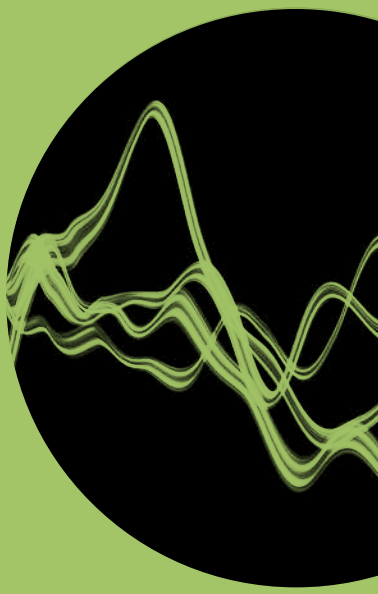
CHARTING THE FUTURE OF INNOVATION | #07·2021

ERICSSON TECHNOLOGY Review



THE NETWORK COMPUTE FABRIC

ERICSSON

The network compute fabric

– ADVANCING DIGITAL TRANSFORMATION WITH EVER-PRESENT SERVICE CONTINUITY

As successive mobile network generations have made connectivity faster and more reliable, cloudification and related virtualization of the underlying infrastructure have revolutionized the IT domain, making affordable cloud services available for widespread use. In 6G, these two major trends come together to create the network compute fabric, the innovation platform of the future.

AZIMEH SEFIDCON, WOLFGANG JOHN, MILJENKO OPSENICA, BJÖRN SKUBIC

In order to fully support emerging use cases such as the Internet of Senses [1], cyber-physical systems and connected intelligent machines [2], the foundation of future network platforms must be built on tight integration between reliable, deterministic connectivity and scalable, affordable and efficient processing capabilities. The result of this tight integration is what we call the network compute fabric.

■ The network compute fabric is one of the four key components of our vision of 6G as an innovation platform [3]. The other three are limitless connectivity, cognitive networks, and trustworthy systems, as shown in *Figure 1*.

The network compute fabric is a fusion of connectivity and computing capabilities, acting as one unified entity, to satisfy both compute and connectivity requirements. Currently, for instance, service anchoring, routing and channel termination are governed by the mobile radio network, while placement and allocation of application tasks are governed by evolving operations support systems and local operating systems. By tightly integrating networking and compute elements, functions like radio-channel handling would operate in concordance with the allocation of related application tasks, opening up for fine-grained optimization possibilities.

Components of the network compute fabric

The left side of *Figure 2* illustrates the four key components of a global, unified network compute

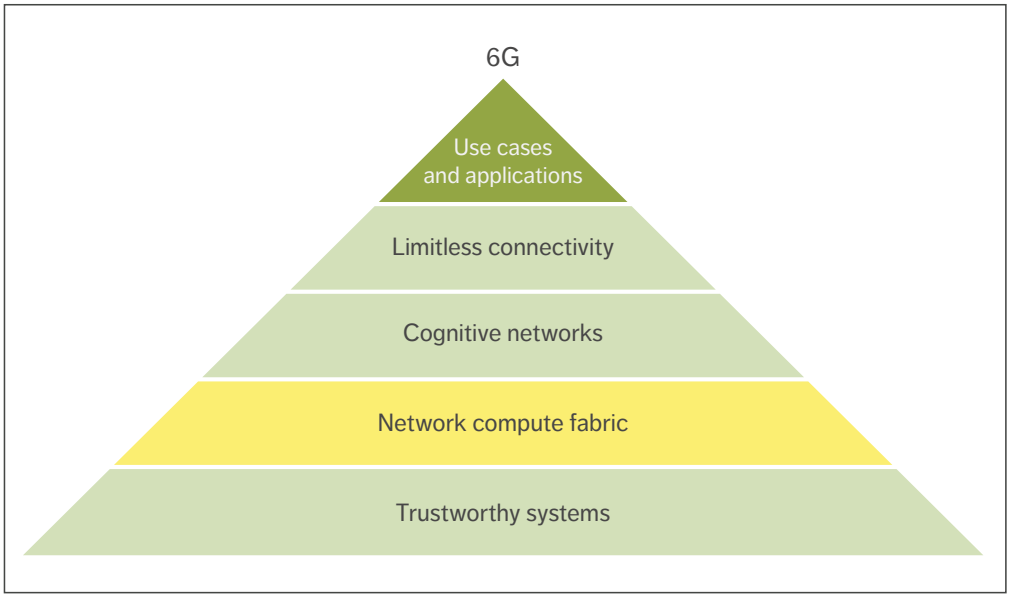


Figure 1 6G as an innovation platform for emerging use cases and applications

fabric: unified ecosystems, unified application management, unified execution environment and unified exposure of network and compute capabilities.

