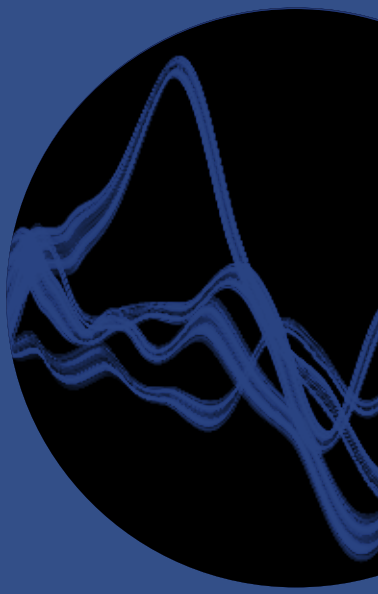
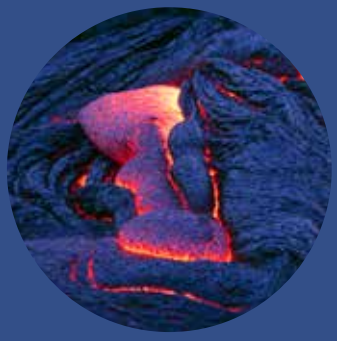
CHARTING THE FUTURE OF INNOVATION | VOLUME 97 I 2018-02

ERICSSON TECHNOLOGY Review



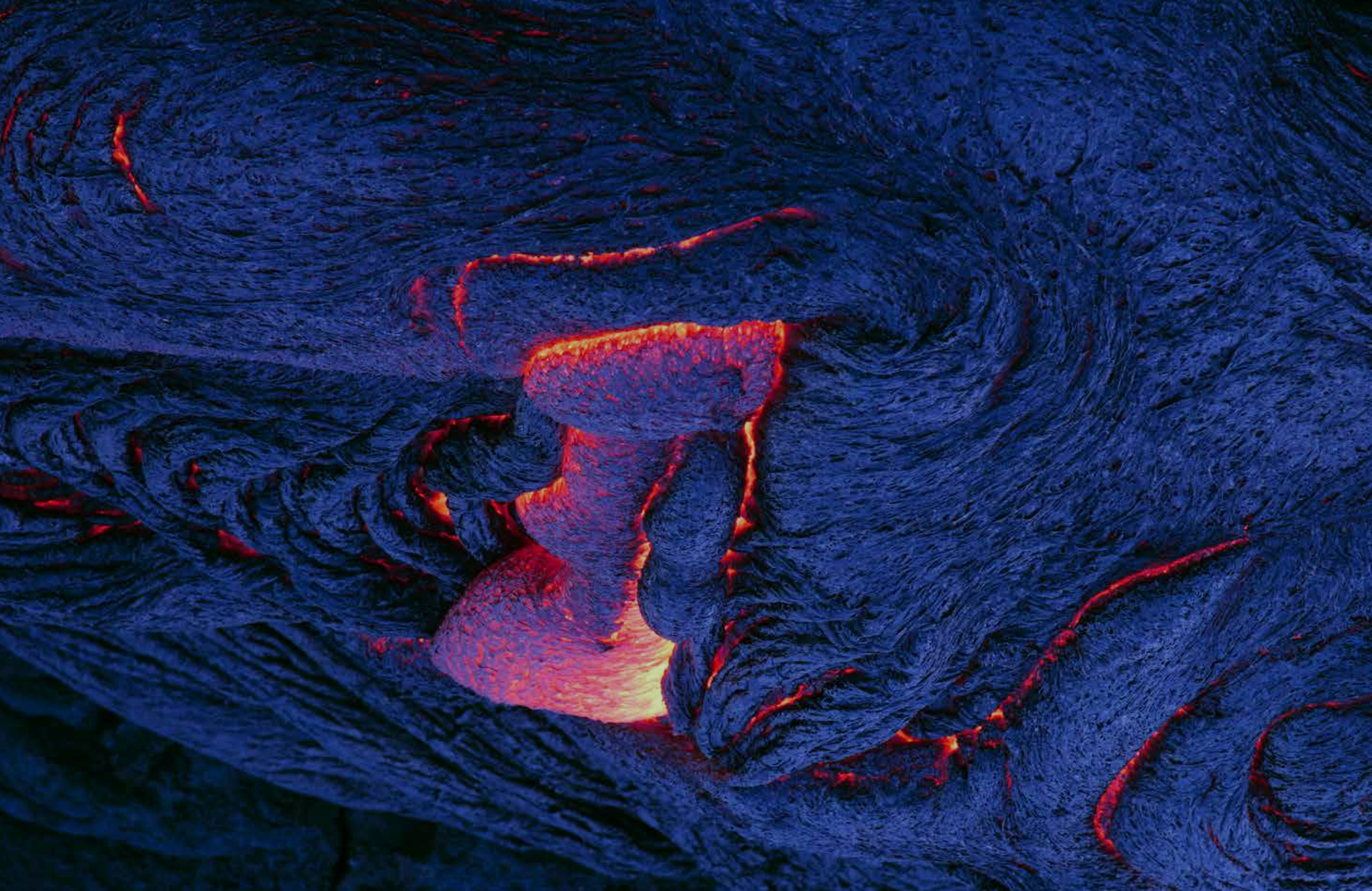
TECHNOLOGY TRENDS AUGMENTING THE CONNECTED SOCIETY

DIGITAL CONNECTIVITY MARKETPLACES



FIXED WIRELESS ACCESS IN LTE AND 5G

ERICSSON



Ericsson Technology Review brings you insights into some of the key emerging innovations that are shaping the future of ICT. Our aim is to encourage an open discussion about the potential, practicalities, and benefits of a wide range of technical developments, and provide insight into what the future has to offer.

ADDRESS
Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 8 7190000

PUBLISHING
All material and articles are published on the Ericsson Technology Review website:
www.ericsson.com/ericsson-technology-review

PUBLISHER
Erik Ekudden

EDITOR
Tanis Bestland (Nordic Morning)
tanis.bestland@nordicmorning.com

EDITORIAL BOARD
Håkan Andersson, Anders Rosengren, Mats Norin, Erik Westerberg, Magnus Buhrgard, Gunnar Thrysin, Håkan Olofsson, Dan Fahrman, Robert Skog, Patrik Roseen, Jonas Högberg, John Fornehed and Sara Kullman

FEATURE ARTICLE
Five technology trends augmenting the Connected Society by Erik Ekudden

ART DIRECTOR
Liselotte Eriksson (Nordic Morning)

PRODUCTION LEADER
Susanna O'Grady (Nordic Morning)

LAYOUT
Liselotte Eriksson (Nordic Morning)

ILLUSTRATIONS
Nordic Morning Ukraine

CHIEF SUBEDITOR
Ian Nicholson (Nordic Morning)

SUBEDITORS
Paul Eade and Penny Schröder-Smith (Nordic Morning)

ISSN: 0014-0171
Volume: 97, 2018

TRANSFORMING TO FIT THE **FUTURE REALITY**

■ **TECHNOLOGY DEVELOPMENT** keeps getting faster and more interconnected, with new innovations appearing every day. As a result, we're swiftly moving toward the realization of the "Augmented Connected Society" – a world characterized by ubiquitous internet access for all, self-learning robots and truly intuitive interaction between humans and machines. But how can our industry best prepare for this future? In my role as CTO, I have the challenging and exhilarating annual task of identifying the five technology trends of the future that are (or will be) most relevant to our industry. You can find my insights and reflections on this year's trends on page 30.

The augmentation of human intelligence is one of the key themes in this year's trends article. Creating the highly automated environment that network operators and digital service providers will need in the future requires the support of intelligent agents that are able to work collaboratively. The two proofs of concept presented in the Cognitive Technologies article in this issue demonstrate how the combination of machine reasoning and machine learning techniques makes it possible to create intelligent agents that are able to learn from diverse inputs, and share or transfer experience between contexts.

I believe that fixed wireless access (FWA) is likely to play an important role on our journey toward a world characterized by ubiquitous internet access for all. Already today, LTE and 5G New Radio are opening up significant commercial opportunities for operators to use FWA to bring the internet to many of the more than one billion households around the world that are still unconnected, as well as to many small and medium-sized businesses. The FWA article in this issue highlights the key principles for combined mobile broadband and FWA deployments, and presents a use case that illustrates the recommended deployment approach.

In my discussions with people from different industries this past year, I've noticed a growing awareness that telecom infrastructure can bring significant value to the digital transformation of industries. However, many providers need help to understand the requirements of this new market and to create solutions for it. The Digital Connectivity Marketplaces article in this issue explains our concept for a platform model in which capabilities from many providers can be effectively packaged and exposed in attractive ways to buyers from different industries.

There's no doubt that operations support systems (OSS) have a critical role to play in the future of our industry. Check out the OSS article in this issue to learn about our future OSS concept, which is built on a solid implementation architecture that enables the use of industrydefined interfaces and open-source modules, as well as integration with full component compatibility.

Finally, I want to encourage those of you who are concerned about bufferbloat to check out the Virtual Active Queue Management (vAQM) article in this issue. In our innovative concept, vAQM is centralized and applied per user and per flow, adapting to link rate fluctuation and considering the current bottleneck link bandwidth to the user. Our testing has shown that when vAQM is centralized upstream, rather than being deployed in the bottleneck nodes as it is in classic AQM, there is a substantial reduction in bufferbloat.

I hope you get a lot of value out of this issue of Ericsson Technology Review, and that the articles can serve as the basis for stimulating future-focused discussions with your colleagues and business partners. If you would like to share a link to the whole magazine or to a specific article, you can find both PDF and HTML versions at https://www.ericsson.com/en/ericsson-technology-review

> "THE AUGMENTATION OF HUMAN INTELLIGENCE IS ONE OF THE KEY THEMES IN THIS YEAR'S TRENDS ARTICLE"

ERIK EKUDDEN
SENIOR VICE PRESIDENT
AND GROUP CTO

OPEN, INTELLIGENT AND

model-driven: evolving OSS

The simplicity required to deploy and manage services in future network and infrastructure operations is driving the need to rethink traditional operations support systems (OSS). To unleash the potential of 5G, Network Functions Virtualization (NFV) and software-defined networking (SDN), future OSS need to be open, intelligent and able to support model-driven automation.

MALGORZATA SVENSSON, MUNISH AGARWAL, STEPHEN TERRILL, JOHAN WALLIN

The long-standing paradigm of bundling vendor and domain-specific management with network functions has been challenged in recent years by a new approach that is built on the concept of an open and model-driven platform. This new approach leverages open source and uses standardized interfaces to enable a horizontal management platform.

Today's service providers expect rapid network deployments, agile introduction of new services and cost-efficient operation and management – requirements that are even more important when planning for future 5G capabilities. Recent technology trends make it possible to redefine traditional support systems at the same time as they enable greater efficiency in development and operations.

For example, model-driven automation eliminates the need for manual interaction by externalizing behavioral aspects of specific domains from application logic. Real-time and near-real-time analytics redefine the implementation and scope of support system functions like performance, fault management, assurance and optimization, where analytics models act on the collected and correlated data, producing insights that are far richer in scope than the outcome of the traditional functions. Machine intelligence (MI) algorithms enable the transition from reactive to proactive decision making.

The DevOps paradigm simplifies development and operational processes. It requires development and deployment-friendly software architecture and tool chains that support automation. Microservice architecture supports the DevOps paradigm and enables highly scalable, extendable and flexible systems.

Ericsson's approach to evolving OSS is based on combining these technologies and solutions to create a modern OSS architecture that is optimized to meet the requirements of service providers in the mid to long term.

Our OSS architecture principles

The key principles guiding Ericsson's OSS architecture concept are that it is service oriented, that the automation is analytics and policy driven, and that it uses virtualization and abstraction of network functions. By following these principles, we can evolve OSS to become programmable and able to leverage the interfaces offered by the production domains.

Definitions of key concepts

- **Analytics** is the discovery, interpretation and communication of meaningful patterns in data. It is the umbrella for AI, ML and MI.
- **Cloud native** is a term that describes the patterns of organizations, architectures and technologies that consistently, reliably and at scale take full advantage of the possibilities of the cloud to support cloud-oriented business models.
- **Microservice** is a small service supporting communication over network interfaces. Several such services constitute an application. Each microservice covers a limited and coherent functional scope and is independently manageable by fully automated life-cycle machinery. A microservice is small enough to be the responsibility of a single, small development team.
- **OSS business layer** is a conceptual layer of the OSS that guides the service provider through the operational support subset of its business process.
- **OSS control layer** is a layer of the OSS containing functionality that controls and manages production domains.
- **Policy** is a function that governs the choices in the behavior of a system. Policy can use a declarative or imperative approach.
- **Programmability** is the ability to externally influence the behavior of a system in a defined manner.
- **Resource** is a means to realize a service, and can be either physical or non-physical.
- **Service** is a means to deliver value to a customer or user; it should be understandable to a customer. Service is also a representation of the implementation aspects to deliver the value.

Terms and abbreviations

AI – artificial intelligence | **API** – application programming interface | **BSS** – business support systems | **ETSI** – European Telecommunications Standards Institute | **MANO** – Management and Orchestration | **MI** – machine intelligence | **ML** – machine learning | **NFV** – Network Functions Virtualization | **ONAP** – Open Network Automation Platform | **OSS** – operations support systems | **PNF** – Physical Network Function | **SDN** – software-defined networking | **SLA** – Service Level Agreement | **TOSCA** – Topology and Orchestration Specification for Cloud Applications | **UDM** – Unified Data Management | **VNF** – Virtual Network Function

> AUTOMATION CAN BE FURTHER ENHANCED BY APPLYING MI, WHICH IS THE COMBINATION OF MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Abstraction, programmability and domain-specific extensions

Networks are complex. As the physical resources within them transform into virtual resources such as Virtual Network Functions (VNFs), virtual switches, virtual networks and virtual cloud resources, they will expose only what is necessary and be managed in a uniform way. This leads to a logical representation of resources and networks, which is referred to as abstraction. The capabilities of these virtual resources are exposed through their interfaces, while their behavior is influenced by these interfaces. This is referred to as programmability. The methodology of abstraction and programmability is continually and recursively applied at all levels of the OSS and networks, from the resources' interfaces to the exposed services.

Abstraction and programmability are the main prerequisites for uniform management of automation, easing the shift from managing networks to managing automation. They enable simplification by exposing only the required capabilities of a domain, while still supporting an efficient interaction between management system scopes, and management systems and networks.

Regardless of the generalization of the behavior between various production domains, the domain-specific aspects remain. Some examples of domain-specific extensions are: managing and optimizing radio frequency, designing and handling data center capacity, or designing L2/L3 service overlay. Specific production domain competencies and processes are required to provide the necessary capabilities in these examples, even though the capabilities are realized by the same OSS.

From automatic management to management of automation

The goal of automation is to provide a zero-touch network, which means that automation moves from automating via scripting what can be done manually, to an autonomic network that takes care of itself and handles previously unforeseen situations. This changes the approach to management, from managing networks manually to managing automation.

A key part of managing automation is the control loop paradigm, as shown in *Figure 1*. This is achieved by designing the insights and policies required for a use case, and the decisions that need to be made. Insights are the outcome when data is processed by analytics. Policies represent the rules governing the decisions to be made by the control loop. The decisions are actuated by COM (Control, Orchestration, Management), which interacts with production domains by requesting actions such as update, configure or heal. The control loops are implemented by multiple OSS functions and can act in a hierarchical way.

By leveraging analytics, MI and policy, control loops become adaptive and are central to assuring and optimizing the deployed services and resources.

MI and autonomic networks

Automation can be further enhanced by applying MI, which is the combination of machine learning (ML) and artificial intelligence (AI). MI adapts to the situation in a network and learns to provide the best insights for a given network situation.

As the level of MI increases, the role of policy shifts from governance of the decisions to be made, to ensuring that the insights produced from MI are appropriate. In this sense, MI is an enabler for autonomic networks.

