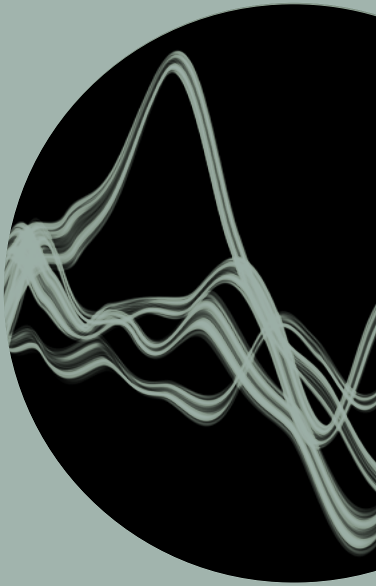
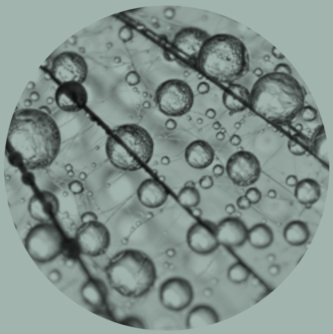
CHARTING THE FUTURE OF INNOVATION | #04·2019

Ericsson Technology Review



SERVICE EXPOSURE IN 5G



ERICSSON

Service exposure:

A CRITICAL CAPABILITY IN A 5G WORLD

Exposure – and service exposure in particular – will be critical to the creation of the programmable networks that businesses need to communicate efficiently with Internet of Things (IoT) devices, handle edge loads and pursue the myriad of new commercial opportunities in the 5G world.

JAN FRIMAN,
MATTIAS EK,
PETER CHEN,
JITENDRA MANOCHA,
JOÃO SOARES

While service exposure has played a notable role in previous generations of mobile technology – by enabling roaming, for example, and facilitating payment and information services over the SMS channel – its role in 5G will be much more prominent.

■ The high expectations on mobile networks continue to rise, with never-ending requests for higher bandwidth, lower latency, increased predictability and control of devices to serve a variety of applications and use cases. At the same time, we can see that industries such as health care and manufacturing have started demanding more customized connectivity to meet the needs of their services. While some of these demands can be met through improved network connectivity capabilities, there are other areas where those improvements alone will not be sufficient.

For example, in recent years, content delivery networks (CDNs) have been used in situations where deployments within the operator network became a necessity to address requirements like high bandwidth. More recently, however, new use-case categories in areas such as augmented reality (AR)/virtual reality (VR), automotive and Industry 4.0 have made it clear that computing resources need to be accessible at the edge of the network. This development represents a great opportunity for operators, enterprises and application developers to introduce and capitalize on new services. The opportunity also extends to web-scale providers (Amazon, Google, Microsoft, Alibaba and so on) that have invested in large-scale and distributed cloud infrastructure deployments on a global scale, thereby becoming the mass-market provider of cloud services.

Several web-scale providers have already started providing on-premises solutions (a combination of full-stack solutions and software-only solutions) to meet the requirements of certain use cases. However, the ability to expand the availability of web-scale services toward the edge of the operator infrastructure would make it possible to tackle a multitude of other use cases as well. Such a scenario is mutually beneficial because it allows the web-scale providers to extend the reach of services that benefit from being at the edge of the network (such as the IoT and CDNs), while enabling telecom operators to become part of the value chain of the cloud computing market.

> “SUCH A SCENARIO... [ENABLES] TELECOM OPERATORS TO BECOME PART OF THE VALUE CHAIN OF THE CLOUD COMPUTING MARKET”

Defining exposure

Exposure in the IT/telecom sphere can be divided into a number of subareas.

Data exposure is the process by which any kind of consumer (human or machine) can access data in a system via secure and controlled mechanisms. Data is normally exchanged in one direction only. Common examples of data exposure include accessing data via an application programming interface (API), downloading a file or retrieving observations from a server.