Unified ecosystems

To enable a wider range of more demanding use cases, the network compute fabric needs to facilitate the transformation of ecosystem engagements. This will be realized through the collaboration of a broad set of actors, driven by the need to mutualize the cost of infrastructure, offer services at larger scale and enable full roaming between providers. Network and cloud providers, application developers, service providers, and device and equipment vendors all have a role to play. Network operators have an opportunity to utilize their distributed network infrastructure as an innovation platform for richer service offerings, combining connectivity with compute in both wide-area and on-premises edge-deployment scenarios [4].

Several ecosystem models may evolve and have already been described in the context of edge computing [5]. There are four basic models for network compute fabric services: standalone, facilitation, aggregation and federation. The standalone model is where the network compute fabric services are offered by one actor in a standalone manner. The facilitation model is where one provider acts to unify fragmented offerings from other providers. The aggregation model is where one provider is active from a business perspective in building and reselling bundles of services from

Terms and abbreviations

CSP – Communication Service Provider | **E2E** – End-to-End

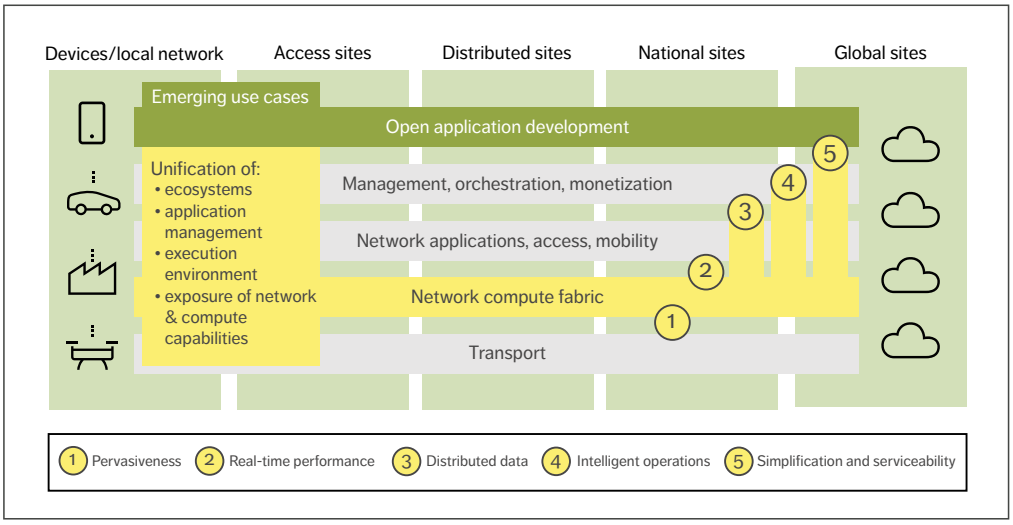


Figure 2 The four key components of the network compute fabric (on the left) enable a new set of features for both network and use-case applications (1-5 on the right)

multiple providers. The federation model is when the providers collaborate within a framework to utilize each other's offerings.

Ecosystem partnerships involve dynamic business agreements that need to be reached regarding the usage of resources between multiple partners. They also involve technical challenges of integrating services from the different actors. Ecosystem partnering can be facilitated by standards that ensure interoperability or by technologies that automate the handling of partner relationships. The right level of harmonization in the ecosystem is key to support scalability as well as innovation.

Unified application management

The network compute fabric will be a highly distributed platform that enables the execution of applications across multiple administrative domains. This innovation platform will require intelligent and data-driven operations to seamlessly span devices and network and cloud domains. An integrated DevOps toolchain will enable the appliance of similar development and operational methodologies across those domains. By separating orchestration from application functionality, the network compute fabric will allow tailored optimization for different network domains without the need for application code change.

Applications are usually composed of one or more services that may have different deployment and performance requirements. With interoperable abstractions and programming models, services can be easily deployed across different domains. With common operation models and tools, service quality can be assured autonomously per domain, considering joint network and compute resources and domain-specific performance requirements. Applications can also be collocated with other applications with dependent performance requirements. Common optimization loops will then enable operational synergies across such applications – that is, data-driven shared knowledge can proactively optimize dependent applications beyond individual optimization models.