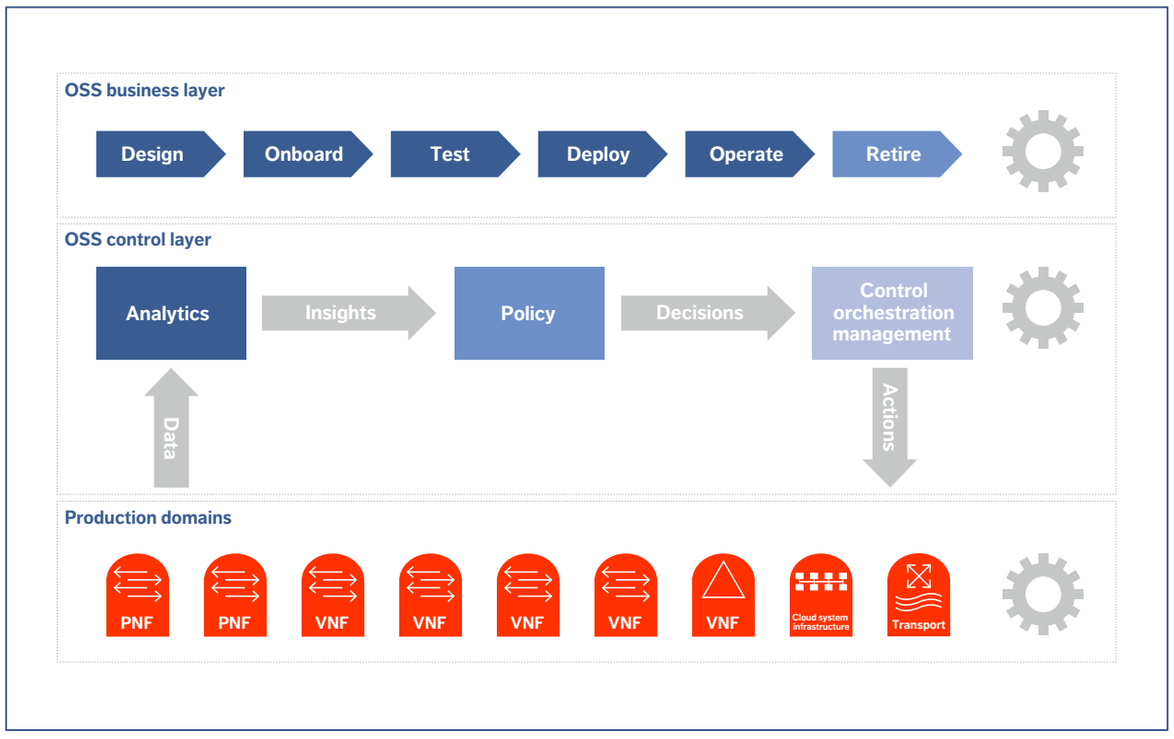


Figure 1 Analytics and policy-driven automation: closed control loop

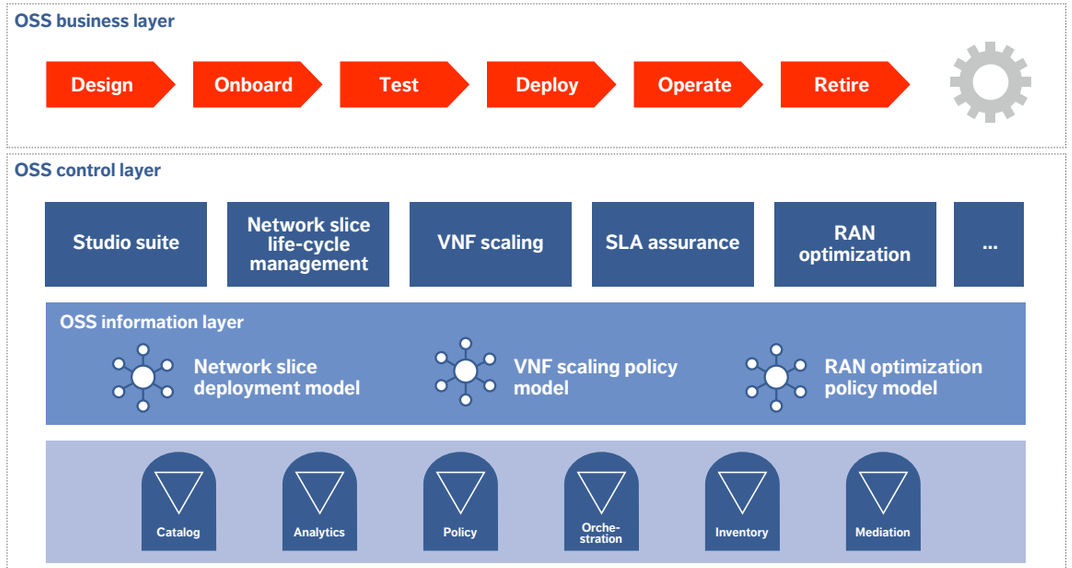


Figure 2 Model-driven OSS architecture

Model-driven architecture

To secure efficient and consistent management, the behavioral aspects of specific domains are externalized from application logic to models. There are models for configuration, orchestration, policy rules and analytics.

Resource and service life-cycle management is a good example of a model-driven approach. As shown in *Figure 2*, the models describe the services and the resources. These models are designed and onboarded in a studio suite. They are then used to test, deploy, operate and retire the services and the resources. The models are based on standard modeling approaches. For example, deployment models are described using HEAT (specifically, orchestration in OpenStack [1]) – and TOSCA [2]. The "operate" process can cover activities such as scale, upgrade, heal, relocate, stop and start.

The configuration and instrumentation of the executing resource and service instances require resource and service views expressed also as models. IETF YANG [3] has emerged as the mainstream modeling language for this purpose.

Policy rule sets define the behavioral aspects of service assurance, resource performance and scaling. Since the models are the way to capture business logic and intellectual property, they will become sellable objects themselves and be subject to commercial agreements.

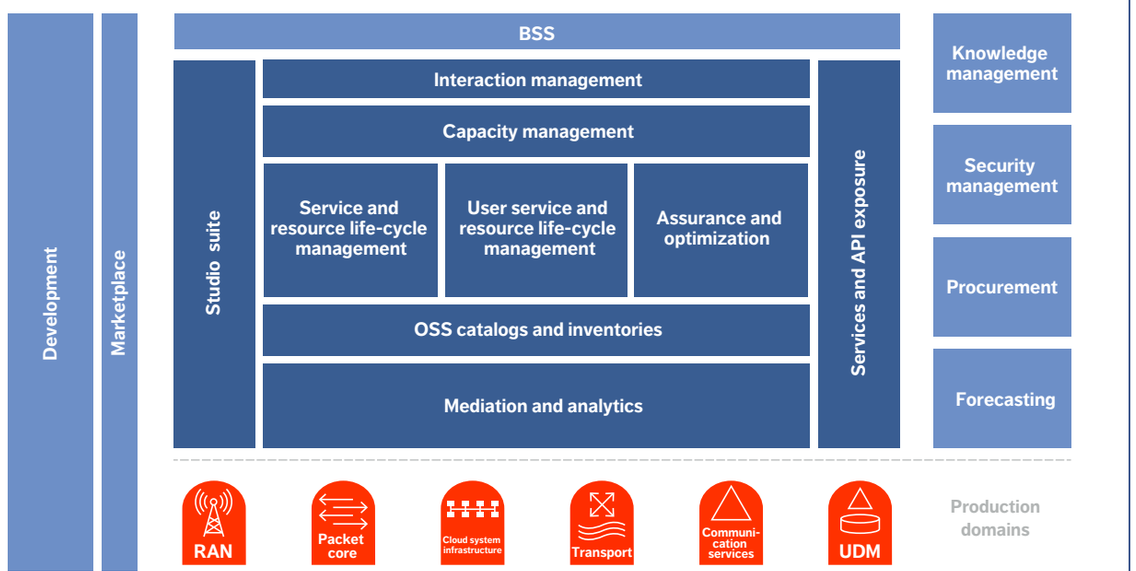


Figure 3 Ericsson's OSS architecture

Modular architecture and implementation capabilities

The optimal OSS of the future will be based both on the principles outlined above and on modular architecture and implementation capabilities.

Studio suite

Figure 3 illustrates the studio suite, which provides a single, coherent environment to design, test, operate, control and coordinate all the business, engineering and operational processes of a service provider. This is done by expressing the processes, interface extensions, policies, deployments and configurations as models and using them for automation of all the OSS functions and the managed production domains. The studio suite framework includes business, engineering and operational studios.

The business studio is the environment where business and operational processes are maintained. The business studio automatically orchestrates the process models by invoking operations on the appropriate functions in the OSS stack.

The engineering studio provides a model-driven platform where the OSS functions and the relevant information and information models associated with those functions are constructed and extended. It lets a user (human or machine) realize various management tasks by extending the OSS functions with use-case specific models for configuration,

orchestration, data and policies. It provides software development kits for defining the extension models and then instantiates the needed components, provisions the components with the models and wires them together to realize the desired use case. An example could be a set of ML functions that provide scalable machinery that can be assigned to multiple uses by applying different configuration and data models.

The operational studio offers capabilities to manage resources and services throughout their life cycle. Resources represent the infrastructure and the networks of production domains. Services use the infrastructure and networks to deliver value to customers. One of the first activities in the resource and service life cycle is to construct a specification of a resource or a service, then assign the required policy rules and configurations defining the behavior of the network, or the infrastructure or the service. Once they are defined, the specifications, policies and configurations are stored in the OSS catalogs and distributed in real time to be used by the OSS functions.

Managing services and resources

Service and resource life-cycle management functions provide a programmable way of managing the services and resources of the production domains based on specifications, policies and configurations stored in the catalogs. The core capabilities are orchestration and policy engines that parse the defined models to realize all the life-cycle steps of the services and resources. The policy and orchestration engines automate the network management by taking over the burden of low-level decision making and execution of life-cycle operations from human users.

These functions are triggered by events generated by surrounding functions like interaction management, assurance and optimization. They provide both closed loop optimization capabilities (where decision making is done based on policies and MI) and an open loop optimization feature (where a human user decides based on a set of automatic recommendations). This dynamic, model-driven and real-time interaction is what will make this future OSS solution different from the traditional one.

We can already see the need for seamless interaction between OSS software development and operations, which suggests that the core DevOps capabilities of continuous integration, delivery, release and deployment on-demand will be an integral part of the future OSS.

OSS inventories will be used to store instantiated resources and services. They track the presence, capacity and configuration of resources and services, as well as their relationships to other resources and services. They provide the data for capacity management, which enables capacity planning, design and monitoring over time. By using analytics in capacity management, it is possible to make it work in real time and be predictive and self-learning.

The purpose of the user service and resource life-cycle management function is to manage resources and services that have been assigned to a user, tenant or subscriber. This capability also supports user, tenant and subscriber provisioning. The closed and open control loop architecture patterns apply here as well. The user service and resource life-cycle management function has the same properties as the service and resource life-cycle management function.

Mediation provides an open and adaptable interface to collect data, process events generated by the production domains, and do the required translation, transformation and filtering. It also provides a set of flexible interfaces to invoke life-cycle operations on the networks and infrastructures of the production domains.

The assurance function makes it possible to check that resources and services behave as they have been designed, published and offered. Assurance makes use of analytics to discover insights about the performance of services and resources. Policy engines are then triggered to act on the insights to determine actions that ensure the desired performance level offered on services and resources.

Our OSS architecture is designed to use interaction management, services and application programming interface (API) exposure to interact with external functions. A few examples of those external functions are development, marketplaces, business support systems, advanced security and knowledge management systems, procurement and forecasting.

The production domain example shown in Figure 3 includes a RAN, packet core and communication services, Unified Data Management, cloud system infrastructure, fixed access and transport networks. Our OSS manages multi-vendor network and infrastructure.

The OSS functions are based on an API-first approach that mandates that an API be defined before the function-exposing services can be implemented. The approach enables parallel development and provides the ability to adapt to external implementations. The API management framework facilitates the effective use of APIs during development, discovery of internal API endpoints at runtime and efficiently controlled exposure of APIs to external applications.

Cloud native

Cloud native is the new computing paradigm that is optimized for modern, distributed systems environments, capable of scaling to tens of thousands of self-healing, multi-tenant nodes. Cloud native systems are container packaged, dynamically managed and microservice oriented. The Ericsson OSS of the future will be based on cloud-native architecture to enable better reuse, simplified operations and improved efficiency, agility and maintainability.

The microservice architecture enables fine-grained reuse by having several small components instead of a single large one. By being loosely coupled and backward compatible, the services can be developed independently, enabling efficient DevOps. It also allows each service to choose the most efficient development and runtime technology for its purpose. The microservice architecture makes it possible to compose various business solutions, then deploy them automatically, significantly reducing deployment cost and time to market. The architecture will use the cloud-native ecosystem as a portability layer to support multiple deployment options.

Security

Supporting multiple deployment models requires multiple security capabilities like credential management, secure communication, encryption, authentication and authorization, which can be selected based on the deployment. Our OSS architecture will include the capabilities needed to support the multiple deployment models and to protect the models.

Embracing industry initiatives

Due to NFV and programmable networks such as SDN, there is huge industry interest in automation to decrease complexity and achieve rapid introduction of services and resources. This has resulted in several significant industry initiatives, most notably

ETSI-MANO [4] and the Open Network Automation Platform (ONAP) [5].

Initiatives that call for standardization provide an environment to develop strong concepts and identify important interfaces, while open source provides reference implementations, as well as implementation interfaces.

ONAP

By bringing together the resources of its members, ONAP has sped up the development of a globally shared architecture and implementation for network automation. ONAP provides both design time and runtime reference implementations, covering service design and creation, catalog, inventory, orchestration and control, policies and analytics. It also includes service management and automation capabilities for Physical/Virtual Network Functions (PNFs/VNFs), transport and cloud infrastructure.

Ericsson's modular and model-driven OSS approach is well aligned with ONAP and extends the ONAP vision to automation beyond network management to include interaction management, workforce management, user service and resource life-cycle management, advanced service optimization and assurance as a few examples. To interact with ONAP, we are in the process of adopting the necessary interfaces and model definitions in our product portfolio (both in OSS and network functions). Ericsson is committed to ONAP and heavily engaged with it across a broad range of topics. We aim to use the ONAP-provided open source implementations in accordance with Ericsson's architecture principles.

Managing multiple production domains requires a model-driven architecture because achieving domain-specific behavior requires the application of appropriate models on a compatible OSS platform. Since a model-driven architecture is essential to drive automation, we are working toward ensuring that our models are usable within ONAP.

Conclusion

As both an established infrastructure and OSS supplier, Ericsson has a vision for OSS in which autonomic networks enable intelligent and automated service and resource life-cycle management. We are using analytics, MI, policy and orchestration technologies to create the OSS of the future, and to offer greater efficiency in development and operations. We apply a DevOps paradigm to our software development to keep us close to customers and speed up our software delivery.

As our approach is model-driven, OSS behavior is externalized from the management platform and materialized through a combination of models and configurations. As a result, the developed models are independent of the execution platform.

Our OSS concept is modular by design and follows microservice architecture principles that make it possible to replace software components with open source implementations. The concept is built on a solid implementation architecture that enables the use of industry-defined interfaces and open source modules, as well as integration with full component compatibility.

Ericsson has embraced open source implementations through active contributions and usage, as well as driving important standardizations. We are committed to driving industry alignment that will help create a healthy ecosystem.

References

1. **OpenStack, Heat Orchestration Template (HOT) Guide, available at:** *https://docs.openstack.org/heat/ocata/template_guide/hot_guide.html*
2. **OASIS, OASIS Topology and Orchestration Specification for Cloud Applications (TOSCA) TC, available at:** *https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca*
3. **Internet Engineering Task Force (IETF), October 2010, YANG – A Data Modeling Language for the Network Configuration Protocol (NETCONF), available at:** *https://tools.ietf.org/html/rfc6020*
4. **ETSI, Network Functions Virtualization, available at:** *http://www.etsi.org/technologies-clusters/technologies/nfv*
5. ONAP, available at: https://www.onap.org/

Further reading

- **Ericsson Technology Review, Technology trends driving innovation – five to watch, 2017, Ekudden, E:** *https://www.ericsson.com/en/publications/ericsson-technology-review/archive/2017/technology-trends-2017*
- **Ericsson Technology Review, Generating actionable insights from customer experience awareness, September 2016, Niemöller, J; Sarmonikas, G; Washington, N:** *https://www.ericsson.com/en/publications/ericsson-technology-review/archive/2016/generating-actionable-insights-from-customer-experience-awareness*
- **Ericsson Technology Review, DevOps: fueling the evolution toward 5G networks, April 2017, Degirmenci, F; Dinsing, T; John, W; Mecklin, T; Meirosu, C; Opsenica, M :** *https://www.ericsson.com/en/publications/ericsson-technology-review/archive/2017/devops-fueling-the-evolution-toward-5g-networks*
- **Ericsson, Gearing up support systems for software-defined and virtualized networks, June 2015:** *https://www.ericsson.com/en/news/2015/6/gearing-up-support-systems-for-software-defined-and-virtualized-networks-*
- **Ericsson, Zero touch networks with cloud-optimized network applications, Darula, M; Más, I:** *https://www.ericsson.com/assets/local/narratives/networks/documents/zero-touch-networks-with-the-5g-cloud-optimized-network-applications.pdf*
- **Ericsson Technology Review, Architecture evolution for automation and network programmability, November 2014, Angelin, L; Basilier, H; Cagenius, T; Más, I; Rune, G; Varga, B; Westerberg, E:** *https://www.ericsson.com/en/publications/ericsson-technology-review/archive/2014/architecture-evolution-for-automation-and-network-programmability*

THE AUTHORS

Malgorzata Svensson
◆ is an expert in OSS. She joined Ericsson in 1996 and has worked in various areas within research and

development. For the past 10 years, her work has focused on architecture evolution. She has broad experience in business process, function and information modelling; 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.

Munish Agarwal
◆ is a senior specialist in implementation architecture

for OSS/BSS. He has 19 years of experience in telecommunications working with product development, common platforms and architecture evolution. He has hands-on experience with mediation, rating, order management and interactive voice response. Agarwal holds a Bachelor of Technology degree from the Indian Institute of Technology Kharagpur.

