICE EXPOSURE PROVIDES A UNIFORM WAY TO HANDLE THE SERVICE APIs AND USER ACCESS RIGHTS

sophisticated routing policies needed for more advanced edge and network slicing use cases.

Managing QoS parameters in mobile networks is important from multiple angles. Edge computing applications (such as automotive and gaming) are sensitive to latency and therefore often require specific QoS conditions at given locations. User mobility and uneven radio conditions at different physical locations may also affect application performance. The 3GPP defines different QoS classes and parameters to serve different kinds of traffic [4, 5]. The enforcement of these rules is executed on the user plane and controlled through the 5G system control plane.

Rather than using URSP-based routing, most of the enterprise use cases have the IP anchor point for all traffic on a selected edge. The location of the IP anchor point is based on the principle that the mobile network selects the breakout point for the mobile device. The application layer selects the application server in the edge cloud, especially in the case of applications running in a third-party cloud, as shown in Figure 2 and Figure 3. The application layer must locate the mobile device and request the network to select the correct breakout point – that is, the point where the application traffic leaves the mobile network administrative domain. After leaving the network domain, the application traffic must be directed to the optimal application server. In the case of data centers provided by HCPs, this process is governed by internet rules and generally executed by HCPs' Domain Name System (DNS) service (Amazon Route 53 or Google DNS, for example).

For the partner-edge go-to-market scenario, service and resource orchestration offers services that enable application developers to take advantage of HCPs' standardized development environments and CSP-provided services to build applications that can utilize network insights and drive network optimization in compliance with enterprise and CSP requirements and use cases. It also offers enterprises, partners and integrators access to services in a simplified way without the need to understand complex 3GPP APIs.

Conclusion

Service exposure is essential for communication service providers (CSPs) to enable close collaboration with enterprises and application developers. By combining service exposure with automated life-cycle management, Ericsson has made it possible for CSPs to establish new revenue streams through mutually beneficial relationships with hyperscale cloud providers. Powerful service and resource orchestration functionality is particularly critical to the automation of 5G services in hybrid, public and private clouds. With the help of service exposure and automated life-cycle management capabilities, CSPs can significantly enhance their ability to capitalize on the rapid digitalization of business and society.

Further reading

- **Ericsson Technology Review, The future of cloud computing: Highly distributed with heterogeneous hardware, May 12, 2020, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/the-future-of-cloud-computing*
- **Ericsson Technology Review, Creating the next-generation edge-cloud ecosystem, February 18, 2020, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/next-generation-cloud-edge-ecosystems*
- **Ericsson Technology Review, Service exposure – a critical capability in a 5G world, May 7, 2019, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/service-exposure-a-critical-capability-in-a-5g-world*
- **Ericsson Technology Review, Distributed cloud – a key enabler of automotive and industry 4.0 use cases, November 20, 2018, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/distributed-cloud*
- **Ericsson, Service automation, available at:** *https://www.ericsson.com/en/service-orchestration/service-automation*
- **Ericsson, Service orchestration, available at:** *https://www.ericsson.com/en/service-orchestration*
- **Ericsson, Network automation, available at:** *https://www.ericsson.com/en/network-automation*

References

1. **Ericsson white paper, Edge computing and deployment strategies for communication service providers, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/edge-computing-and-deployment-strategies-for-communication-service-providers*
2. **Ericsson Technology Review, Cognitive processes for adaptive, intent-based networking, November 11, 2020, Niemöller, J; Mokrushin, L; Mohalik, S.K.; Vlachou-Konchylaki, M; Sarmonikas, G, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/adaptive-intent-based-networking*
3. **Ericsson Technology Review, The network compute fabric – advancing digital transformation with ever-present service continuity, June 30, 2021, Sefidcon, A; John, W; Opsenica, M; Skubic, B, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/network-compute-fabric*
4. **3GPP, Specification 29.122, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3239*
5. **3GPP, Specification 23.502, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3145*

The authors would like to thank the following people for their contributions to this article: Carlos Bravo, Ignacio Más, Stephen Terrill, Magnus Buhrgard, Patrick Maguire, Christer Jakobsson and Jan Backman.

THE AUTHORS

Malgorzata Svensson

◆ is an expert in operations support systems (OSS). She joined Ericsson in 1996 and has worked in various areas within research and development. 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.

Benedek Kovács

◆ joined Ericsson in 2005 as a software developer and tester. Today his work focuses 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 (BME) in Hungary.

Elisabeth Mueller

◆ joined Ericsson in 2006 and has held many business support systems (BSS) related roles in areas including system design, system management and solution architecture. She is currently an expert for BSS E2E systems focusing on 5G/Internet of Things BSS architecture. Mueller holds several patents as well as an M.Sc. in mathematics from Johannes Gutenberg University in Mainz, Germany.

Massimiliano Maggiari

◆ is an Ericsson fellow and a senior expert in OSS architecture. Since joining the company in 2006, he has held many roles across product development and product management in the OSS domain. Maggiari holds numerous patents related to OSS and control plane-based networking, as well as an M.Sc. in electronic engineering from the University of Genoa, Italy.

Róbert Szabó

◆ joined Ericsson in 2013 and currently works as a principal researcher at Research Area Cloud Systems and Platform, where he focuses on distributed cloud, zero-touch automation and Network Functions Virtualization. Before joining Ericsson, he worked as an associate professor at BME in Hungary. Szabó holds both a Ph.D. in electrical engineering and an MBA from BME.

ISSN 0014-0171
284 23- 3371 | Uen

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