Generalized operational pipelines and tools with programmable infrastructure will also enable the

offloading of the common orchestration functionality from applications to the fabric. This will enable simpler application programming models with a stronger focus on the core application functionality.

With common data pipelines and data governance models including data protection, data flows can be combined across resource, functional and administrative domains for multi-layer data classifications that can bring aggregated and more accurate optimizations beyond single domain maturity.

Unified execution environment

The unified execution environment will act as an operating system, providing fundamental functionalities and services on top of distributed and heterogeneous network, compute and storage assets [6]. Evolving the ideas of serverless computing, it will facilitate the development and deployment of distributed applications on top of this infrastructure. The application has access to a compute service that always appears local, despite dynamic network changes or user/data mobility events. The unified execution environment will simplify the development of distributed applications by offering several capabilities.

Seamless task mobility will be enabled by evolved container formats based on portable bytecode such as WebAssembly and associated system interfaces. This will allow the platform to dynamically schedule workloads on nodes regardless of varying hardware and system software setups. As a result, several optimizations can be performed with limited overhead, such as moving computations close to a data source or consumer. This can be useful for a variety of reasons, including having computational tasks follow mobile users, offloading workloads from the device to preserve device energy and moving computations for an optimized cost/performance trade-off.

A distributed data infrastructure with a range of built-in consistency mechanisms will allow developers to choose from the right-sized trade-off between consistency levels and their associated resource costs. This will offer application tasks seamless access to data stored in multiple locations, as if from a single local access. As the application components will be distributed between several locations in a pipeline fashion, the platform would be in full control of both the compute and data locations, as well as traffic management and prioritization. As a result, data access can be optimized along the computation pipeline even considering network characteristics and usage profiles, while obliging legal or operational data policies.

Unified exposure of network and compute capabilities

The computational environment in the network compute fabric will be heterogeneous, which will increase with emerging computational hardware complementing traditional network equipment. One example is hardware acceleration technologies that provide compute and storage capabilities specialized for certain workload tasks, such as graphics processing units, field-programmable gate arrays and storage-class persistent memories.

By adapting tools and languages like OneAPI [7], developers will be able to write a single-source implementation of an algorithm that will be able to run on different hardware choices. Selection of actual processing or memory/storage instance will be based on availability and performance requirements, making it easy for developers to create performant applications that can run anywhere.

The network compute fabric also opens up for extensive in-network computation. Modern transport networking equipment will no longer be limited to packet transport, but also able to provide programmable compute capabilities, opening up for new divisions of application functionality where some tasks, such as QoS, scheduling, policing,

> THE RIGHT LEVEL OF HARMONIZATION IN THE ECOSYSTEM IS KEY TO SUPPORT SCALABILITY AS WELL AS INNOVATION

encoding and recoding can be performed directly in the data plane.

The network compute fabric will enable efficiency gains through transparent shortcuts inside the infrastructure between application components and the network and compute platform. Once applications are collocated with network functions on the same host, rack or cluster, parts of the network and operating system stacks can be bypassed. Depending on the situation, different technologies can be used, such as modern interconnect technologies and protocols in combination with fast replication services like Derecho [8].

Features of the network compute fabric

The unification of ecosystems, application management, execution environments and exposure of network and compute capabilities will lead to the creation of a network compute fabric that can deliver the five advanced features shown on the right side of Figure 2: pervasiveness, real-time performance, distributed data, intelligent operations, and simplification and serviceability.

Pervasiveness

The network compute fabric will provide ubiquitous compute resources, fully integrated with the network to complement connectivity, spanning from central parts all the way out to devices. This will enable distributed applications that interact with physical reality to benefit from having their tasks close to data sources and data consumers. Examples of where this will be useful include radio beamforming, closed-loop control of mission-critical processes and the intelligent aggregation of large amounts of data.

> THE NETWORK COMPUTE FABRIC WILL ENSURE OPTIMIZED APPLICATION EXECUTION WITH REAL END-TO-END PERFORMANCE GUARANTEES

The network compute fabric will ensure optimized application execution with real end-to-end (E2E) performance guarantees, spanning both the connectivity and compute domains. In this context, pervasive should be interpreted as "everywhere" [9]. Billions of networked smartphones and powerful connected devices in combination with cloud and edge resources form an immensely powerful, unified compute infrastructure that has the potential to enable completely new, unprecedented applications and services that could not be easily accomplished with traditional, siloed infrastructures.