Stephen Terrill
◆ has 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, 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 OSS, and he has been engaged in open source on ONAP's Technical Steering Committee. Terrill holds an M.Sc., a B.E. (Hons.) and a B.Sc. from the University of Melbourne, Australia

Johan Wallin
◆ is an expert in network management. He joined Ericsson in 1989 and has worked in several areas in R&D in various positions covering design, product management and system

management. He has broad experience of systems architecture in various mobile telephony systems including GSM, TDMA and CDMA. For the past 10 years, he has been working with network management architectures from an overall Ericsson perspective. He holds an M.Sc. in computer engineering and computer science from the Institute of Technology at Linköping University, Sweden.

The authors would like to acknowledge the contribution their colleagues John Quilty, Bo Åström, Ignacio Más, Jan Friman, and Jaco Fourie made to the writing of this article.

LOW LATENCY, HIGH FLEXIBILITY

Virtual AQM

Virtual Active Queue Management is an innovative approach to buffer management that does not require deployment in the network nodes where the buffering takes place. Instead, it is deployed as a centralized network function, which allows for simpler configuration and increased flexibility compared to traditional Active Queue Management deployments.

MARCUS IHLAR,
ALA NAZARI,
ROBERT SKOG

To minimize the effects of bandwidth fluctuations in mobile networks, radio base stations are typically provisioned with large buffers. However, since excessive buffering of packets causes a large perceived delay for the application consuming the data, it is important for the network to detect and properly manage cases in which buffering becomes excessive.

Active Queue Management (AQM) is a well-established technique for managing the frequent and substantial fluctuations in bandwidth and delay that are common in networks. It has been applied to network buffers for many years with the purpose of detecting and informing data senders about incipient traffic congestion. Traditional AQM tends to work well in fixed networks, where bottleneck link bandwidth varies little over short time durations. However, the deployment of traditional AQM is not widespread in mobile networks.

The problem of bufferbloat

Bufferbloat is an undesirable phenomenon prevalent in packet networks in which excessive buffering results in unnecessarily large end-to-end latency, jitter and throughput degradation. It occurs due to large queues that absorb a huge amount of traffic in a congested path and usually causes a long queuing delay. To prevent packet losses, networks nodes rely on large buffers.

Bufferbloat occurs mostly at edge network nodes and defeats the built-in TCP congestion avoidance mechanism, which relies on dropped packets to find the ideal send rate for a given end-to-end link. Mobile networks are provisioned with large buffers to handle bursty traffic, user fairness and radio channel variability. Even though these oversized buffers prevent packet loss, the overall performance degrades, as bufferbloat has major impacts on congestion control at endpoints. It misleads loss-based congestion control algorithms, resulting in packet overshooting on the sender side. Over-buffering also results in large jitter of end-to-end latency, which potentially brings in timeout events and eventually leads to throughput degradation.

Application impact

High latency and jitter cause significant problems for interactive applications such as VoIP, gaming and videoconferencing. Web browsing is also highly dependent on latency, and page load times grow linearly as round-trip time (RTT) increases.

Video streaming with adaptive bitrate through a continuous sequence of small HTTP downloads suffers when packets are lost. Under good conditions, video is transferred over the network at the same rate as it is played back. If the network can't keep up with video playback, the player usually switches to a lower video rate representation and reasonable video quality can be achieved with less bandwidth.

The main challenges in video streaming are packet loss and excessive buffering. When a packet is lost, the video player stops receiving data until the lost packet has been retransmitted, even though data packets continue to arrive. Video players normally buffer video segments, allowing them to continue playing from their buffer while downloading video segments that might involve retransmissions. Buffering is normally used for video on demand and timeshifted video but is limited for live video.

Terms and abbreviations
ACK – acknowledgement | **AMBR** – aggregate maximum bitrate | **AQM** – Active Queue Management | **CoDel** – Controlled Delay | **CPE** – Customer-premises equipment | **DN** – Data Network | **DNS** – Domain Name System | **DSCP** – Differentiated Services Code Point | **DRR** – Deficit Round Robin | **EPC** – Evolved Packet Core | **EPS** – Evolved Packet Switched System | **FQ** – Flow Queue | **FQ-CoDel** – Flow Queue Controlled Delay | **GBR** – guaranteed bitrate | **IETF** – Internet Engineering Task Force | **PDN GW**– public data network gateway | **PDU** – protocol data unit | **QFI** – QoS Flow Identifier | **QUIC** – Quick UDP Internet Connections | **RTT**– round-trip time | **SDF** – Service Data Flow | **SMF** – Session Management Function | **SYN** – synchronization | **TCP** – Transmission Control Protocol | **UE** – user equipment | **UPF** – user plane function | **vAQM** – virtual Active Queue Management

However, if multiple TCP packets are lost, TCP may slow down its sending rate to well below the data rate of the video, forcing the player into the rebuffering state.

In short, bufferbloat degrades application performance and negatively impacts the user experience. There is therefore a pressing need to control latency and jitter in order to provide desirable QoS to users.

Mitigation

Traditional AQM mitigates bufferbloat by controlling queue length through dropping or marking packets from the bottleneck buffer when it becomes full or when the queuing delay exceeds a threshold value. AQM reacts to congestion by dropping packets and thereby signaling to the sender that it should restrict the sending data rate due to the imminent overload. The short wait time also improves the response time for the transport protocol error handling.

> WITH FQ-CODEL, IT IS POSSIBLE TO REDUCE BOTTLENECK DELAYS BY SEVERAL ORDERS OF MAGNITUDE

Controlled Delay

Controlled Delay (CoDel) is an emerging AQM scheme designed by the IETF (Internet Engineering Task Force) [1]. Its goal is to contain queuing latency while maximizing the throughput. CoDel attempts to limit bufferbloat and minimize latency in saturated network links by distinguishing good queues (the ones that empty quickly) from bad queues that stay saturated and slow. CoDel features three major innovations that distinguish it from prior AQMs. First, it uses the local minimum queue as a measure of the standing/persistent queue. Second, it uses a single state-tracking variable of the minimum delay to see where it is relative to the standing queue delay. Third, instead of measuring queue size in bytes or packets, it measures the packet sojourn time in the queue and uses it as a metric to detect incipient congestion.

CoDel works as follows: upon arrival, a packet is enqueued if there is room in the queue. While enqueuing, a timestamp is added to the packet. When dequeuing, the packet sojourn time is calculated. CoDel is either in a dropping state or a non-dropping state. If the packet sojourn time remains above the target value (default 5ms) for a specified interval of time (default 100ms), CoDel enters the dropping state and starts dropping packets at the head of queue. While CoDel is in dropping state, if the packet sojourn time falls below the target or if the queue does not have sufficient packets to fill the outgoing link, CoDel leaves the dropping state.

The CoDel parameters target and interval are fixed parameters and their values have been chosen based on the observations from several experiments. Interval is used to ensure that the measured minimum delay does not become too stale. It should be set at about the worst-case RTT through the bottleneck to give endpoints sufficient time to react. The default value is 100ms. The target parameter is the acceptable queue delay for the packets, including capacity to cope with traffic spikes. A target of 5ms works well in most situations; lower values can reduce the throughput. If the target is more than 5ms, there is minor or no improvement in utilization.

Flow Queue Controlled Delay

Flow Queue Controlled Delay (FQ-CoDel) [2] is a standard AQM algorithm that ensures fairness in CoDel. It is useful because AQM algorithms working on single queues exacerbate the tendency of unfairness between the flows. FQ-CoDel is a hybrid scheme that classifies flows into one of multiple queues, applying CoDel on each queue and using modified Deficit Round Robin (DRR) scheduling to share link capacity between queues.

With FQ-CoDel, it is possible to reduce bottleneck delays by several orders of magnitude. It also provides accurate RTT estimates to elephant flows, while allowing priority access to shorter flow packets [3].

In combination with improvements in Linux network stacks, CoDel and FQ-CoDel can eliminate the problem of bufferbloat on Ethernet, cable, digital subscriber lines and fiber, resulting in network latency reductions of two to three orders of magnitude, as well as improvements in goodput [4].

In mobile networks, however, AQM needs to adjust to link bandwidth fluctuation that varies for different users. This is due to the fact that bottleneck link bandwidth can vary significantly in mobile networks even over short time durations, which causes queue build-up that can add significant delays when link capacity decreases.

Our innovative concept for virtual AQM (vAQM) is applied per user and per flow, adapting to link rate fluctuation and considering the current bottleneck link bandwidth to the user. Bottleneck link bandwidth is determined by correlating the number of bytes in flight with RTT measurements. vAQM is similar to FQ-CoDel but it uses the measured time-varying link bandwidth to the user as the dequeuing rate instead of non-varying deficit/quantum as in FQ-CoDel.

Virtual AQM

Traditional AQM is typically designed to operate on queues directly. However, due to deployment constraints and other factors, it is not always straightforward to deploy new AQMs at the network edges where congestion is likely to be experienced. One way of getting around this problem is to move the bottleneck to a more central point in the network and apply the AQM there. In a fixed network, it is easy to achieve this by shaping traffic to a rate slightly below the bottleneck peak capacity. Since mobile networks exhibit vastly different properties of fluctuating bandwidth, a more refined approach is required to apply centralized AQM.

> OUR INNOVATIVE CONCEPT FOR VIRTUAL AQM IS APPLIED PER USER AND PER FLOW

Virtual AQM consists of two separate but interrelated components: a bottleneck modeler and a queue manager. A separate vAQM instance is required for each bottleneck. In LTE, this corresponds roughly to a bearer (dedicated or default), while in 5G it corresponds to a Protocol Data Unit (PDU) session or a QoS flow. To effectively model the bottleneck, the vAQM measures the RTT of packets between itself and the user equipment (UE), as well as the amount of inflight data.

Virtual AQM maintains four variables: Path RTT, Average RTT, Target Queue Delay and Target RTT. Path RTT represents the propagation delay between the vAQM and the UE at the receiving end, without queuing delay. It is calculated by feeding RTT measurement samples to a min filter. Due to the varying nature of radio networks, the Path RTT must be occasionally revalidated.

Average RTT is an exponentially weighted moving average of the measured RTT, which represents the current observed delay. The use of a weighted moving average reduces the impact of inherent noise in the RTT signal, which typically arises from path layer characteristics such as Radio Link Control layer retransmissions.

A certain bottleneck queue delay must be tolerated to ensure smooth delivery of data over the radio. Target Queue Delay represents the delay that is tolerated on top of Path RTT before vAQM takes some sort of action. Target Queue Delay is defined as a fraction of Path RTT. The sum of the Path RTT and the Target Queue Delay is the Target RTT.

These four variables are used to determine whether queuing is taking place in network bottlenecks. The information can be used in two distinct ways: either as direct input to an AQM algorithm or as input to a bitrate shaper.

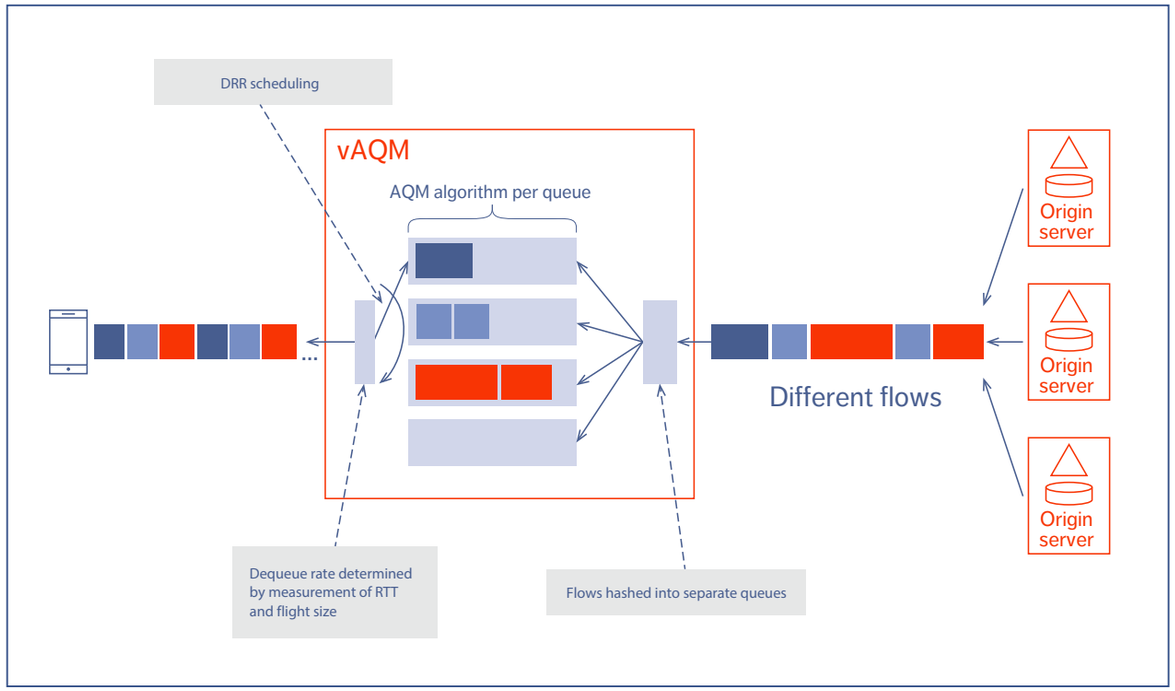


Figure 1: vAQM with dynamic bitrate adjustment and flow separation

vAQM with dynamic bitrate adjustment

In vAQM with dynamic bitrate adjustment, the traffic shaper dynamically adjusts its dequeuing rate based on the measured sending rate and the relation between Average RTT and Path RTT. The rate of incoming data is measured and applied to a max filter, while the average RTT is less than or equal to Target RTT. When Average RTT exceeds the Target RTT, the shaper will set the dequeue rate to the largest bitrate observed while Average RTT was below target. Furthermore, the difference between Average RTT and Target RTT is used as an additional bounding factor on the dequeue rate.

Internal queues will start to form if packets arrive at the vAQM at a higher rate than the dequeue rate. In such cases, a standard AQM algorithm such as FQ-CoDel can be applied to those queues. Path RTT can be used as input to the AQM algorithm to determine the CoDel interval and other variables.

Figure 1 demonstrates how vAQM assigns different flows to different queues managed by CoDel and dequeues with DRR using the current rate to the UE.

vAQM without bitrate adjustment

The vAQM without bitrate adjustment approach sacrifices the flow isolation that is achieved by FQ for reduced implementation and runtime complexity. This approach uses the RTT measurements as direct input to an AQM algorithm without applying the shaping step.

The AQM makes its drop/mark decisions based on the relation between Average RTT, Path RTT and the Target Queue Delay. If the Average RTT continuously exceeds the Target RTT for an interval amount of time, the AQM enters a dropping state, like that in CoDel. The interval is determined as a function of Path RTT. If multiple flows are traversing the same bottleneck, a drop will occur on the flow that is determined to be the most significant contributor of queue delay. This is effectively determined by comparing the flight sizes of the respective flows.

Handling of non-responsive flows

The basic principle of AQM is that it provides feedback to the congestion control mechanism at the sending endpoint. However, there is a possibility that senders will not react as expected to the signals that the AQM provides. The vAQM can detect such flows and apply flow-specific shaping to a level where the flow does not generate congestion in the bottleneck. This feature makes it safe to deploy and experiment with new congestion control algorithms, as it reduces the potential harm to the network and other flows sharing the same bottleneck.

Rate and RTT measurement

Virtual AQM relies on passive measurements of throughput and RTT to detect congestion. TCP is unencrypted at the transport level and can be measured trivially by storing sequence numbers and observing acknowledgements (ACKs) on the reverse path.

When the transport layer network protocol QUIC (Quick UDP Internet Connections) is used, a cryptographic envelope hides most transport mechanisms. Most notably, the ACK frame is completely hidden from on-path observers. A specific mechanism that allows passive on-path measurement of RTT is being proposed in QUIC standardization [5]. This mechanism makes use of a single bit in the QUIC short header, the value of the bit cycles with a frequency corresponding to the RTT between two communications endpoints. However, at the time of writing, neither the QUIC protocol nor the described mechanism is a finished standard.

The additional benefits of vAQM

The benefits of traditional AQM are well known, namely: fast delivery of short-lived flows and increased performance of latency-critical applications. There is an increased responsiveness to short-lived flows due to scheduling at flow level and minimal queue delay. For example, Domain Name System (DNS) responses and TCP SYN/ACK packets do not have to share a queue with a TCP download of bulk data.

In recent years, it has become increasingly common for different types of applications to share a single bottleneck buffer. Latency-critical applications that run in parallel with bulk download applications such as video streaming or file downloads typically experience severe performance degradation. By ensuring that the standing queues at the bottleneck are small or non-existent, large flows will not degrade latency.

In addition to the benefits of traditional AQM, vAQM is adaptive to link rate fluctuation and considers the current bottleneck link bandwidth to the user. It also provides the benefits of centralized AQM and modularity for increased flexibility. vAQM is able to simplify the deployment of AQM in an operator network because embedding it in a single (aggregation) node is sufficient to enable the management of the entire network downstream from the point of deployment. This means that legacy routers and radio schedulers do not need to be upgraded to include AQM. While vAQM currently uses a scheme resembling FQ-CoDel, it is built to support pluggable AQM modules, so that advances in the field of AQM and congestion control can have rapid impact in live networks.

vAQM testing

We have tested our vAQM concept on the Ericsson lab LTE network, embedding it in a multi-service proxy that inspects traffic to determine optimal bottleneck link capacity to UEs.