Service exposure goes beyond data exposure to also include the ordering of execution of operations in the underlying system. Using an API to initiate operations and/or processes is a good example of service exposure. Services can be invoked bidirectionally by triggering events, for example. Data can also be updated via a service.

Service exposure can be applied in a domain, as in **network exposure**, which exposes both data and services of the network. Enterprise resource planning (ERP) and customer relationship management (CRM) are other examples of domains where service exposure can be applied.

To maintain security, the details of the underlying system are typically hidden in exposure scenarios.

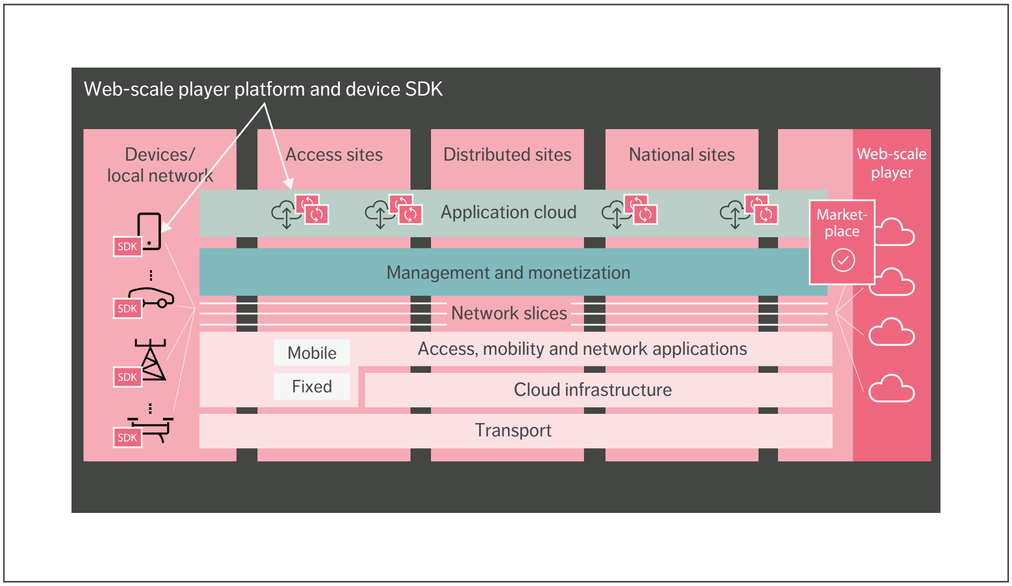


Figure 1 Collaboration with web-scale providers on telecom distributed clouds

Figure 1 illustrates how a collaboration with web-scale providers on telecom distributed clouds could be structured. We are currently exploring a partnership to enable system integrators and developers to deploy web-scale player application platforms seamlessly on telecom distributed clouds. Distributed cloud abstraction on the web-scale player marketplace encompasses edge compute, latency and bandwidth guarantee and mobility. Interworking with IoT software development kits (SDKs) and device management provides integration with provisioning certificate handling services and assignment to distributed cloud tenant breakout points.

In the mid to long term, service exposure will be critical to the success of solutions that rely on edge computing, network slicing and distributed cloud. Without it, the growing number of functions, nodes, configurations and individual offerings that those solutions entail represents a significant risk of increased operational expenditure. The key benefit of service exposure in this respect is that it makes it possible to use application programming interfaces (APIs) to connect automation flows and artificial intelligence (AI) processes across organizational, technology, business-to-business (B2B) and other borders, thereby avoiding costly manual handling. AI and analytics-based services are particularly good candidates for exposure and external monetization.

Key enablers

The 5G system architecture specified by 3GPP has been designed to support a wide range of use cases based on key requirements such as high bandwidth/throughput, massive numbers of connected devices and ultra-low latency. For example, enhanced mobile broadband (eMBB) will provide peak data rates above 10Gbps, while massive machine-type communications (mMTC) can support more than 1 million connections per square kilometer. Ultra-reliable low-latency communications (uRLLC) guarantees less than 1ms latency.