Real-time performance

Current and emerging use cases often require a combination of stringent real-time characteristics, such as low latency, high throughput, high reliability and scalability. E2E performance requirements demand that deterministic and reliable connectivity offerings are complemented with corresponding compute capabilities. Connectivity with an integrated real-time compute stack will enable the network platform to host and manage the critical tasks of real-time applications to guarantee true E2E performance.

Use cases such as robotics and ubiquitous augmented reality or virtual reality control imply strict latency bounds in combination with distribution and elasticity requirements. These are difficult to achieve with segmented control loops of individual platform components and application tasks. Additionally, virtualization technologies predominant in the cloud often add latency, decrease throughput and cause jitter in data flows. Such unpredictable characteristics affect the E2E performance of applications.

Edge computing involves bringing computing capabilities either closer to data sources or to application front ends. By building on this principle and taking it even further, utilization of shorter control loops will eliminate latency toward central services. However, data packets still pass through numerous layers of network and virtualization

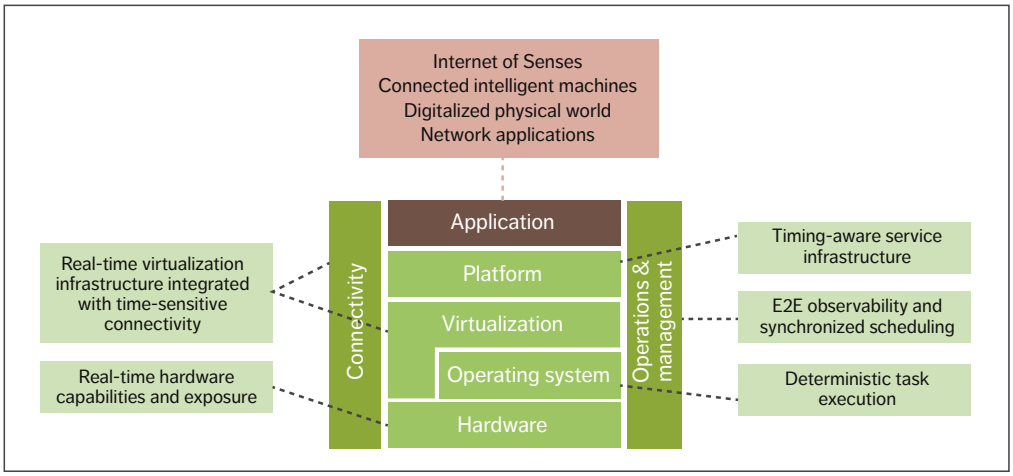


Figure 3 Real-time cloud stack supporting the network compute fabric performance characteristics

layers with fragmented control loops. A truly reliable real-time solution requires coordinated control of network and compute capabilities in a holistic way with full awareness of the application. Moreover, hypervisors and the network must operate in concert and be synchronized to avoid jitter and timing variations. This will require the network and compute components to be tightly related.

New network technologies such as Ultra-Reliable Low Latency Connectivity provide predictable wireless communication, while Time-Sensitive Networking for wired communication can bring network latency down to a few milliseconds. A compute infrastructure building on time-sensitive networks will integrate real-time properties on all layers of the stack, as shown in *Figure 3*. Data-plane services such as messaging and storage will provide guarantees on latency bounds, and services on the control and management planes have to support time-sensitive operations accordingly. Key components of such fabric are predictable networking and compute technologies, coupled with advanced resource management and scheduling features, realized at the operating system and hypervisor levels.

Distributed data

Unified data management together with computing across the network is the key to reducing data latency and transfer volumes while not impacting the quality of data-driven applications including artificial intelligence. The network compute fabric offers capabilities to utilize the available compute, storage and network resources more efficiently.

Allowing data to be processed closer to the data source significantly improves round-trip times and time to action/response for mission-critical use cases. Raw data does not have to be transmitted across the network but can be processed locally, and thereby be transformed to models or insights. As a consequence, only necessary real-time business insights, equipment maintenance predictions or other actionable data would be shared across network locations, as input to further federated learning.

Furthermore, the existence of a distributed data infrastructure within the unified execution environment makes it possible to utilize data shortcuts when application tasks access state or other types of data. These shortcuts can be realized with the help of modern cluster interconnect technologies that bypass the kernel and parts of the networking stack, such as RDMA (Remote Direct

> AUTOMATED MANAGEMENT LOOPS WILL MAKE THE FABRIC THE FOUNDATION OF NEW AUTONOMOUS APPLICATIONS

Memory Access), PCIe NTB (Peripheral Component Interconnect express Non-Transparent Bridging) and so on.