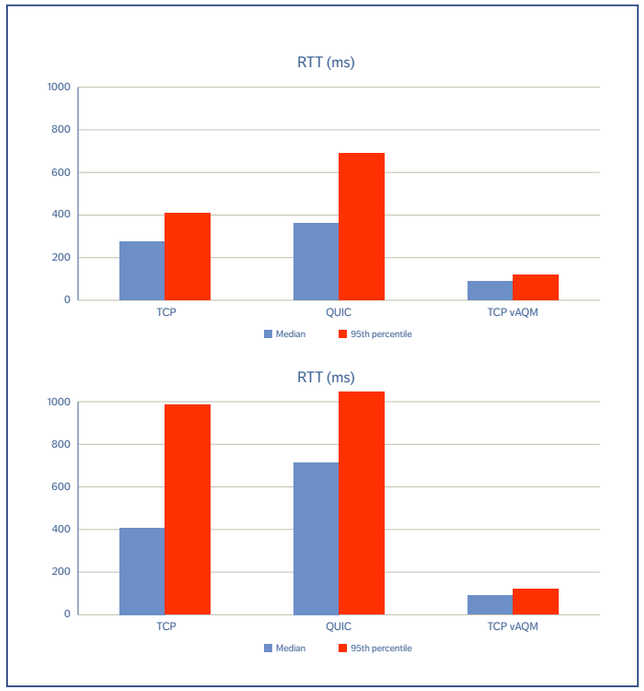


Figure 2: vAQM in full signal strength testing scenario (top) and in weak signal strength scenario (bottom)

We conducted the testing according to two multiuser cell scenarios: at full-signal strength and at weak-signal strength.

The test application consisted of loading a Google Drive page and downloading a 16MB image. We used three mobile devices to generate the background traffic for the multiuser cell, with each of them downloading 1GB files, so that we could test latency under load for multiple streams. The page was fetched using both TCP and QUIC, and the eNodeB was configured with a drop-tail queue. The server was configured to use the CUBIC congestion control algorithm for both TCP and QUIC.

We ran the tests for each scenario in three different ways:

- TCP with vAQM disabled
- QUIC with vAQM disabled
- TCP with vAQM enabled

The results, shown in *Figure 2*, demonstrate a reduction of the median RTT by roughly three times with vAQM enabled in the strong-signal scenario. In the weak-signal scenario, the median RTT with vAQM is not much different from the strong-signal scenario, whereas the traffic without AQM displays increased latency with a 95th percentile RTT of approximately one second for both QUIC and TCP traffic.

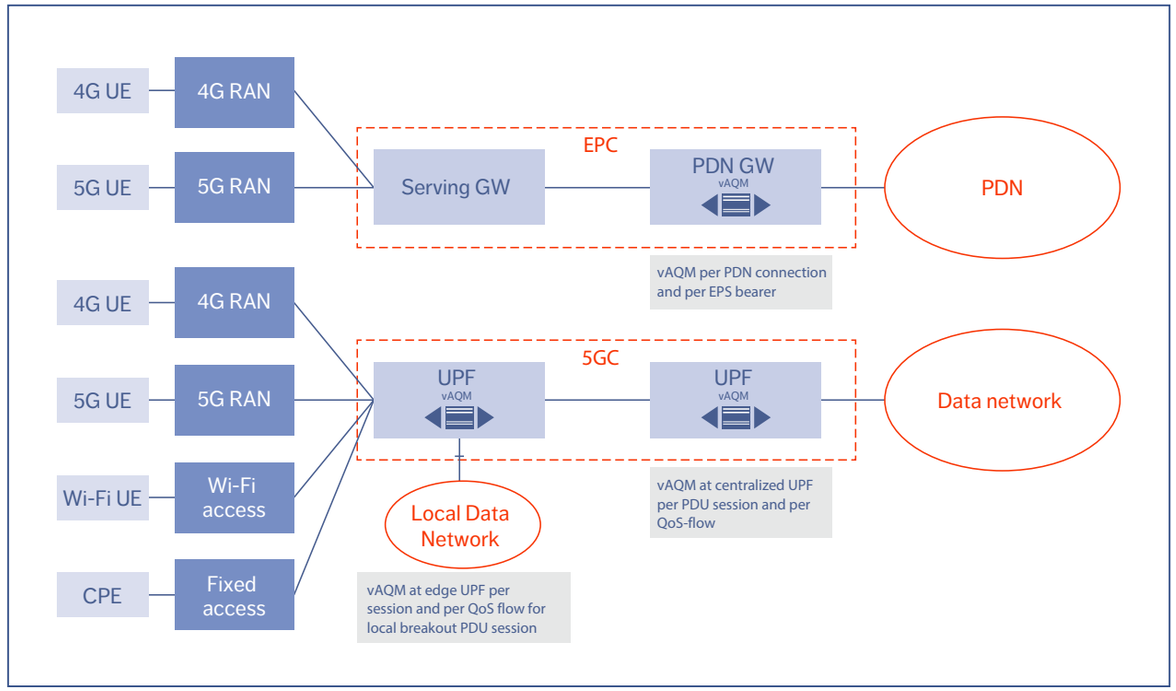


Figure 3: vAQM at PDN GW of EPC and at UPF of 5GC

vAQM in packet core

Figure 3 illustrates vAQM in two scenarios: Packet Data Network Gateway (PDN GW) of Evolved Packet Core (EPC) in 4G and in the User Plane Function (UPF) of 5G core.

In 5G, vAQM will be deployed as a network function by the UPF of the 5G core. It will be configured per PDU session, per SDF (Service Data Flow)/QoS flow – that is, the guaranteed bitrate (GBR) or non-GBR QoS flow. The non-GBR QoS flows of one PDU session will be aggregated and handled by one vAQM, whereas each GBR QoS flow will be allocated a vAQM with a dedicated virtual buffer.

The Session Management Function (SMF) will configure vAQM and its virtual buffers for the QoS flows during the PDU session establishment/modification phase. Configuration will occur when SMF provides the UPF with a QoS Enforcement Rule that contains information related to QoS enforcement of traffic including an SDF template (in other words, packet filter sets), the QoS-related information including GBR and maximum bitrate, and the corresponding packet marking information – in other words, the QoS Flow Identifier (QFI), the transport-level-packet-marking value. vAQM is self-learning, making it fully automated without any parametrization. When sending the packets on the PDU session to the UE, vAQM will read out the packets from the queues, recalculate the rate to the UE and adjust the bottleneck. In cases

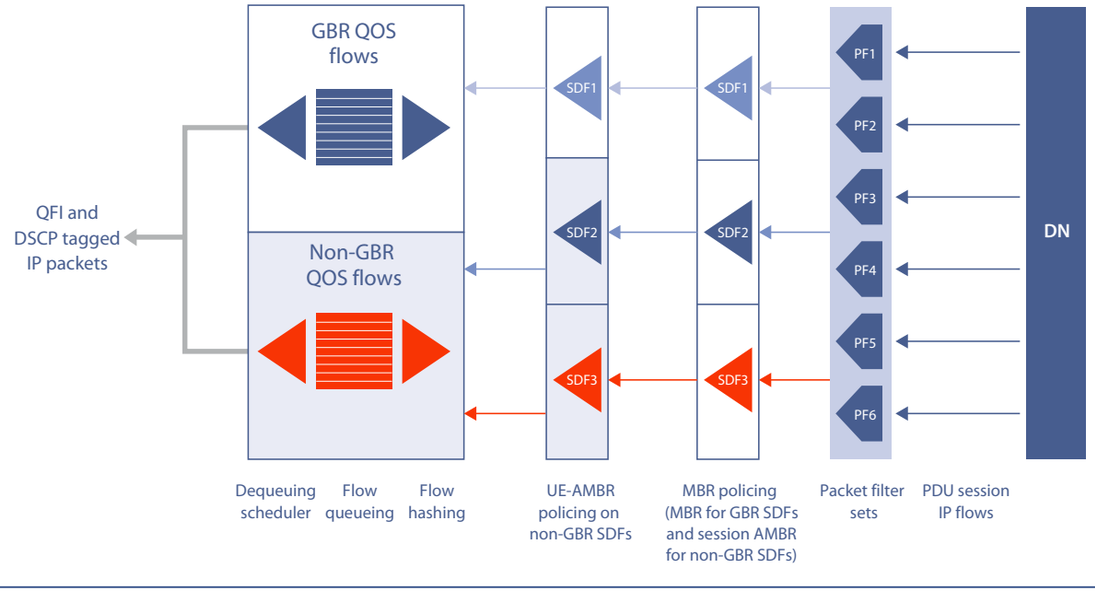


Figure 4: vAQM for the PDU session QoS flows at the UPF

where the PDU session has a local breakout, vAQM is configured at the edge UPF serving the local breakout; in other words, at the uplink classifier or branching point.

In downlink, as shown in *Figure 4*, the UPF binds the IP flows of the PDU session to SDFs/QoS flows based on the SDF templates using the IP packet filter set (for example, 5 tuples). The UPF then enforces the PDU session aggregate maximum bitrate (AMBR) across all non-GBR QoS flows of the PDU session. It also enforces the GBR for the GBR QoS flows. The UPF finally tags the packets with QoS Flow Identities (QFIs) and hands over the packets to vAQM, which is responsible for the QoS flow.

The different vAQM instances associated with QoS flows can contain implementations of completely different AQM algorithms or be configured with different default parameters.

Conclusion

The ongoing evolution of networks and applications is increasing the requirements on end-to-end performance. Specifically, it is important to be able to serve data at high rates without causing unnecessary delay due to bloated buffers. Our testing has shown it is possible to mitigate bufferbloat substantially through the use of a virtual form of AQM (vAQM) that is centralized upstream rather than being deployed in the bottleneck nodes, as it is in classic AQM. This approach greatly simplifies the deployment and configuration of AQM in mobile networks, and it ensures consistent behavior across bottleneck nodes, which in many cases are supplied by a multitude of vendors. In light of these benefits, vAQM will be an important user plane function in 5G core.

THE AUTHORS

Marcus Ihlar

◆ is a system developer in the field of traffic optimization and media delivery. He joined Ericsson in 2013 and initially did research on information-centric networking. Since 2014, he has been working on transport layer optimization and media delivery. Ihlar also participates actively in IETF standardization in the Transport Area Working Group. He holds a B.Sc. in computer science from Stockholm University in Sweden.

Ala Nazari

◆ is a media delivery architecture expert. He joined Ericsson in 1998 as a specialist in datacom, working with GPRS, 3G and IP transport. More recently, he has worked as a senior solution architect and engagement manager. Prior to joining Ericsson, Nazari spent several years at Televerket Radio working with mobile and fixed broadband and transport. He holds an M.Sc. in computer science from Uppsala University in Sweden

Robert Skog

◆ is a senior expert in the field of media delivery. After earning an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm in 1989, he joined Ericsson's two-year trainee program for system engineers. Since then, he has mainly been working in

the service layer and media delivery areas, on everything from the first WAP solutions to today's advanced user plane solutions. In 2005, Skog won Ericsson's prestigious Inventor of the Year Award.

References

1. **Controlled Delay Active Queue Management, RFC8289, January 2018, available at:** *https://tools.ietf.org/html/rfc8289*
2. **The Flow Queue CoDel Packet Scheduler and Active Queue Management Algorithm, RFC8290, January 2018, available at:** *https://www.ietf.org/mail-archive/web/ietf-announce/current/msg17315.html*
3. **CoDel Overview, available at:** *https://www. bufferbloat.net/projects/codel/wiki*
4. **Making WiFi Fast (blog), Dave Täht, available at:** *http://blog.cerowrt.org/post/make-wifi-fast/*
5. **The Addition of a Spin Bit to the QUIC Transport Protocol, IETF Datatracker (2017), available at:** *https://datatracker.ietf.org/doc/draft-trammell-quic-spin*

Further reading

- **Ending the Anomaly: Achieving Low Latency and Airtime Fairness in WiFi (conference paper), Toke Høiland-Jørgensen et al., USENIX ATC (2017), available at:** *https://www.usenix.org/system/files/conference/atc17/atc17-hoiland-jorgensen.pdf*
- **Ending the Anomaly: Achieving Low Latency and Airtime Fairness in WiFi (slide set), Toke Høiland-Jørgensen et al., USENIX ATC (2017), available at:** *https://datatracker.ietf.org/meeting/99/materials/slides-99-iccrg-iccrg-presentation-1/*

FIVE TECHNOLOGY TRENDS AUGMENTING THE CONNECTED SOCIETY

Rapid advancements in the use of machines to augment human intelligence are creating a new reality in which we increasingly interact with robots and intelligent agents in our daily lives, both privately and professionally. The list of examples is long, but a few of the most common applications today are found in education, health care, maintenance and gaming.

My vision of the future network is an intelligent platform that enables this new reality by supporting the digitalization of industries and society. This network platform consists of three main areas: 5G access, automation through agility, and a distributed cloud. A set of intelligent network applications and features is key to hiding complexity from the network's users, regardless of whether they are humans or machines.

The ability to transfer human skills in real time to other humans and machines located all around the world has the potential to enable massive efficiency gains. Autonomous operation by machines with self-learning capabilities offers the additional advantage of continuous performance and quality enhancements. High levels of cooperation and trust between humans and machines are essential. Building and maintaining trust will require decision transparency, high availability, data integrity and clear communication of intentions.

The network platform I envision will deliver truly intuitive interaction between humans and machines. In my view, there are five key technology trends that will play critical roles in achieving the vision:

#1 The realization of zero touch
#2 The emergence of the Internet of Skills
#3 Highly adaptable, cyber-physical systems
#4 Trust technologies for security assurance
#5 Ubiquitous, high-capacity radio

BY ERIK EKUDDEN, CTO

FEATURE ARTICLE

#1 THE REALIZATION OF ZERO-TOUCH

THE ZERO-TOUCH networks of the future will be characterized by the fact that they require no human intervention other than high-level declarative and implementation-independent intents. On the road to zero touch both humans and machines will learn from their interactions. This will build trust and enable the machines to adjust to human intention.

Compute and intelligence will exist in the device, in the cloud and in various places in the network. The network will automatically compute the imperative actions to fulfill given intents through a closed loop operation. Today's complex networks are designed for operation by humans and the complexity is expected to increase. As machine learning and artificial intelligence continue to develop, efficiently integrating learning and reasoning, the competence level of machine intelligence will grow.

AUGMENTATION OF HUMAN INTELLIGENCE

The realization of zero touch is an iterative process in which machines and humans collaborate reciprocally. Machines build intelligence through continuous learning and humans are assisted by machines in their decision-making processes. In this collaboration, the machines gather knowledge from humans and the environment in order to build models of the reality. Structured knowledge is created from unstructured data with the support of semantic web technologies, such as ontologies. The models are created and evolved with new knowledge to make informed predictions and enhance automated decision making.

To maximize human trust and improve decision quality, there is a need for transparency in the machine-driven decision-making process. It is possible to gain insights into a machine's decision process by analyzing its internal model and determining how that model supported particular decisions. This serves as a basis for generating explanations that humans can understand. Humans can also evaluate decisions and provide feedback to the machine to further improve the learning process. The interaction between humans and machines occurs using natural language processing as well as syntactical and semantic analysis.

ROBOTS AND AGENTS COLLABORATE WITH HUMANS

In a collaborative scenario, a robot will be able to anticipate human intentions and respond proactively. For example, an assembly-line robot would automatically adapt its pace to the skills of its human coworkers. Such interactions require the introduction of explainable artificial intelligence to cultivate human trust in robots. Robots will work alongside humans to aid and to learn. Robots can also interact with other digitalized components or digital twins to receive direct feedback. However, further advancements in robot design and manufacturing will be needed to improve their dexterity.

A software agent in a zero-touch network acts in the same way as a human operator. The agent should be able to learn the role in real time, as well as the pattern and the proper actions for a given task. In particular, it should be able to handle a wide range of random variations in the task, including contaminated data from the real world that originates from incidents and mistakes. These agents will learn through a combination of reinforcement learning (where the agent continually receives feedback from the environment) and supervised/unsupervised learning (such as classification, regression and clustering) from multiple data streams. An agent can be pre-trained in a safe environment, as within a digital twin, and transferred to a live system. Domain knowledge is a key success factor when applying agents to complex tasks.

Techniques such as neural networks offer significant advantages in learning patterns, but the current approach is too rigid. Differential plasticity is another technique that looks promising.

#2 THE EMERGENCE OF THE INTERNET OF SKILLS

THE INTERNET OF SKILLS allows humans to interact in real time over great distances – both with each other and with machines – and have similar sensory experiences to those that they experience locally. Current application examples include remote interactive teaching and remote repair services. A fully immersive Internet of Skills will become reality through a combination of machine interaction methods and extended communication capabilities. Internet of Skills-based systems are characterized by the interplay of various devices with sensing, processing and actuation capabilities near the user.

Current systems lack the audio, visual, haptic and telecommunication capabilities necessary to provide a fully realistic experience. To enable the Internet of Skills, the interplay between humans and robots, and between humans and virtual content, is of particular importance. Both industry and consumers are showing great interest and openness in using these new capabilities.