Fulfilling these eMBB, mMTC and uRLLC requirements necessitates significant changes to both the RAN and the core network. One of the most significant changes is that the core network functions (NFs) in the 5G Core (5GC) interact with each other using a Service-based Architecture (SBA). It is this change that enables the network programmability, thereby opening up new opportunities for growth and innovation beyond simply accelerating connectivity.

Service-based Architecture

The SBA of the 5GC network makes it possible for 5GC control plane NFs to expose Service-based Interfaces (SBIs) and act as service consumers or producers. The NFs register their services in the network repository function, and services can then be discovered by other NFs. This enables a flexible deployment, where every NF allows the other authorized NFs to access the services, which provides tremendous flexibility to consume and expose services and capabilities provided by 5GC for internal or external third parties. This support of the services subscription makes it completely different to the 4G/5G Evolved Packet Core network.

Because it is service-driven, SBA enables new service types and supports a wide variety of diversified service types associated with different technical requirements. 5G provides the SBI for different NFs (for example via SBI HTTP/2 Restful APIs). The SBI can be used to address the diverse service types and highly demanding performance requirements in an efficient way. It is an enabler for short time to market and cloud-native web-scale technologies.

The 3GPP is now working on conceptualizing 5G use cases toward industry verticals. Many use cases can be created on-demand as a result of the SBA.

Distributed cloud infrastructure

The ability to deploy network slices – an important aspect of 5G – in an automated and on-demand manner requires a distributed cloud infrastructure. Further, the ability to run workloads at the edge of the network requires the distributed cloud infrastructure to be available at the edge. What this essentially means is that distributed cloud deployments within the operator network will be an inherent part of the introduction of 5G. The scale, growth rate, distribution and network depth (how far out in the network edge) of those deployments will vary depending on the telco network in question and the first use cases to be introduced.

As cloud becomes a natural asset of the operator infrastructure with which to host NFs and services (such as network slicing), the ability to allow third parties to access computing resources in this same infrastructure is an obvious next step. Contrary to the traditional cloud deployments of the web-scale players, however, computing resources within the operator network will be scarcer and much more geographically distributed. As a result, resources will need to be used much more efficiently, and mechanisms will be needed to hide the complexity of the geographical distribution of resources.

> “CORE NETWORK FUNCTIONS IN THE 5GC INTERACT WITH EACH OTHER USING A SERVICE-BASED ARCHITECTURE”

Cloud-native principles

The adoption of cloud-native implementation principles is necessary to achieve the automation, optimized resource utilization and fast, low-cost introduction of new services that are the key features of a dynamic and constrained ecosystem. Cloud-native implementation principles dictate that software must be broken down into smaller, more manageable pieces as loosely coupled stateless services and stateful backing services. This is usually achieved by using a microservice architecture, where each piece can be individually deployed, scaled and upgraded. In addition, microservices communicate through well-defined and version-controlled network-based interfaces, which simplifies integration with exposure.

Three types of service exposure

There are three main types of service exposure in a telecom environment:

- network monitoring
- network control and configuration
- payload interfaces.

Examples of network monitoring service exposure include network publishing information as real-time statuses, event streams, reports, statistics, analytic insights and so on. This also includes read requests to the network.

Service exposure for network control and configuration involves requesting control services that directly interact with the network traffic or request configuration changes. Configuration can also include the upload of complete virtual network functions (VNFs) and applications.

Examples of service-exposure-enabled payload interfaces include messaging and local breakout, but it should be noted that many connectivity/payload interfaces bypass service exposure for legacy reasons. Even though IP connectivity to devices is a service that is exposed to the consumer, for example, it is currently not achieved via service exposure. The main benefit of adding service exposure would be to make it possible to interact with the data streams through local breakout for optimization functions.