A distributed data infrastructure also makes it possible to offer advanced data services such as local caching mechanisms and data-centric near-memory and near-storage computing technologies, which can be very convenient for both developers and operations.

Intelligent operations

Manual and reactive operation practices often lead to suboptimal efficiency and performance. In the automated systems that are in use today, business intent is translated centrally to predefined workflows and then executed by individual domains. To achieve optimal efficiency and performance, however, it is essential to manage orchestration and operations across the entire ecosystem. This can be achieved by enhancing automated control and management processes, and applying data-driven learning and cognitive functions as part of unified cross-domain optimization models. With the network compute fabric, new operational models that can proactively evolve with the system will be able to manage the dynamics of network payloads and application workloads in real time.

The network compute fabric will utilize cognitive capabilities and data-driven automation to manage synergies across network, compute and storage domains. Distributed intelligence based on telemetry data from the network and compute infrastructure as well as application performance indicators will realize short feedback cycles for instant run-time optimizations of application deployment and infrastructure configurations.

Automated and self-maintained management loops will make the fabric the foundation for new autonomous applications. Examples of automation steps include:

- Predictive deployment strategies for applications and platform services
- Predictive resource availability and collocation for different domains
- Intelligent sharing and scheduling of rare resources such as hardware accelerators
- Qualitative trade-off analysis of conflicting business and operational intents
- Autonomous application deployment and life-cycle management
- Predictive fault management and platform resiliency.

Simplification and serviceability

The placement decisions for the distributed application components will be based on service coverage, E2E service requirements and infrastructure capabilities. The placement will also be adapted dynamically based on conditions of resource availability.

Such dynamic distribution of application components places a new set of requirements on the serviceability of applications. For instance, traffic flows between individual application components have to be managed automatically in order to maintain ongoing sessions. DevOps pipelines across ecosystem partners need to be an integral part of the fabric. This includes developer-friendly tracing and debugging capabilities for vastly distributed applications. These types of features and services need to be built into the fabric and exposed through convenient application programming interfaces to the application operators and developers.

The network compute fabric will offer possibilities to specify high-level business intents for services or applications. For instance, simple low-code/no-code interfaces will allow application providers to declare requirements associated with the deployment of the

distributed application. Across the service graph of an application, there may be performance requirements of varying stringency that make it possible for the fabric to automatically strike a balance between application performance requirements and operational costs.

Business opportunities of the network compute fabric

While the monetizable differentiators for each player will vary significantly, the network compute fabric will enable communication service providers (CSPs), developers and many other types of players to offer new services in a wide variety of contexts, including the Internet of Senses and connected intelligent machines.

Depending on the interplay of harmonized telecom and IT domains, these differentiators could be monetized based on the services the mobile network is expected to provide. The majority of 6G use cases will generate massive amounts of data, which will enable novel insights using the intelligent algorithms embedded in the network compute fabric.

The 6G world will require a shift in how CSPs do business in terms of exposing services to new types of customers. These services need to be bundled or aggregated from multiple domains into relevant offerings. The offering may range from pure connectivity services, to integrated connectivity and compute, to other value-added services. These types of services may be exposed directly to the customers or delivered by ecosystem partners that aggregate services for customers.

The network compute fabric will enable new business models by integrating the services of different ecosystem actors and providing new technologies for as-a-service models. Automated contract negotiation and fulfillment to support sales, delivery and charging operations can reduce business friction within various ecosystem constellations. Much of the interaction between the players will happen in software. New technologies based on distributed ledgers and smart contracts can enable new brokerless partnership models between providers.

Application-supporting services can play an important role in the ecosystem by simplifying application development for application providers. Further development of technologies and synergies across ecosystems will allow for additional business model innovation.

Conclusion

Our vision of the network compute fabric is based on expected 6G use cases such as the Internet of Senses and cyber-physical systems that will require a new set of capabilities beyond connectivity. These capabilities will not only serve demanding use cases but will also transform the network into an innovation platform characterized by its cognitive behavior, trustworthiness and adaptability.

The principal components of the network compute fabric facilitate unification across ecosystems, application management, execution environment, and exposure of network and compute capabilities, regardless of the heterogeneity of the infrastructure. The key features of the network compute fabric are pervasiveness, real-time performance, distributed data, intelligent operations, and simplification and serviceability.

The network compute fabric will be a pervasive, globally interconnected compute and storage platform that facilitates optimized handling of application components while giving the impression of locality. Real-time infrastructure and services together with unified data access are core elements of the network compute fabric, supporting distributed intelligence as well as simplification of deployment across heterogeneous infrastructure and administrative domains.