HUMAN SKILLS DELIVERED WITHOUT BOUNDARIES

An authentic visual experience requires real-time 3D video capturing, processing and rendering. These capabilities make it possible to create a 3D representation of the captured world and provide the experience of being immersed in a remote or virtual environment. While today's user devices don't yet provide the necessary resolution, field of view, depth perception, wearability and positioning capabilities, the quality and performance of these technology components is steadily improving.

Spatial microphones will be used to separate individual sound sources in the space domain. This implies that there will be an increased amount of data needed to capture the audio spatial aspects. Spatial audio rendering performance is very much tied to efficient head-related filter models. New formats for exchanging spatial audio streams have been specified and compression techniques are being developed.

Haptic components allow users to feel shapes, textures, motion and forces in 3D. Devices will also track the motions and forces applied by the user during interaction. With current technologies the user needs to wear or hold a physical device, but future ultrasound based haptic devices will offer a contact-free solution. Standardization efforts for haptic communication will allow for a quicker adoption of haptic capabilities.

INSTANT INTERACTION AND COMMUNICATION

Communication between humans and machines will become more natural, to the point that it is comparable to interpersonal interaction. Natural user interfaces such as voice and gesture will be commonplace. The use of vision-based sensors will allow for an intuitive type of interaction. To better understand human-machine interaction there is a need to evolve the understanding of kinesiology, ergonomics, cognitive science and sociology, and to incorporate them into algorithms and industrial design. This would make it easier to convey a machine's intent before it initiates actions, for example.

Large volumes of 3D visual information impose high network capacity demands, making ultra-low latency and high bandwidth communication technologies essential. Enabling the best user experience requires the use of network edge computers to process the large volumes of 3D visual, audio and haptic information. This setup saves device battery lifetime and reduces heat dissipation, as well as reducing network load.

#3 HIGHLY ADAPTABLE CYBER-PHYSICAL SYSTEMS

A CYBER-PHYSICAL system is a composite of several systems of varying natures that will soon be present in all industry sectors. It is a self-organizing expert system created by the combination of model of models, dynamic interaction between models and deterministic communication. A cyber-physical system presents a concise and comprehensible system overview that humans can understand and act upon.

The main purpose of a cyber-physical system is to control a physical process and use feedback to adapt to new conditions in real time. It builds upon the integration of computing, networking and physical processes. An example of a cyber-physical system is a smart factory where mechanical systems, robots, raw materials, and products communicate and interact. This interaction enables machine intelligence to perform monitoring and control of operations at all plant levels.

SYNERGISTIC INTEGRATION OF COMPUTATION, NETWORKING AND PHYSICAL PROCESSES

The main challenge is the orchestration of the networked computational resources for many interworking physical systems with different levels of complexity. Cyber-physical systems are transforming the way people interact with engineered systems, just as the internet has transformed the way people interact with information. Humans will assume responsibility on a wider operating scale, supervising the operation of the mostly automated and self-organizing process.

A cyber-physical system contains different heterogeneous elements such as mechanical, electrical, electromechanical, control software, communication network and human-machine interfaces. It is a challenge to understand the interaction of the physical, cyber and human worlds. System models will define the evolution of each system state in time. An overarching model will be needed to integrate all the respective system models while contemplating all possible dynamic interactions. This implies a control program that delivers a deterministic behavior to each subsystem. Current design tools need to be upgraded to consider the interactions between the various systems, their interfaces and abstractions.

MODEL OF MODELS CREATES THE CYBER-PHYSICAL SYSTEM

Within the cyber-physical system all system dynamics need to be considered through a model that interacts with all the sub-models. Many factors impact the dynamics of the interactions between the systems, including latency, bandwidth and reliability. For a wireless network, factors such as the device location, the propagation conditions and the traffic load change over time. This means that networks need to be modeled in order to be integrated in the model of models.

The time it takes to perform a task may be critical to enable a correctly functioning system. Physical processes are compositions of many things occurring in parallel. A model of time that is consistent with the realities of time measurement and time synchronization needs to be standardized across all models.

EXAMPLE: INDUSTRY 4.0

The factory of the future implements the concept of Industry 4.0, which includes the transformation from mass production to mass customization. This vision will be realized through large-scale industrial automation together with the digitalization of manufacturing processes.

Humans assume the role of supervising the operation of the automated and self-organizing production process. In this context it will be possible to recognize all the system models that need to interact:

- Physical and robotic systems such as conveyors, robotic arms and automated guided vehicles
- Control systems such as robot controllers and programmable logic controllers for production
- Software systems to manage all the operations
- Big data and analytics-based software systems
- Electrical systems to power machines and robots
- Communication networks
- Sensors and devices.

The master model consists of and interacts with all the listed processes above, resulting in the realization of the final product.

#4 TRUST TECHNOLOGIES FOR SECURITY ASSURANCE

TRUST TECHNOLOGIES will provide mechanisms to protect networks and offer security assurance to both humans and machines. Artificial intelligence and machine learning are needed to manage the complexity and variety of security threats and the dynamics of networks. Rapidly emerging confidential computing – together with possible future multi-party computation – will facilitate secure cloud processing of private and confidential data. Performance and security demands are driving the development of algorithms and protocols for identities.

The use of cloud technologies continues to grow. Billions of new devices with different capabilities and characteristics will all be connected to the cloud. Many of them are physically accessible and thus exposed and vulnerable to attack or to being misused as instruments of attack. Digital identities are needed to prove ownership of data and to ensure that services only connect to other trustworthy services. Flexible and dynamic auditing and compliance verification are required to handle new threats. Furthermore, there is a need for automated protection that adapts to operating modes and performs analytics on the system in operation.

PROTECTION DRIVEN BY ARTIFICIAL INTELLIGENCE

Artificial intelligence, machine learning and automation are becoming important tools for security. Machine learning addresses areas such as threat detection and user behavior analytics. Artificial intelligence assists security analysts by collecting and sifting through threat information to find relevant information and computing responses. However, there is a need to address the current lack of open benchmarks to determine the maturity of the technology and permit comparison of products.

While the current trend is to centralize data and computation, security applications for the Internet of Things and future networks will require more distributed and hierarchical approaches to support both fast local decisions and slower global decisions that influence local policies.

CONFIDENTIAL COMPUTING TO BUILD TRUST

Confidential computing uses the features of enclaves – trusted execution environments and root of trust technologies. Code and data is kept confidential and integrity protection is enforced by hardware mechanisms, which enable strong guarantees that data and processing are kept confidential in the cloud environment and prevent unauthorized exposure of data when doing analytics. Confidential computing is becoming commercial in cloud systems. Research is underway to overcome the remaining challenges, including improving the efficiency of the trusted computing base, reducing context switch overheads when porting applications and preventing side channel information leakage.

Multi-party computation enables parties to jointly compute functions over their combined data inputs while keeping those inputs private. In addition to protecting the confidentiality of the input data, multi-party protocols must guarantee that malicious parties are not able to affect the output of honest parties. Although multi-party computation is already used in special cases, its limited functionality and high computation complexity currently stand in the way of wide adoption. Time will tell if it becomes as promising as confidential computing.

PRIVACY REQUIRES SECURE IDENTITIES

Digital identities are crucial to maintaining ownership of data and for authenticating and authorizing users. Solutions that address identities and credentials for machines are equally important. The widespread use of web and cloud technologies has made the need for efficient identity solutions even more urgent. In addition, better algorithms and new protocols for the transport layer security provide improved security, lower latency and reduced overhead. Efficiency is particularly important when orchestrating and using identities for many dynamic cloud systems, such as those realized via microservices, for example.

When quantum computers with enough computational power are available, all existing identity systems that use public-key cryptography will lose their security. Developing new secure algorithms for this post-quantum cryptography era is an active research area.

#5 UBIQUITOUS, HIGH-CAPACITY RADIO

THE WIRELESS access network is becoming a general connectivity platform that enables the sharing of data anywhere and anytime for anyone and anything. There is good reason to believe that rapidly increasing data volumes will continue in the foreseeable future. Ultra-reliable connectivity requires ultra-low latency, which will be needed to support demanding use cases. The focus will be on enabling high data rates for everyone, rather than support for extremely high data rates that are only achievable under specific conditions or for specific users.

A few technologies will need to be enhanced in order to create a ubiquitous, high-capacity radio network. The common denominator for these technologies is their capability to enable and utilize high frequencies and wide bandwidth operations. Coverage is addressed through beamforming and flexibility in device interworking. The challenge is to support data volumes and demanding-traffic use cases, without a corresponding increase in cost and energy consumption.

DEVICES ACT AS NETWORK NODES

To enhance device coverage, performance and reliability, simultaneous multi-site connectivity across different access technologies and access nodes is required. Wireless technology will be used for the connectivity between the network nodes, as a complement to fiber-based networks. Device cooperation will be used to create virtual large antenna arrays on the device side by combining the antennas of multiple devices. The borderline between devices and network nodes will be more diffuse.

Massive heterogenous networks will have a much more mesh-like connectivity. Advanced machine learning and artificial intelligence will be key to the management of this network, enabling it to evolve and adapt to new requirements and changes in the operating environment.

NO SURPRISE – EXPONENTIAL INCREASED DATA RATES

Meeting future bit rate demands will require the use of frequency bands above 100 GHz. Operation in such spectrum will enable terabit data rates, although only for short-range connectivity. It will be an implementation challenge to generate substantial output power and handle heat dissipation, considering the small dimensions of THz components and antennas. Spectrum sharing will be further enabled by beamforming, which is made possible by the high frequency.

Integrated positioning will be enabled by high-frequency and wide-bandwidth operation in combination with very dense deployments of network nodes. High-accuracy positioning is important for enhanced network performance and is an enabler for new types of end-user services. The positioning of mobile devices, both indoor and outdoor, will be an integrated part of the wireless access networks. Accuracy will be well below one meter.

A NEW TRADE-OFF BETWEEN ANALOG AND DIGITAL RADIO FREQUENCY HARDWARE

For the past 20 years there has been a continuous trend toward moving functionality from the analog to the digital radio frequency domain. However, the trend is reversed for very wide band transmission at very high frequencies, in combination with a very large number of antennas. This means that a new implementation balance and interplay between the analog and digital radio frequency domains will emerge. Increasingly sophisticated processing is already moving over to the analog domain. This will soon also include utilizing correlations between different analog signals received on different antennas, for example. The compression requirements on the analog-to-digital conversion is reduced. The split between analog and digital radio frequency hardware implementation will change over time as technology and requirements evolve. ✪

Digital connectivity marketplaces

TO ENRICH 5G AND IoT VALUE PROPOSITIONS

One of the key growth opportunities for the telecom industry is to provide network capabilities that support the digital transformation underway in most businesses and industries. Already today, we have a powerful technology foundation in place, and this will become even stronger with 5G. Now is the ideal time to evolve the business side of the equation toward platform business models, which will enable the telecom industry to prosper in multisided business ecosystems as well.

MALGORZATA SVENSSON, LARS ANGELIN, CHRISTIAN OLROG, PATRIK REGÅRDH, BO RIBBING

Digitalization is undoubtedly one of the most significant forces for change of our time, characterized by a wide variety of smart devices and applications that touch all possible areas of life. Now, new technologies such as the Internet of Things (IoT), artificial intelligence, wearables and robotics are driving an accelerating digital transformation in which core processes across virtually all businesses and industries are becoming more distributed, connected and real-time optimized.

Businesses need networks that can provide them with the foundation to operate the new transforming business models. Digital transformation requires network capabilities to support a broad spectrum of different user scenarios. Coverage, speed, latency and security are some of the key parameters, along with new network functions for effective processing of data and distribution of applications and functions far out in the network. 5G brings the ability to realize a wide variety of connectivity and network services to meet the performance requirements of tomorrow's digital industries.

With networks capable of supporting an unlimited set of services and use cases, it has become increasingly clear that traditional mass-market services will soon be a thing of the past. Looking forward, networks must be viewed as a horizontal foundation or platform on which businesses can consume networks just as other ICT capabilities are consumed. The business model of the future will look more like current aaS models for other parts of the ICT sector. The current telecom business model must evolve significantly to work in this new market reality to place telecom companies in a better position to capitalize on new business opportunities.

The case for business model transformation

For many years, the telecom industry has successfully been operating in what is best described as a vertical business model. In this model, telecom operators have delivered end-to-end (E2E) standardized services to consumers under long-term contracts and competed within national boundaries. The collaboration between communication service providers has generally been limited to interoperability and roaming for the purpose of global service reach and to drive technology and market scale. It is reasonable to expect the traditional modus of collaboration to continue to serve the industry in the years ahead, but it is also clear that this framework is not fit to explore the full value of network infrastructure. Rich configuration variances and more complex services will in many cases push demand beyond the capability of a single

Definitions

- **Consumer:** a user associated with a company that purchases services/products exposed by the platform/marketplace. A consumer can be an individual and/or an organization.
- **Producer:** a supplier of services/products to a platform/marketplace. A producer can be an individual and/or an organization.
- **Platform provider:** an organization that manages the platform/marketplace.
- **Platform/marketplace:** the entity that exposes the services of the platform. The exposed services are a composition of platform services and services supplied by producers.
- **Digital connectivity marketplace:** the marketplace that offers various connectivity services and connectivity service enablers to enterprises. The enterprises may require IoT capabilities.
- **Cloud services:** a set of services including storage and computation.
- **Digitalization:** the use of digital technologies to change a business model and provide new revenue and value-producing opportunities. It is the process of moving to a digital business.

Terms and abbreviations

aaS – as a Service | **B2B** – business-to-business | **B2C** – business-to-consumer | **BSS** – business support systems | **DCP** – Device Connection Platform | **E2E** – end-to-end | **IoT** – Internet of Things | **OSS** – operations support systems | **SLA** – Service Level Agreement

provider, as well as beyond the scope of what is possible to realize with traditional interoperability and federation models.

In light of this, it is necessary to evolve toward a new collaboration framework – one that offers service providers an environment where they can dynamically aggregate capabilities from multiple sources into the tailored solution they require to enable their unique digitalization journey.

Such a model must excel in serving the unique and varying needs of consumers as effectively as possible and include value contribution from innovation partners. It must also meet a new set of requirements to effectively create and capture value. Some of the most critical aspects are:

- simplicity for service providers to define and configure their own connectivity and network solution regardless of underlying infrastructure and player boundaries.
- openness for service providers to easily integrate and dynamically adjust their solution, including the needed management, into their digital business processes.
- well-structured exposure of enhanced connectivity and network capabilities such as distributed cloud, edge analytics, and so on, for rapid use-case innovation and value creation.
- easy exchange of digital assets such as safe and secure trade of data across different players.
- attractive complement for operators to explore currently unaddressable business/markets beyond the vertical business they traditionally run today.
- creation of an environment where innovation thrives to the benefit of all business model participants.

To effectively meet these and other requirements, it is obvious that the establishment of such a business model is beyond the ability of a single player or company. Instead, it makes more sense to look at a business that is designed like a platform where a multitude of players, producers and consumers can coexist and gain mutual business outcomes. In this scenario, the platform is both a marketplace that supports the business of its stakeholders in the most effective way and also an infrastructure that facilitates and simplifies these multisided relationships with a range of key functionalities.

> MULTISIDED PLATFORMS BRING TOGETHER TWO OR MORE DISTINCT BUT INTERDEPENDENT GROUPS OF CONSUMERS AND SUPPLIERS

Multisided connectivity platforms

Multisided platforms bring together two or more distinct but interdependent groups of consumers and suppliers. Such platforms are of value to one group (consumers, for example) only if the other group (suppliers, in this case) is also present. A platform creates value by facilitating interactions and matchmaking between different groups. A multisided platform grows in value to the extent that it attracts more users and it matchmakes one consumer's needs with services and products offered by multiple suppliers; a phenomenon known as the network effect.

The telecom industry already offers a vast variety of connectivity services, and 5G technologies will accelerate the development of new ones [1] characterized by the different capabilities that the various industries require. The connectivity services of 5G will be realized by means of network slicing. We believe that tailored connectivity services with

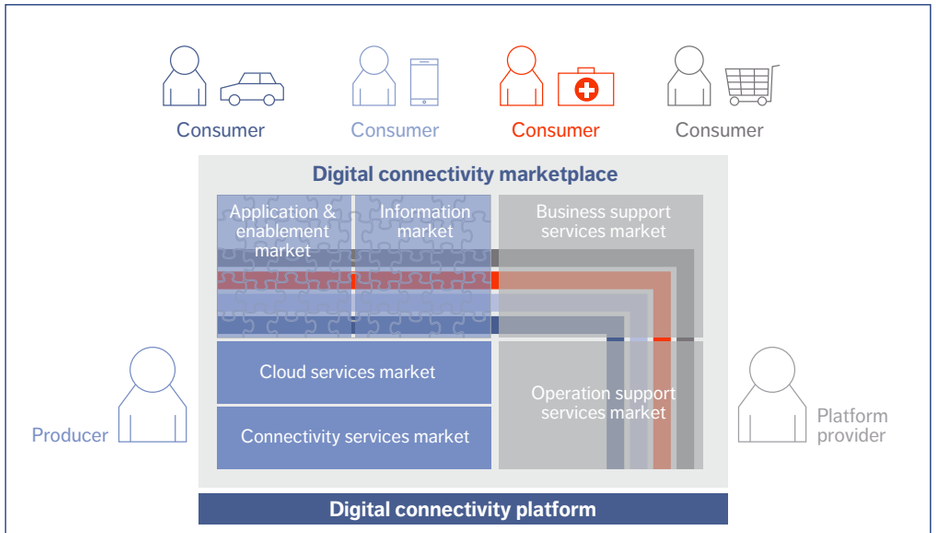


Figure 1: Digital connectivity marketplace: multisided platform

diverse characteristics will be necessary for enterprises within the industries to digitalize their operational and business processes [2].