Leveraging software development kits

At Ericsson, we are positioning service exposure capabilities in relation to developer workflows and practices. Developers are the ones who use APIs to create solutions, and we know they rely heavily on SDKs. There are currently advanced developer frameworks for all sorts of advanced applications including drones, AR/VR, the IoT, robotics and gaming. Beyond the intrinsic value in exposing native APIs, an SDK approach also creates additional value in terms of enabling the use of software libraries, integrated development environments (IDEs) plug-ins, third-party provider (3PP) cloud platform extensions and 3PP runtimes on edge sites, as well as cloud marketplaces to expose these capabilities.

Software libraries can be created by prepackaging higher-level services such as low-latency video streaming and reverse charging. This can be achieved, for example, by using the capabilities of network exposure functions (NEF) and service capability exposure functions (SCEF), creating ready-to-deploy functions or containers that can be distributed through open repositories, or even marketplaces, in some cases. This possibility is highly relevant for edge computing frameworks.

Support for IDE plug-ins eases the introduction of 3PP services with just a few additional clicks. Selected capabilities within 3PP cloud platform extensions can also create value by extending IoT device life-cycle management (LCM) for cellular connected devices, for example. The automated provisioning of popular 3PP edge runtimes on telco infrastructure enables 3PP runtimes on edge sites.

> “CLOUD MARKETPLACES ARE AN IDEAL PLACE TO EXPOSE ALL OF THESE CAPABILITIES”

Finally, cloud marketplaces are an ideal place to expose all of these capabilities. The developer subscribes to certain services through their existing account, gaining the ability to activate a variety of libraries, functions and containers, along with access to plug-ins they can work with and/or the automated provisioning required for execution.

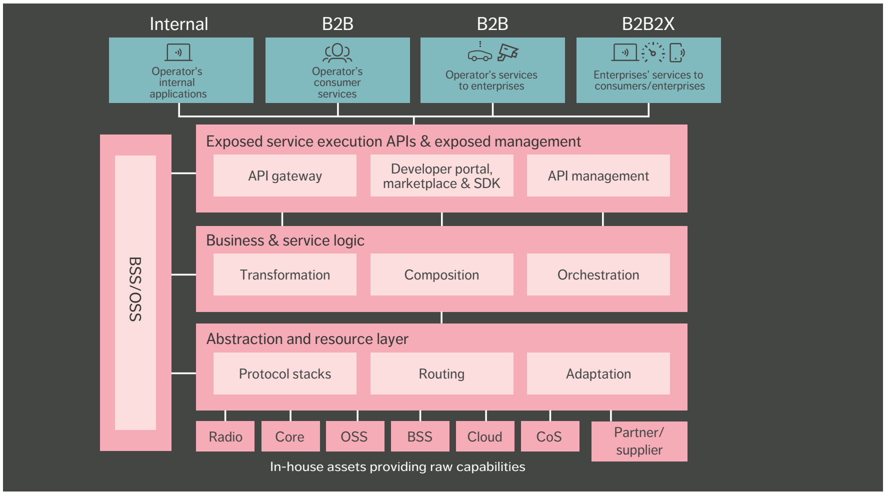

Internal
B2B
B2B
B2B2X
Operator's internal applications
Operator's consumer services
Operator's services to enterprises
Enterprises' services to consumers/enterprises
Exposed service execution APIs & exposed management
API gateway
Developer portal, marketplace & SDK
API management
Business & service logic
Transformation
Composition
Orchestration
Abstraction and resource layer
Protocol stacks
Routing
Adaptation
BSS/OSS
Radio
Core
OSS
BSS
Cloud
CoS
Partner/ supplier
In-house assets providing raw capabilities

Figure 2 Functional architecture for service exposure

Functional architecture for service exposure

The functional architecture for service exposure is built around four customer scenarios:

- internal consumers
- business-to-consumers (B2C)
- business-to-business (B2B)
- business-to-business-to-business/consumers (B2B2X).

In the case of internal consumers, applications for monitoring, optimization and internal information sharing operate under the control and ownership of the enterprise itself. In the case of B2C, consumers directly use services via web or app support. B2C examples include call control and self-service management of preferences and subscriptions. The B2B scenario consists of partners that use services such as messaging and IoT communication to support their business. The B2B2X scenario is made up of more complex value chains such as mobile virtual network operators, web scale, gaming, automotive and telco cloud through web-scale APIs.