References

1. **Ericsson report, 10 hot consumer trends 2030, available at:** *https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/10-hot-consumer-trends-2030*
2. **Ericsson report, Connected intelligent machines, available at:** *https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/10-hot-consumer-trends-2030-connected-intelligent-machines*
3. **Ericsson white paper, Ever-present intelligent communication, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/a-research-outlook-towards-6g*
4. **Ericsson blog, 5G and cloud: How telecom can architect the next cloud era, February 4, 2021, Ekudden, E, available at:** *https://www.ericsson.com/en/blog/2021/2/5g-and-cloud*
5. **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*
6. **Ericsson Technology Review, The future of cloud computing: Highly distributed with heterogeneous hardware, May 12, 2020, John, W; Sargor, C; Szabo, R; Javed Awan, A; Padala, C; Drake, E; Julien, M; Opsenica, M, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/the-future-of-cloud-computing*
7. **OneAPI website, available at:** *http://www.oneapi.com/*
8. **The Derecho Project website, available at:** *https://derecho-project.github.io/*
9. **Ericsson blog, Why the world needs a pervasive Network Compute Fabric, February 18, 2021, John, W; Persson, P; Sefidcon, A, available at:** *https://www.ericsson.com/en/blog/2021/2/pervasive-network-compute-fabric*

Further reading

- **Ericsson website, Network compute fabric, available at:** *https://www.ericsson.com/en/future-technologies/network-compute-fabric*
- **Ericsson blog, What is computing fabric in the network?, available at:** *https://www.ericsson.com/en/blog/2020/4/computing-fabric-network*

THE AUTHORS

Azimeh Sefidcon

◆ joined Ericsson in 2010. Since then she has held various technology development and leadership roles in the areas of mobile and core networks and software and system design. She currently serves as research director for cloud systems and platforms, where she sets the research agenda for the future vision of the network compute fabric, technologies for edge cloud, network-friendly cloud evolution and future computing platforms. Sefidcon holds a Ph.D. in mobility and IP from Concordia University in Montreal, Canada.

Wolfgang John

◆ is a principal researcher at Ericsson Research in Stockholm. His current research focuses primarily on edge and distributed cloud computing systems and platform concepts for both telco and IT applications. Since joining Ericsson in 2011, he has also done research on network function virtualization, software-defined networking and network management. John holds a Ph.D. in computer engineering from Chalmers University of Technology in Gothenburg, Sweden, and has coauthored more than 50 scientific papers and reports, as well as several patent families.

Miljenko Opsenica

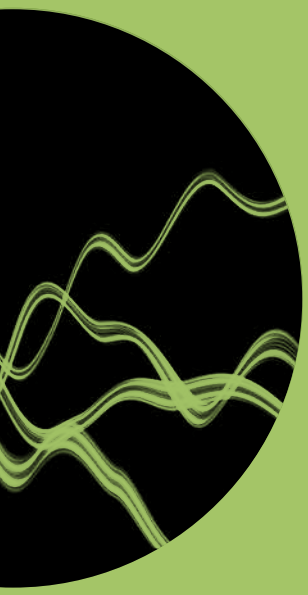
◆ joined Ericsson in 1998 and currently serves as a principal researcher at Ericsson Research in Finland (NomadicLab), where he is working on cloud architectures and technologies, orchestration frameworks and automation. Opsenica also leads Ericsson Research's integrated connectivity and edge program, which focuses on integrated architecture and concepts for enabling collaborative edge ecosystems. He holds an M.Sc. in electrical engineering and computing from the University of Zagreb in Croatia. He has coauthored several scientific papers and reports and holds numerous patents.

Björn Skubic

◆ is a research leader in cloud systems and platforms at Ericsson Research in Stockholm. He leads a research team that focuses on service enablement, service exposure and monetization of telco services in edge ecosystems. He joined Ericsson in 2008 and has

been active in the areas of fixed networks, transport and infrastructure for mobile networks and energy efficiency. Skubic holds a Ph.D. in physics from Uppsala University in Sweden and has coauthored more than 60 scientific papers as well as several book chapters and patent families.

The authors would like to thank Henrik Voigt, Carlos Bravo, Benedek Kovács, Hannes Medelius, Andrew Williams, Fetahi Wuhib, Joacim Halén and Zoltán Turányi for their contributions to this article.



ISSN 0014-0171
284 23- 3360 | Uen

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