Digitalization of enterprises' business and operation requires not only connectivity services but also various cloud-based services like computation and storage, where companies' operations and business applications can be hosted, deployed and operated, or where enterprises can benefit from the cloud services' capabilities offered in the software as service business model.

The full potential of 5G connectivity and cloud services can be reached when they are exposed on fully digital multisided connectivity platforms (marketplaces), as shown in *Figure 1*. In this way, the services – and consequently the industries – will benefit from the presence of a vast range of other services such as information and application enablement services, as well as business and operation support capabilities. The connectivity marketplace will complement the federation model for connectivity and bridge the gap where the existing federation model is not enough.

The digital connectivity marketplace will enable new business models by providing capabilities for the exchange of information and connectivity enabling services to ecosystems of consumers and suppliers.

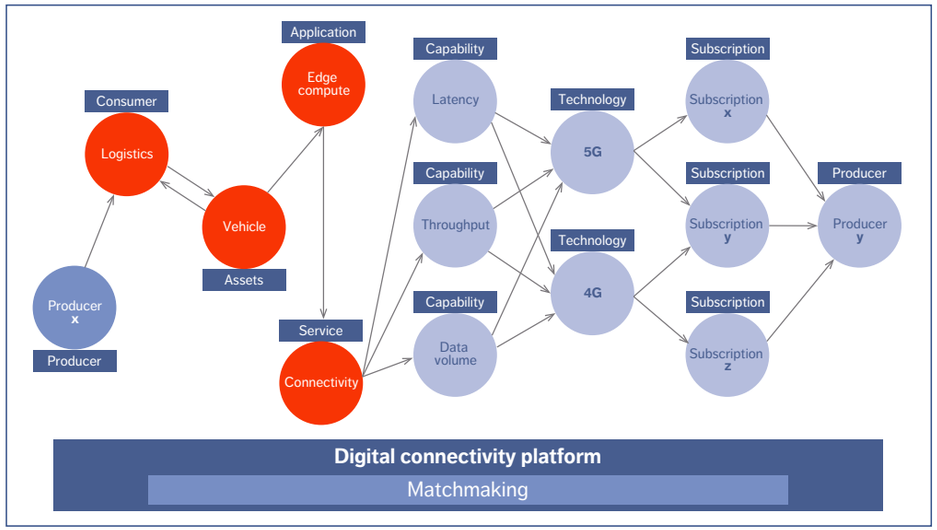


Figure 2: Graph model: how industry requirements can be linked with connectivity services

Our vision of the future digital connectivity marketplace is that it will offer interaction between producers and consumers by exposing the connectivity and cloud services on one side, and collecting the requirements from consumers and matchmaking them against the capabilities of the services on the other. The platform will facilitate the interactions by providing business and operation support services to both producers and consumers.

The services exposed in the marketplace could be supplied by multiple producers owning the networks and infrastructures required to produce these services. The consumers representing various industries will require connectivity and cloud services to digitalize their business and operational processes [2].

Consumer services offered by producers

As discussed above, the future multisided connectivity marketplace will offer capabilities to expose connectivity and cloud services that may be provided by multiple producers. In the context of 5G the following types of connectivity services may be exposed: massive machine type communication (also known as massive IoT), mission-critical machine type communication (also known as mission-critical IoT) and extreme mobile broadband services. There will be a need to complement the connectivity services provided by others to be able to offer the desired value proposition and satisfy industry demands.

Security will remain one of the top priorities as societies continue to connect everything from autonomous trucks to elevators and ventilation systems. Isolation techniques such as automatic network slicing are likely to play a key role in reducing security risks, together with machine-learning-based ingress firewalls, possibly embedded as virtual network functions in operators' networks.

Making the platform attractive and efficient to operate

At Ericsson, we foresee the digital marketplace as the foundation of an ecosystem of many consumers and providers from a multitude of industries. This implies that industries will require capabilities to express their requirements in languages and contexts that are specific to their business and operational environments, as opposed to needing telco and cloud-domain knowledge to be able to choose or define the requirements on the connectivity and cloud services.

The multisided character of a platform – and its matchmaking abilities – will make it well suited to provide these capabilities. The platform will match industry-specific requirements with service capabilities.

The matchmaking function uses the relatively old research area of graph theory, boosted by its explosion in usage and implementation in social networks in recent years. The model in *Figure 2* shows an example of how consumer requirements, modeled as graph edges, are linked with services, then technologies and consequently subscriptions. By defining each service as a vertex in a graph where edges constitute a consumer/producer relationship, it is possible to model complex value chains or relationships between services.

> SECURITY WILL REMAIN ONE OF THE TOP PRIORITIES AS SOCIETIES CONTINUE TO CONNECT

Given a set of requirements modeled as vertexes with consumption needs as "dangling edges," we can find an efficient subset of the graph which fulfills these needs, if at all possible. By also modeling non-functional aspects such as location, latency and throughput, we get a generically applicable mechanism. This gives rise to tremendous possibilities where a multitude of loosely connected players together enable an efficient E2E service, as a dynamically constructed value chain, capable of reaching far beyond what is currently possible. There are many implementation options, including capabilities for life-cycle management to ensure that consumers and producers, joined by market forces and the platform, provide a suitable mix of stability and innovation.

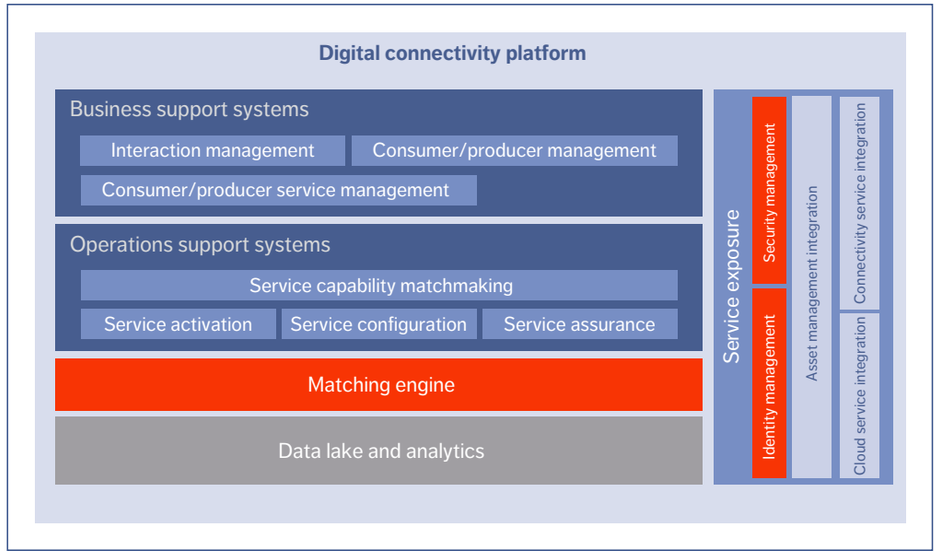


Figure 3: Digital connectivity platform

Data lakes and analytics are used to execute matchmaking, as shown in *Figure 3*. Analytics services will make the future platform data driven, as they will be exposed for use by either consumers or producers to help them evolve their services. At the same time, the consumers or producers provide the platform with extended insight into how producers' services are used, potentially helping improve the recommendations supporting service discovery and matchmaking.

A common solution for providing identification is yet another important component of the platform to truly support scale and innovative combinations. Federation of identities is a key service when considering new service combinations. Authorization for a connection between two things will likely take both software fingerprinting and behavioral analysis into account if this functionality is available in a "thing management" system in a trusted relationship.

Security solutions will enable the safe and secure trade of data between different players and the exchange of digital assets [3] . Trust facilities will enable secure transactions between consumers and producers. Analytics insight in the shape of fraud management, combined with a novel form of analytics insight predicting likelihood of successful service activation, will be the key for the platform provider to be able to manage risk when bridging the trust divide.

Service exposure will secure openness for consumers to easily integrate and dynamically adjust their solutions and services, including the needed management, into their digital business processes.

Business support platform

The business support platform comprises many things for both producers and consumers: their entire life cycle, interactions and management while in a business relationship with the platform; the entire service life cycle of the onboarded producer and consumer services sets; and self-services for the aforementioned.

One of the most critical capabilities is contracting, due to the fact that the mix of services, SLAs and pricing schemes is complex and contracts are frequently negotiated and renegotiated. The business support platform simplifies business by using only bilateral contracts between parties – that is, a business-to-business (B2B) contract between any two of the three: the producer, the consumer and the marketplace provider. B2C contracts are applicable if the consumer is a private individual and not an enterprise. While business value chains are often referred to as B2B2x, the contracting arrangement supporting them is a set of linked B2B and B2C contracts.

In addition, the marketplace allows for a party to act as a consumer to purchase services, orchestrate them into a new service, and offer the orchestrated service on the platform as a producer. A party in a producer role can also act as a consumer for its resource needs. This results in a series of B2C and B2B contracts, which is the essence of a vibrant ecosystem built on the platform. The contracting and contract management processes are fully automated.

Operations support platform

The consumer and producer services that enable the network capabilities such as distributed cloud and edge analytics will be exposed using service exposure, thereby enabling innovation and value creation.

The dependencies between services are automatically inferred through service-level assurance, and the combined analytics of the different services provides insight into performance at several levels. This may allow fine-tuning of SLAs based on heuristics and help fuel innovation by managing and providing transparency with regard to risk. A network or service outage must be dealt with swiftly, as they may be used in consuming enterprises' mission-critical production. Automatic healing, based on orchestration and policies, is one way to to mitigate outage situations.

The service configuration and activation will make it possible for enterprises to define and configure their connectivity services settings simply and independently of underlying infrastructure and player boundaries.

Device Connection Platform

One of the first realizations of a multisided platform is Ericsson's Device Connection Platform (DCP). This is a platform where mobile network connectivity services are exposed to be consumed by applications. Its fundamental purpose is to make connectivity available for enterprises to include in their offerings and provide the means to manage this connectivity. Two examples of typical enterprises in this context would be an original equipment manufacturer of some sort (such as an automotive company or a camera manufacturer) or a service provider of a service where a device is an integral part of the offering (such as a utility company using automatic meter reading or a point-of-sales terminal provider).

As mobile networks are always of a geographically-limited nature, normally by country or license area, there is a need to harmonize the way the enterprise's connectivity is used across several networks. Ideally, the enterprise needs to have the same connectivity experience wherever it launches and runs its service. The only practical way to achieve this is to standardize on a centralized platform, which brings together the connectivity of all providers and exposes it in a uniform way. This makes the experience truly harmonized, in terms of functionality as well as service levels and operational procedures. Furthermore, it also minimizes the cost of integration between the enterprise systems and the access networks, saving money both initially and continuously, as well as shortening time to market for the enterprise's service.

Conclusion

The ongoing digitalization of businesses and industries is one of the key growth opportunities

for the telecom industry. Telecom infrastructure, encompassing both current networks and soon-to-be-launched 5G, is rapidly evolving into a very powerful resource that can bring significant value to the digital transformation of most industries. Looking ahead, some of the telecom industry's key value levers are: quality differentiated connectivity; distributed computing and storage; analytics of data flows; and security solutions. With the technology already in place, the main challenge is to fully understand the needs of a new breed of consumers, and use that information to organize the business, develop new relationships and establish efficient operations.

At Ericsson, we believe that the best way to overcome this challenge is to establish a platform model where capabilities from many providers can be effectively packaged and exposed in attractive ways to buyers from different industries. This approach creates a new and much-needed role for the telecom industry as the provider of the marketplace and brings opportunities for multisided business relations and transactions to prosper.

On one side of the platform are the businesses and industries that are consuming services from the underlying infrastructure, and on the other side are providers of both network assets and enabling functions. The platform organizes the relationships for optimal fit and usefulness for the players involved. In so doing, the platform provides a number of useful functions with one of the key values being the scale of business and the networking effect that it has to offer. Instead of competing for the same consumers with the same offer, operators that restructure according to a logic that enables full participation in an ecosystems platform will find themselves well positioned to fully capitalize on new market opportunities.

References

1. **Ericsson white paper, January 2017, 5G – enabling the transformation of industry and society, available at:** *https://www.ericsson.com/en/white-papers/5g-systems--enabling-the-transformation-of-industry-and-society*
2. **Ericsson white paper, October 2017, Telecom IT for the digital economy, available at:** *https://www.ericsson.com/en/publications/white-papers/telecom-it-for-the-digital-economy*
3. **Ericsson Technology Review, November 2017, End-to-end security management for the IoT, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2017/end-to-end-security-management-for-the-iot*

Further reading

- **Ericsson white paper, November 2015, Operator service exposure – enabling differentiation and innovation, available at:** *https://www.ericsson.com/en/white-papers/operator-service-exposure--enabling-differentiation-and-innovation*
- **Ericsson Technology Review, April 2017, Tackling IoT complexity with machine intelligence, available at:** *https://www.ericsson.com/en/publications/ericsson-technology-review/archive/2017/tackling-iot-complexity-with-machine-intelligence*
- **Ericsson press release, December 2017, 2018 hot consumer trends: technology turns human, available at:** *https://www.ericsson.com/en/press-releases/2017/12/2018-hot-consumer-trends-technology-turns-human*

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. For the past 10 years, her work has focused on architecture evolution. Svensson has broad experience in

business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an M.Sc. in technology from the Silesian University of Technology in Gliwice, Poland.

Patrik Regårdh
◆ is head of strategy for Solution Area OSS, where his work focuses on market development, industry dynamics and driving strategies and initiatives for Ericsson's digital business. He joined the company in 1994 and his previous positions have ranged from strategy and

business development to account management. Currently based at the global headquarters in Stockholm, he has also worked extensively in Brazil, Thailand and Germany. Regårdh holds an M.Sc. from KTH Royal Institute of Technology in Stockholm, Sweden.

Christian Olrog
◆ is an expert in cloud service delivery architecture. He joined Ericsson in 1999 and has been a technical leader and architect in many different areas, ranging from embedded client

development and wireless enterprise connectivity with a security focus to cloud and operations support systems/business support systems (OSS/BSS). Olrog currently sits in the Technology & Industry group at Business Area Digital Services at Ericsson. He holds an M.Sc. in general physics from KTH Royal Institute of Technology.

Lars Angelin
◆ is an expert in BSS at Business Area Digital Services. He has more than 30 years of experience in the areas of concept development, architecture and strategies within the telco and education industries. He joined Ericsson in 1996 as a research engineer, and in 2003 he moved to a position as concept developer in the machine-to-machine and OSS/BSS areas. Since 2006, Angelin has been focusing on BSS – specifically business support, enterprise architectures and the software architectures to implement the systems. He holds an M.Sc. in engineering physics and a Tech. Licentiate in tele-traffic theory from the Faculty of Engineering at Lund University in Sweden, and an honorary Ph.D. from Blekinge Institute of Technology, Sweden.

Bo Ribbing
◆ is head of Product Management for Connectivity Management. He joined Ericsson in 1991 and has worked in the networks business since 1995. During this time, he has gained broad international experience, particularly

from Latin America and Asia Pacific, and has held several management positions. Ribbing holds an M.Sc. in applied physics from Linköping University in Sweden.

The authors would like to thank Frans de Rooij for his contribution to this article.

Cognitive technologies

IN NETWORK AND BUSINESS AUTOMATION

Forward-looking network operators and digital service providers require an automated network and business environment that can support them in the transition to a new market reality characterized by 5G, the Internet of Things, virtual network functions and software-defined networks. The combination of machine learning and machine reasoning techniques makes it possible to build cognitive applications with the ability to utilize insights across domain borders and dynamically adapt to changing goals and contexts.

JÖRG NIEMÖLLER, LEONID MOKRUSHIN

The need to support emerging technologies will soon lead to radical changes in the operations of both network operators and digital service providers, as their businesses tend to be based on a complex system of interdependent, manually-executed processes. These processes span across technical functions such as network operation and product development, support functions such as customer care, and business-level functions such as marketing, product strategy planning and billing. Manually-executed processes represent a major challenge because they do not scale sufficiently at a competitive cost.

Automation is an essential part of the solution. At Ericsson, we envision a new infrastructure for network operators and digital service providers in which intelligent agents operate autonomously with minimal human involvement, collaborating to reach their overall goals. These agents base their decisions on evidence in data and the knowledge of domain experts, and they are able to utilize knowledge from various domains and dynamically adapt to changed contexts.

Cognitive technologies

Software that is able to operate autonomously and make smart decisions in a complex environment is referred to as an intelligent agent (a practical implementation of artificial intelligence and machine learning). It perceives its environment and takes actions to maximize its success in achieving its goals. The term cognitive technologies refers to a diverse set of techniques, tools and platforms that enable the implementation of intelligent agents.