Figure 2 illustrates the functional architecture for service exposure. It is divided into three layers that each act as a framework for the realization. Domain-specific functionality and knowledge are applied and added to the framework as configurations, scripts, plug-ins, models and so on. For example, the access control framework delivers the building blocks for specializing the access controls for a specific area.

The abstraction and resource layer is responsible for communicating with the assets. If some assets are located outside the enterprise – at a supplier or partner facility in a federation scenario, for example – B2B functionality will also be included in this layer.

The business and service logic layer is responsible for transformation and composition – that is, when

> “ COMMON EXPOSURE FUNCTIONS [CAN BE DEPLOYED] BOTH IN A DISTRIBUTED WAY AND INDIVIDUALLY ”

there is a need to raise the abstraction level of a service to create combined services.

The exposed service execution APIs and exposed management layer are responsible for making the service discoverable and reachable for the consumer. This is done through the API gateway, with the support of portal, SDK and API management.

Business support systems (BSS) and operations support systems (OSS) play a double role in this architecture. Firstly, they serve as resources that can expose their values – OSS can provide analytics insights, for example, and BSS can provide “charging on behalf of” functionality. At the same time, OSS are responsible for managing service exposure in all assurance, configuration, accounting, performance, security and LCM aspects, such as the discovery, ordering and charging of a service.

One of the key characteristics of the architecture presented in Figure 2 is that the service exposure framework life cycle is decoupled from the exposed services, which makes it possible to support both short- and long-tail exposed services. This is realized through the inclusion and exposure of new services through configuration, plug-ins and the possibility to extend the framework.

Another key characteristic to note is that it is possible to deploy common exposure functions both in a distributed way and individually – in combination with other microservices for efficiency reasons, for example. Typical cases are distributed cloud with edge computing and web-scale scenarios such as download/upload/streaming where the edge site and terminal are involved in the optimization.

The exposure framework is realized as a set of loosely connected components, all of which are cloud-native compliant and microservice based, running in containers. There is not a one-size-fits-all deployment – some of the components are available in several variants to fit different scenarios. For example, components in the API gateway support B2B scenarios with full charging but there are also scaled-down versions that only support reporting, intended for deployment in internal exposure scenarios.

Other key properties of the service exposure framework are:

Terms and abbreviations

3PP – Third-party Provider | **5GC** – 5G Core | **AI** – Artificial Intelligence | **API** – Application Programming Interface | **AR** – Augmented Reality | **B2B** – Business-to-Business | **B2BCX** – Business-to-Business-to-Business/Consumers | **B2C** – Business-to-Consumers | **BSS** – Business Support Systems | **CDN** – Content Delivery Network | **CoS** – Communication Services | **CRM** – Customer Relationship Management | **eMBB** – Enhanced Mobile Broadband | **ERP** – Enterprise Resource Planning | **IDE** – Integrated Development Environment | **IoT** – Internet of Things | **LCM** – Life-cycle Management | **mMTC** – Massive Machine-type Communications | **NEF** – Network Exposure Functions | **NF** – Network Function | **ONAP** – Open Network Automation Platform | **OSS** – Operations Support Systems | **SBA** – Service-based Architecture | **SBI** – Service-based Interface | **SCEF** – Service Capability Exposure Functions | **SDK** – Software Development Kit | **uRLLC** – Ultra-reliable Low-latency Communications | **VNF** – Virtual Network Function | **VR** – Virtual Reality