The model of mind shown in *Figure 1* illustrates the main tasks of an intelligent agent, and thus the main concerns of cognitive technologies. The model describes the process of deriving an action or decision from input and knowledge.

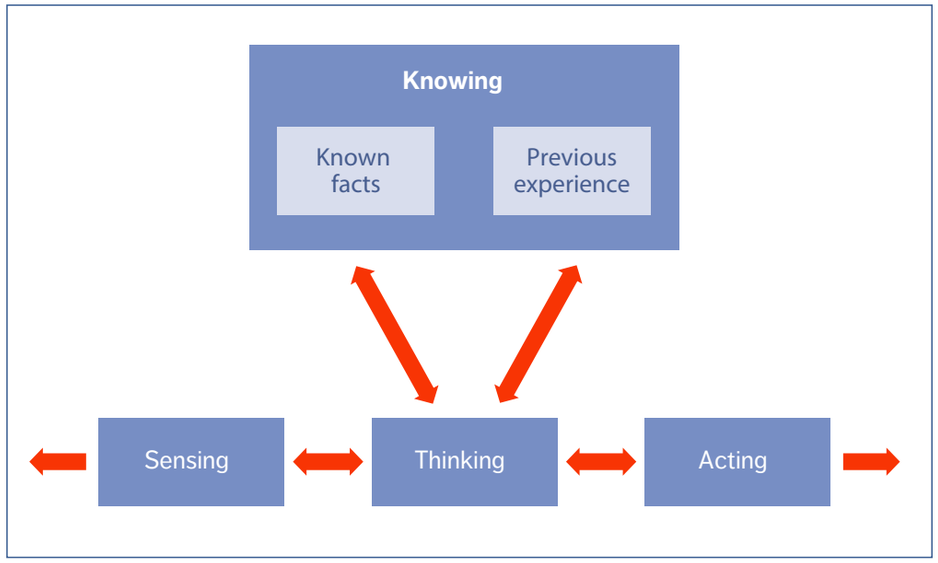


Figure 1: The model of mind

An intelligent agent needs a model of the environment in which it operates. Technologies used to capture information about the environment are diverse and use-case dependent. For example, natural language processing enables interaction with human users; network probes and sensors deliver measured technical facts; and an analytics system processes data to provide relevant insights.

The purpose of intelligent agents is to perform

Terms and abbreviations

CPI – Customer Product Information | **eTOM** – Enhanced Telecom Operations Map | **SID** – Shared Information/Data | **SLI** – Service Level Index | **TOVE** – Toronto Virtual Enterprise

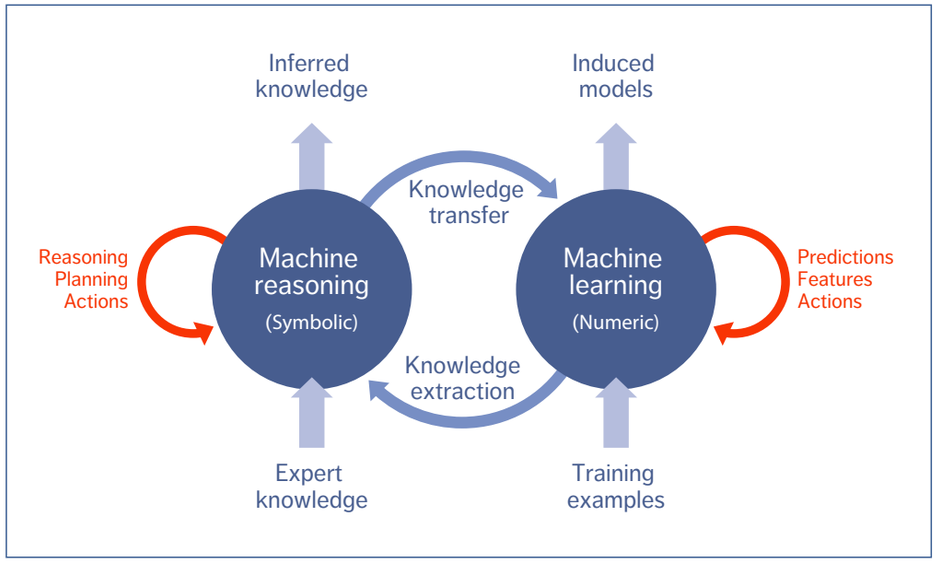


Figure 2: Machine reasoning and machine learning [1]

actions and communicate solutions. Acting complements sensing in interaction with the environment. The choice of techniques and tools is equally diverse and use-case dependent. For example, speech synthesis enables convenient communication with users, robotics involves mechanical actuation, and an intelligent network manager can act by executing commands on the equipment or changing configuration parameters.

The thinking phase in the model of mind is the source of the intelligence in an intelligent agent. Thinking can be implemented, for example, as a logic program in Prolog, in an artificial neural network, or in any other type of inference engine, including machine-learned models.

The thinking phase derives its decisions from facts and previous experiences stored in a knowledge base. The key is a machine-readable knowledge representation in the form of a model. Graph databases and triple stores are frequently used for efficient storage. Formal knowledge definition can be achieved using concepts of RDF (the Resource Description Framework) or description languages, such as UML (the Universal Markup Language) or OWL (the Web Ontology Language).

Machine learning and machine reasoning

There are two technological pillars on which an intelligent agent can be based: machine learning and machine reasoning (illustrated in *Figure 2*). Both involve making predictions and planning actions toward a goal. Each has its own strengths and weaknesses.

Machine learning relies on statistical methods to numerically calculate an optimized model based on the training data provided. This is driven by wanted characteristics of the model, such as low average error or the rate of false positive or negative predictions. Applying the learned numerical model to new data leads to predictions or action recommendations that are statistically closest to the training examples.

An example of a learned model is the Service Level Index (SLI) [2] [3] implemented in Ericsson Expert Analytics, which predicts a user's level of satisfaction. The training input is measurements from network probes that show the QoS delivered to the user combined with surveys in which users state their level of satisfaction. The learned model predicts this satisfaction level from new QoS measurements.

Machine reasoning generates conclusions from symbolic knowledge representation. Widely used techniques are logical induction and deduction. It relies on a formal description of concepts in a model, often organized as an ontology. Knowledge about the environment is asserted within the model by connecting abstract concepts and terminology to objects representing the entities to be used and managed. For example, "customer satisfaction," "user" and "quantifies" are abstract concepts. Based on these, we can assert that "Adam" is a user and "4" is the SLI value representing his satisfaction. We can further assert inference rules: "SLI quantifies satisfaction," "SLI below 5 is low," "low satisfaction causes churn". Based on this knowledge, a machine-reasoning process would logically conclude that Adam is about to churn. It would trace the reason to the low SLI value.

Hybrid approaches to symbolic neural networks also exist. These are deep neural networks with a numeric and statistics-based core and an implicit mapping of the model's numeric variables to a symbolic representation.

Designing intelligent agents

Autonomous intelligent agents support human domain experts by fully taking over the execution of operational tasks. Doing this convincingly requires them to react and execute faster than humans and be able to overcome unexpected situations, while making fewer errors and scaling to a high number of managed assets and tasks.

Intelligent agents are developed and deployed in a software life cycle. As such, they profit from the encapsulation provided by a microservice architecture, comprehensive and performant data routing and management, and a dynamically scalable execution environment. The ability to create an optimal thinking core for an intelligent agent requires a good understanding of the fundamental characteristics of machine learning and machine reasoning.

> AUTONOMOUS INTELLIGENT AGENTS SUPPORT HUMAN DOMAIN EXPERTS BY FULLY TAKING OVER THE EXECUTION OF OPERATIONAL TASKS

The role of abstraction

A person uses abstraction to distill essential information from the input presented. Abstraction provides focus and easier-to-grasp concepts as a base for reasoning and decisions. It also facilitates communication.

Interacting with a person or with another intelligent agent requires an intelligent agent to have the ability to operate on the same level of abstraction with a shared understanding of concepts and terminology. This includes, for example, how goals are formulated and how the intelligent agents present insights and decisions.

> BUSINESS STRATEGY PLANNING IS A GOOD EXAMPLE OF A HIGHLY ABSTRACT DOMAIN

Machine-learned models are numerical. They manage abstraction by mapping meaning to numerical values. This constitutes an implicit translation layer between the numerical representation and the abstract semantics.

Ontology-based models are symbolic. Within an ontology, objects are established and linked to each other using predicates. Machine reasoning draws inference from this representation by logical induction and deduction.

The symbolic representation assigned to objects, predicates and numeric values is convention. It is chosen to use the same abstraction and the same terminology as the domain it reflects. This facilitates an intuitive experience when users create and maintain the knowledge base.

Business strategy planning is a good example of a highly abstract domain. It deals with concepts such as growth, churn, customers, satisfaction and policy. Numerical data needs to be interpreted to deliver a meaningful contribution at this level. An intelligent agent performing this interpretation of data is a valuable assistant in business-level processes.

The introduction of intelligent agents will not make domain experts unnecessary. Instead, the task of the expert shifts from direct involvement in operational processes to maintenance of the models that dictate the operation of autonomous agents.

The abstraction of the models contributes to the efficiency of the domain expert. A practical example is the design of decision processes of expert systems proposing actions. These systems reach an answer by checking a tree of branching conditions. Even with a small number of variables, manually designing these conditions is a time-consuming and unintuitive task. An intelligent agent can compile the tree from knowledge about the reasons for proposing an action. Managing the abstract rules is a considerably more intuitive because the abstraction rises to the level the expert is used to thinking at.

Obtaining and managing knowledge

The intelligent digital assistant example (see proof of concept #1 on page 8) demonstrates an automated process that contributes knowledge. The assistant is generated from product manuals written in natural language by a document-crawler application. Based on existing knowledge, it identifies and classifies the information provided in the documents. It asserts this information as additional knowledge. Furthermore, site data stored in catalogs and inventories is automatically and continuously asserted in the knowledge base. This keeps the knowledge up-to-date, and the reasoning results adapt dynamically to changed facts.

The intelligent digital assistant also uses image recognition. It identifies physical elements and the current situation from images and asserts its findings in the knowledge base. This demonstrates a transformation of numeric data into symbolic knowledge. Deep-learning based neural networks are particularly successful at this task of identifying patterns in data and classifying them symbolically.

The intelligent digital assistant's use of image recognition and its ability to read natural language documents show that not all knowledge for machine reasoning needs to originate from a human domain expert. Machine-learning-based processes can add knowledge and keep it up-to-date based on what is learned from data.

In this respect, it is important to differentiate between data and knowledge. Data is values as provided by the environment. Knowledge is the interpretation of these values with respect to the semantics that are applied to give the data its meaning. Data and information models categorize data objects. Analytics creates further knowledge from multiple data elements and the domain context. A knowledge base preserves this knowledge for reasoning. When facing continuously changing data, a swarm of specialized intelligent agents can keep the knowledge up-to-date.

In machine learning, the learned model is the knowledge, and training examples are the main source. Domain experts are involved in selecting variables and data sources, and in configuring the learning processes according to use-case goals and constraints. The success of learning – and consequently, the performance of a learning-based intelligent agent – mainly depends on the availability and quality of training data.

Reinforcement learning is a variant of machine learning that learns from a set of rules and a simulation of the environment. Therefore, it does not necessarily depend on example data. However, the learned model is also not based on experience.

The manual design of knowledge by domain experts remains a major source of knowledge for machine reasoning. The domain experts create a stable core framework of asserted terminology and concepts. Based on this, they express their domain expertise by asserting further concepts and inference rules. They also design the applications that assess data source and automatically assert knowledge. This requires staff to be well trained in knowledge management, with efficient processes and tools for knowledge life-cycle management. A well-designed meta model establishes a standard for consistent knowledge representation. Any knowledge management competence gap can usually be filled by knowledge engineers, who can listen to the domain experts and transfer their knowledge into a model.

A major task in modeling is assembling a knowledge base according to use-case requirements. Ontologies can integrate and interconnect any formally defined model allowing extensive reuse. For example, data and information models used in application programming interface design constitute a foundation for asserting data objects. eTOM [4] and SID [5] are industry-standard models contributing common telecommunication terminology. TOVE [6] [7] or Enterprise Ontology [8] can cover business concepts. They were used in the business analytics orchestration example [9] (see page 9) for interpreting business-level questions.

An important part of the knowledge of autonomous intelligent agents is their goals. The domain expert uses goals to tell the intelligent agent what it is supposed to accomplish. Ideally, they are formulated as abstract business-level goals derived directly from the business strategy of the organization. This requires broad knowledge and adaptability to be built into the intelligent agents, but it promises a high level of autonomy.

> IT IS IMPORTANT TO DIFFERENTIATE BETWEEN DATA AND KNOWLEDGE

Proof of concept

#1: INTELLIGENT DIGITAL ASSISTANT

The intelligent digital assistant (see ***Figure 3***) is designed to assist field technicians who service base stations [10]. The technician interacts with the assistant through a mobile device. The assistant uses augmented reality to derive the base station type, configuration and state through object detection and visible light communication. For example, it can read the status LED of the device. The assistant provides instructions and visual guidance to the technician during maintenance operations. It downloads contextual data about the site and requests any additional information that could not be retrieved automatically through a query and answer dialogue.

The intelligent digital assistant is currently a proof of concept implemented by Ericsson Research. We have implemented and deployed the machine-reasoning system on backend servers. The system collects sensed input, analyses symptoms and presents corresponding maintenance procedures as a proposed series of actions. Domain experts have manually designed the procedural knowledge for problem resolution. Additionally, a document crawler automatically reads operational documentation, which allows the assistant to present documents that are relevant for the current tasks to the technician for reference.

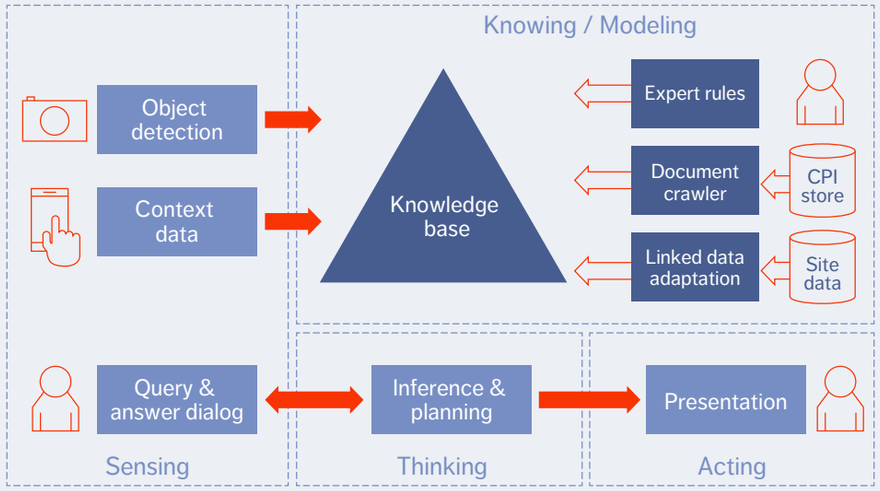


Figure 3: Intelligent digital assistant

Proof of concept

#2: BUSINESS ANALYTICS ORCHESTRATION

The business analytics orchestration use case (see ***Figure 4***) was implemented at Ericsson as a proof of concept within a master thesis project [9]. It demonstrates how the abstract level of business concepts can be linked with the technical level of data-driven analytics, so that intelligent agents can operate across the levels. The use case starts with a business question that can be solved through analytics. An intelligent agent acts as a business consultant, providing analytics-based assistance to a user. It analyzes the question, plans the needed analytics and orchestrates the execution of suitable analytics applications. When the results are available, the intelligent agent reasons about their meaning in the context of the question and explains the answer to the user.

The inference is based on a knowledge base that contains a combination of a business concept ontology and abstract service descriptions of analytics applications. It was built using existing and freely available business ontologies combined with manually-designed knowledge.

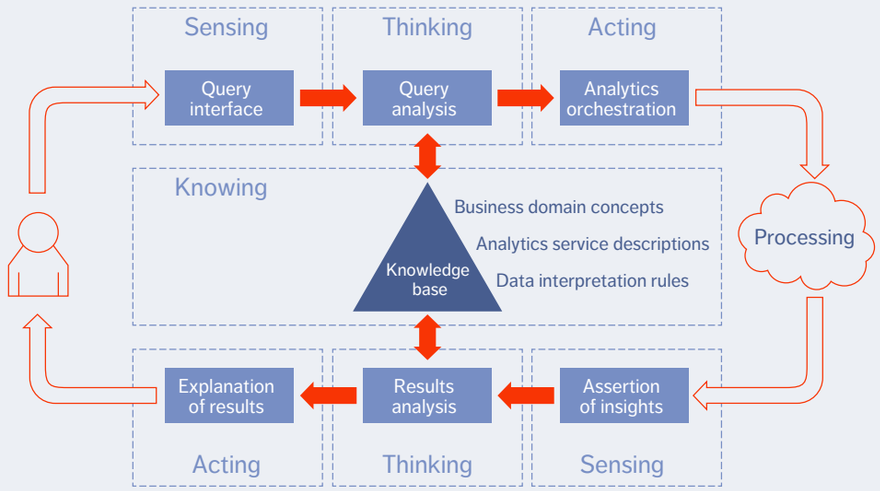


Figure 4: Business consulting through analytics

Machine learning and machine reasoning hybrid solutions

Good decisions and plans are often based on understanding multiple domains. For example, experts in network operation know about network incidents and the appropriate procedures to solve them. They can analyze technical root causes and apply corrective and preventive actions. The same experts usually also know some facts about the broader business environment. Knowing about financial goals and Service Level Agreements helps them to prioritize tasks. By understanding the application domain of a device or the concerns of a user, they can customize the solution. They might also know about marketing efforts or products in development and proactively provide consulting. All this knowledge allows an expert to make the right decisions. For intelligent agents, it is a challenge to operate with the same amount of diverse knowledge and to provide an equally diverse range of actions.