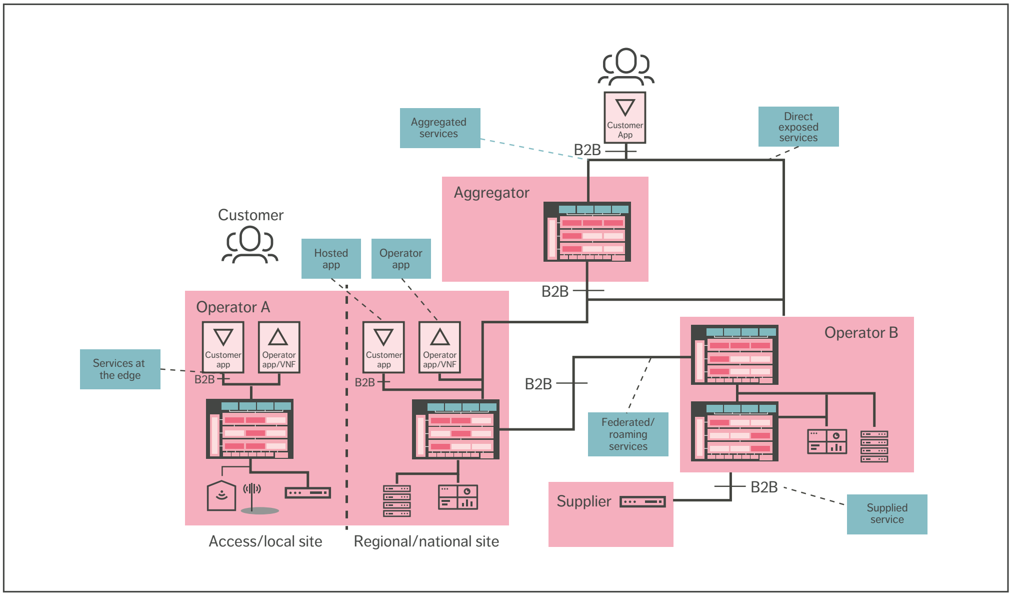


Figure 3 Service exposure deployment (dark pink boxes indicate deployed components)

- scalability (configurable latency and scalable throughput) to support different deployments
- diversified API types for payload/connectivity, including messaging APIs (request-response and/or subscribe-notify type), synchronous, asynchronous, streaming, batch, upload/download and so on
- multiple interface bindings such as restful, streaming and legacy
- multivendor and partner support (supplier/federation/aggregator/web-scale value chains)
- security and access control functionality.

Deployment examples

Service exposure can be deployed in a multitude of locations, each with a different set of requirements that drive modularity and configurability needs. *Figure 3* illustrates a few examples.

In the case of Operator B in Figure 3, service exposure is deployed to expose services in a full B2B context. BSS integration and support is required to handle all commercial aspects of the exposure and LCM of customers, contracts, orders, services and so on, along with charging and billing. Operator B also uses the deployed B2B commercial support to acquire services from a supplier.

In the case of Operator A, service exposure is deployed both at the central site and at the edge site to meet latency or payload requirements. Services are only exposed to Operator A's own applications/VNFs, which limits the need for B2B support. However, due to the fact that Operator A hosts some applications for an external partner, both centrally and at the edge, full B2B support must be deployed for the externally owned apps.

The aggregator in Figure 3 deploys the service exposure required to create services put together by more than one supplier. Unified Delivery Network and web-scale integration both fall into this category. As exposure to the consumer is done through the aggregator, this also serves as a B2B interface to handle specific requirements. Examples of this include the advertising and discovery of services via the portals of web-scale providers.

A subset of B2B support is also deployed to provide the service exposure that handles the federation relationship between Operator A and Operator B, in which both parties are on the same level in the ecosystem value chain.

Conclusion

There are several compelling reasons for telecom operators to extend and modernize their service exposure solutions as part of the rollout of 5G. One of the key ones is the desire to meet the rapidly developing requirements of use cases in areas such as the Internet of Things, AR/VR, Industry 4.0 and the automotive sector, which will depend on operators' ability to provide computing resources across the whole telco domain, all the way to the edge of the mobile network. Service exposure is a key component of the solution to enable these use cases.

Recent advances in the service exposure area have resulted from the architectural changes introduced in the move toward 5G and the adoption of cloud-native principles, as well as the combination of Service-based Architecture, microservices and container technologies. As operators begin to use 5G technology to automate their networks and support systems, service exposure provides them with the additional benefit of being able to use automation in combination with AI to attract partners that are exploring new, 5G-enabled business models. Web-scale providers are also showing interest in understanding how they can offer their customers an easy extension toward the network edge.

Modernized service exposure solutions are designed to enable the communication and control of devices, providing access to processes, data, networks and OSS/BSS assets in a secure, predictable and reliable manner. They can do this both internally within an operator organization and externally to a third party, according to the terms of a Service Level Agreement and/or a model for financial settlement.

Service exposure is an exciting and rapidly evolving area and Ericsson is playing an active role in its ongoing development. As a complement to our standardization efforts within the 3GPP and Industry 4.0 forums, we are also engaged in open-source communities such as ONAP (the Open Network Automation Platform). This work is important because we know that modernized service exposure solutions will be at heart of efficient, innovative and successful operator networks.

Further reading

- **Ericsson web page, Service enablement, available at:** *https://www.ericsson.com/en/portfolio/digital-services/cloud-core/service--enablement*
- **Ericsson web page, Cloud core exposure server, available at:** *https://www.ericsson.com/en/portfolio/digital-services/cloud-core/cloud-unified-data-management-and-policy/cloud-core-exposure-server*
- **Ericsson web page, Cloud packet core, available at:** *https://www.ericsson.com/en/portfolio/digital-services/cloud-core/cloud-packet-core*

THE AUTHORS

Jan Friman

◆ is an OSS/BSS expert in the Architecture and Technology team within Business Area Digital Services, where he is driving the architecture of service exposure. Since joining Ericsson in 1997, he has held various OSS/BSS-related positions within the company's R&D, system management and strategic product management organizations. He holds an M.Sc. in computer science from Linköping University, Sweden.

Mattias Ek

◆ joined Ericsson in 1996 and currently serves as a strategic product manager. He has extensive experience in service delivery platforms and service enablement domains, specializing in consumer interaction, mobile commerce and consumer self-service. His focus in recent years has shifted toward exposure and enablement solutions for cellular IoT, massive IoT and machine-type communications. Today, Ek leads the IoT Enabler and Network Exposure team in Solution Area Packet Core with responsibility for commercial and product strategies.

Peter Chen

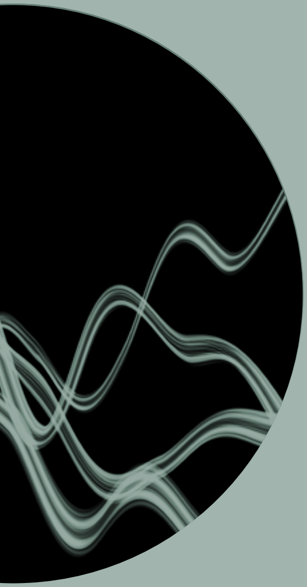
◆ is the technical product manager leading the technical solution and evolution for the network exposure area in Product Development Unit UDM & Policy. He has been working in different areas within the core network at Ericsson since 2006 including IMS, voice over Wi-Fi and Unified Data Management (UDM), and he has contributed more than 10 patents in these areas in recent years. He holds a B.Sc. in materials science and engineering from Dalian University of Technology, China.

Jitendra Manocha

◆ is strategic product manager (5G Core) in Solution Area Packet Core within Business Area Digital Services, where he is responsible for the Cloud Core Exposure Server, a component of Ericsson's 5G Cloud Core solution. He joined Ericsson in 2004 and has held various leading positions in product lines, R&D and services. He holds an M.Sc. from KTH Royal Institute of Technology in Stockholm, Sweden.

João Soares

◆ is a solution manager for distributed cloud, leading Ericsson's strategic solution development for edge computing. Before joining the company in 2014, he worked for Portugal Telecom (now Altice Portugal), during the introduction of cloud technologies within the operator's network. He holds both an M.Sc. and a Ph.D. in electronics and telecommunications engineering from the University of Aveiro, Portugal.



ISSN 0014-0171
284 23-3328 | Uen

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