The role of machine reasoning

The knowledge used in machine reasoning is pure data decoupled from the implementation of the inference engine. Changes in behavior and extensions of scope must therefore be reached by changing the model data rather than the implementation of the intelligent agent. Therefore, machine-reasoning models are well suited to integrating ontologies and inference rules from multiple domains, if formal and semantic consistency is preserved.

Ideally, a layer of core concepts and terminology common to all domains should be used to anchor domain-specific models. This allows inference engines to traverse across domain borders and draw conclusions from all constituent domain models. If the models from different domains already use similar concepts, but define them differently, a "glue" model can relate them by introducing knowledge about the differences.

The drawbacks of the multi-domain knowledge base described here are the complexity of maintaining model consistency and the performance of the inference generation due to the number of knowledge elements to process.

The role of machine learning

In machine learning, each additional domain contributes yet another set of variables adding further numerical dimensions to the model. This introduces challenges such as the need for training examples that contain consolidated data samples from all domains. There is also an increase in the number of data points required to reach acceptable statistical characteristics. The combination of more dimensions and higher data volume increases the processing cost. Furthermore, each change in scope requires a full life-cycle loop including data selection, implementation, deployment and learning until a new model is available for productive use.

Considering these challenges, machine-learned models are best suited to be specialists in confined tasks. A secondary layer of models can then build on the specialist insights and evaluate them in a broader context. The second tier operates on higher abstraction with concepts from multiple domains. However, since training examples at this level are broad in scope, they tend to be hard to obtain. Domain experts are still available, though, so using machine reasoning is always feasible. In general, machine learning excels at inference that results from processing large amounts of data, while machine reasoning works very well in drawing conclusions from broad, abstract knowledge.

Hybrid solutions

The result is an environment comprised of orchestrated or choreographed intelligent agents. Coordination and collaboration is done through the knowledge. A machine-learned model can contribute its findings through asynchronous assertion. A mapping application is designed to monitor the numeric output of a machine-learned model or analyze the learned numeric model itself. When new output is generated, or a new version of the model is available, the mapping application interprets it in the domain context, determines its meaning and generates a respective symbolic representation. This constitutes new knowledge that is asserted into the knowledge base.

Alternatively, an application incorporating a machine-learned model can be linked directly into the knowledge base acting as a proxy for a knowledge object. A reasoning process would call the linked application when the respective knowledge is needed. The application generates a reply based on all currently available data.

Both methods create a hybrid of machine learning and machine reasoning that enables dynamic adaptation of the reasoning results based on learning and the latest data. Asynchronous assertion acts like a domain expert continuously updating knowledge. A knowledge proxy application synchronously generates knowledge on demand. However, this comes at the cost of delaying the reasoning process.

Symbolic neural networks

Symbolic neural networks specialize in learning about the relationships between entities. They implicitly abstract from an underlying statistical model, which allows them to answer abstract questions directly. One example is image processing combining multiple machine-learned models. One model identifies the objects seen. Another learns about the relationship between the objects. A third has learned to interpret questions asked. Due to the implicit abstraction and use of symbolic representation, the insights generated by these models would integrate seamlessly into a knowledge base and further reasoning. However, getting reference data for learning is a challenge in this scenario and would usually be dependent on human experts creating samples. As this setup has machine learning at its core, it also does not scale well to a high number of concerns and variables. Nevertheless, it can find and contribute knowledge about new relationships that was hitherto unknown to experts.

Tiered implementation

The tiered implementation approach uses machine learning on the layer of specialist models and machine reasoning for consolidation across domains. This assignment of roles reflects strengths of the technology families, although a different selection is possible depending on the use case and environment. For example, machine learning may be successfully applied for cross-domain consolidation if training data is available. And machine reasoning can implement a specialist intelligent agent, for example, if it incorporates the manually-designed rules of a human domain expert.

> "COGNITIVE TECHNOLOGIES WILL MAKE THIS FLEXIBLE, AUTONOMOUS ENVIRONMENT A REALITY"

Conclusion

Intelligent agents with the ability to work collaboratively present the best opportunity for network operators and digital service providers to create the extensively automated environment that their businesses will require in the near future. Cognitive technologies – and in particular a combined use of machine reasoning and machine learning – provide the technological foundation for developing the kind of intelligent agents that will make this flexible, autonomous environment a reality. These agents will have a detailed semantic understanding of the world and their own individual contexts, as well as being able to learn from diverse inputs, and share or transfer experience between contexts. In short, they are capable of dynamically adapting their actions to a broad range of domains and goals.

References

1. **Acadia University, On Common Ground: Neural-Symbolic Integration and Lifelong Machine Learning (research paper), Daniel L. Silver, available at:** *http://daselab.cs.wright.edu/nesy/NeSy13/silver.pdf*
2. **Ericsson Technology Review, Generating actionable insights from customer experience awareness, September 30, 2016, Niemöller, J; Sarmonikas, G; Washington N, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2016/generating-actionable-insights-from-customer-experience-awareness*
3. **Annals of Telecommunications, Volume 72, Issue 7-8, pp. 431-441, Subjective perception scoring: psychological interpretation of network usage metrics in order to predict user satisfaction, 2017, Niemöller, J; Washington, N, abstract available at:** *https://link.springer.com/article/10.1007%2Fs12243-017-0575-6*
4. **TM Forum, GB921 Business Process Framework (eTOM), R17.0.1, available at:** *https://www.tmforum.org/resources/suite/gb921-business-process-framework-etom-r17-0-1/*
5. **TM Forum, GB922 Information Framework (SID), Release 17.05.1, available at:** *https://www.tmforum.org/resources/suite/gb922-information-framework-sid-r17-0-1/*
6. **Berlin: Springer-Verlag, pp. 25-34, The TOVE project towards a common-sense model of the enterprise, Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, 1992, Fox, M.S., available at:** *https://link.springer.com/chapter/10.1007/BFb0024952*
7. **University of Toronto, TOVE Ontologies, available at:** *http://www.eil.utoronto.ca/theory/enterprise-modelling/tove/*
8. **Cambridge University Press, The Knowledge Engineering Review, Vol. 13, Issue 1, pp. 31-89, The Enterprise Ontology, March 1998, King, M; Moralee, S; Uschold, M; Zorgios, Y, abstract available at:** *https://www.cambridge.org/core/journals/knowledge-engineering-review/article/enterprise-ontology/17080176D5F06DEAEA8DBB2BAA9F8398*
9. **Tilburg University, Mediating Insights for Business Needs, A Semantic Approach to Analytics Orchestration (master's thesis), June 2016, Alhinnawi, B.**
10. **Ericsson Mobility Report 2018, Applying machine intelligence to network management, Stephen Carlsson, available at:** *https://www.ericsson.com/en/mobility-report/reports/june-2018*

THE AUTHORS

Jörg Niemöller
◆ is an analytics and customer experience expert in solution area OSS. He joined Ericsson in 1998 and spent several years at Ericsson Research, where he gained experience of machine-reasoning technologies and developed an understanding of their business relevance. He is currently driving the introduction of these technologies into Ericsson's portfolio of Operations Support Systems / Business Support Systems solutions. Niemöller holds a degree in electrical engineering from TU Dortmund University in Germany and a Ph.D. in computer science from Tilburg University in the Netherlands.

Leonid Mokrushin
◆ is a senior specialist in cognitive technologies at Ericsson Research. His current focus is on investigating new opportunities within artificial intelligence in the context of industrial and telco use cases. He joined Ericsson Research in 2007 after postgraduate studies at Uppsala University, Sweden, with a background in real-time systems. He received an M.Sc. in software engineering from Peter the Great St. Petersburg Polytechnic University, Russia, in 2001.

Further reading

- **CIO, Artificial intelligence is about machine reasoning – or when machine learning is just a fancy plugin, November 3, 2017, Rene Buest, available at:** *https://www.cio.com/article/3236030/machine-learning/artificial-intelligence-is-about-machine-reasoning-or-when-machine-learning-is-just-a-fancy-plugin.html*
- **Microsoft Research, From machine learning to machine reasoning – An essay, February 13, 2013, Léon Bottou, available at:** *https://www.microsoft.com/en-us/research/wp-content/uploads/2014/01/mlj-2013.pdf*

LEVERAGING LTE AND 5G NR NETWORKS FOR

Fixed wireless access

Globally, there is a huge underserved market for broadband connections, with more than one billion households still unconnected. The growth in high-speed mobile broadband coverage enabled by LTE and 5G New Radio is opening up much more commercially attractive opportunities for operators to use fixed wireless access to deliver broadband services to homes and small and medium-sized enterprises.

HÅKAN OLOFSSON, ANDERS ERICSSON, FREDRIC KRONESTEDT, SVEN HELLSTEN

Unlike the country-wide decisions typically made for mobile broadband (MBB), decisions about fixed broadband and targeted fixed wireless access (FWA) deployments tend to be made at the local market level, and operators have a critical role to play. A number of different drivers govern local market attractiveness, as outlined in Ericsson's recently published FWA handbook [1].

We have organized the FWA market opportunities into three distinct segments that we call 'Wireless Fiber', 'Build with Precision', and 'Connect the Unconnected'. Each of these has different characteristics mainly based on the offering, the availability of fixed access and the corresponding average revenue per user (ARPU) that can be expected from customers [1]. The Wireless Fiber segment consists of those cases in which there is a need for very high-rate offerings and capacity as a direct alternative to high-end fixed broadband. The ambition is to provide fiber-like speeds and handle households' TV needs, matched with a correspondingly high ability to pay. Typical sold date rates are 100 to 1,000+ Mbps and monthly ARPU levels of USD 50-100. The FWA sweet spot for this segment is typically suburban environments.

The Build with Precision segment is comprised of those cases where there is competition from performance-limited fixed broadband alternatives, such as xDSL. Here, the need is for high data rate and capacity, with a corresponding level of ARPU. Typical sold data rates are 50 to 200Mbps and monthly ARPU levels are around USD 20-60. The FWA sweet spot for this segment is in suburban or rural villages or towns that are currently underserved. Some more sparsely populated areas are also addressable.

The Connect the Unconnected segment is made up of cases in which fixed broadband competition is virtually non-existent, and smartphones that use MBB are the dominant way of accessing the internet. User expectations of access speed are relatively low. Typical sold data rates are 10 to 100Mbps and monthly ARPU levels are around USD 10-20. Even though ARPU levels are limited in this segment, it has a FWA sweet spot that stretches from urban environments to rural villages, due to limited investment needs.

Subscriptions, data rates and consumption

The paradigms for fixed broadband and MBB are different, both in terms of subscription offerings and dimensioning. Fixed broadband subscriptions tend to focus on maximum data rates that are achieved under normal circumstances – that is, at low to medium load. The user traffic is often shaped so that it does not exceed the sold data rate. Hence, for fixed broadband, the sold data rate is the normal value that household subscribers relate to.

By contrast, for MBB, peak rates are sometimes used for marketing, and normally the network transmits the maximum rate that the mobile device can handle. Monthly data buckets dominate the subscription paradigm, and additional monetization is achieved through upgrades to larger data buckets, all the way to unlimited data. Hence, for MBB, monthly data buckets are the normal subscription value that mobile subscribers relate to.

Terms and abbreviations

ADSL – Asymmetric Digital Subscriber Line | **ARPU** – Average Revenue per User | **CAT** – Category (in LTE) | **CPE** – Customer Premises Equipment | $\mathbf{d_{av}}$ – Average Busy-hour Data Consumption | **DL** – Downlink | **DSL** – Digital Subscriber Line | **FDD** – Frequency Division Duplex | **FWA** – Fixed Wireless Access | **MBB** – Mobile Broadband | **MIMO** – Multiple-input, Multiple-output | **mmWave** – Millimeter Wave | **NR** – New Radio | $\mathbf{R_{min}}$ – Minimum Data Rate | **TDD** – Time Division Duplex | **Tx/Rx** – Radio Transmitter/Radio Receiver | **WCDMA** – Wideband Code Division Multiple Access | **xDSL** – DSL family (e.g. ADSL)

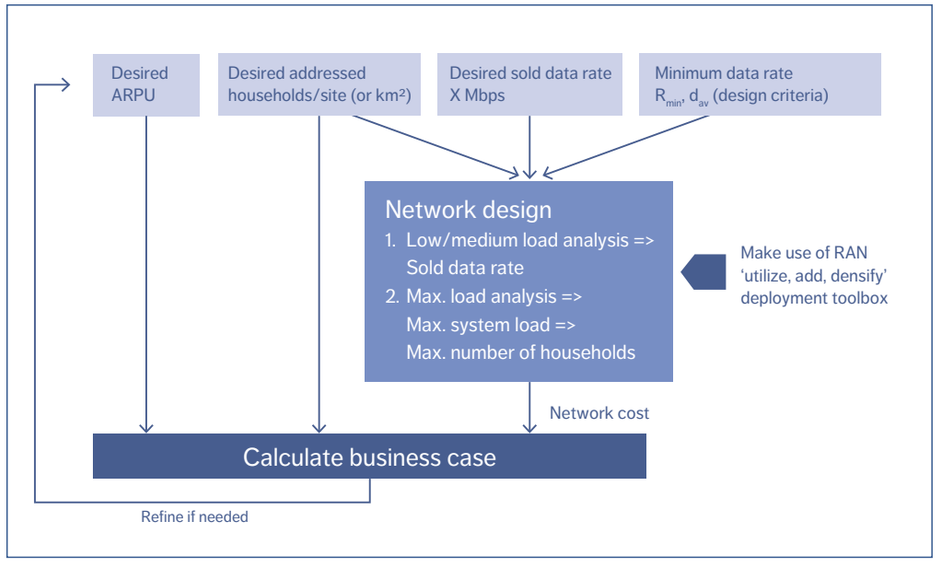


Figure 1: FWA deployment analysis flow

It is important that both consumers and operators (fixed and mobile) understand this crucial difference. Our view is that FWA will inherit the subscription paradigms of fixed broadband rather than those of MBB. That is, households should pay for FWA on the basis of data rate and not be concerned about data consumption.

Last-hop dimensioning

In FWA the last hop is wireless, so all the characteristics of a wireless network apply to the dimensioning. Unlike fiber, but similar to digital subscriber line loop length, there will be varying connection quality to different households. And, unlike fixed broadband overall, the last hop is radio and therefore shared, which means that speeds will degrade with increasing network load. All these characteristics must be taken into account when dimensioning an FWA network. Further, since Ericsson promotes the sharing of assets with MBB (when available), we recommend that FWA is brought into general RAN dimensioning.

Note that for fixed broadband, FWA and MBB alike, there is transport aggregation above the last hop, which is dimensioned according to standard principles and can also contribute to a varying user experience.

In short, while FWA will inherit the subscription paradigm of fixed broadband, due to the radio properties of the last hop to households, it must use modified dimensioning methods and terms from the MBB paradigm.

Figure 1 illustrates a typical FWA analysis flow. It starts with input on the subscription and offering, including dimensioning criteria, which triggers a selected, maximally efficient network design that depends on the offering ambitions and network starting point. A business case can be calculated by balancing the resulting cost items of the deployment with the extra revenues foreseen.

FWA toolbox

An existing mobile radio network, normally designed for voice and MBB, is an excellent base for offering an FWA service. Depending on the radio network starting point and the operator's ambitions for FWA, there is a toolbox available to make the network capable of handling a combination of voice, MBB and FWA.

These tools fall into three main categories: utilize, add and densify. The particular needs of each local situation can be met by deploying a well-planned mix of these tools.

Utilize existing radio network assets

The ability to utilize existing radio network assets is a fundamental advantage that sets mobile operators apart from start-ups or greenfield competitors in the FWA market. However, the advantage is only fully realized if all relevant RAN assets are efficiently combined for voice, MBB and FWA. If the operator chooses not to utilize existing assets built for voice and MBB, the number of economically viable local areas for FWA will be smaller, and the operator risks facing unnecessary competition with standalone FWA providers.

The radio network assets that should be utilized include existing radio sites, spare capacity in deployed spectrum (including associated equipment), and acquired but undeployed spectrum. Existing radio sites are critical assets, whether they are operator-owned or rented. The 'tool' of utilizing existing sites is not used by itself, but in combination with other actions to make those more cost-efficient. Spare capacity in deployed spectrum and associated deployed radio, baseband and transport network equipment is quite common in FWA target areas, and making use of it requires no new capital expenditure. Acquired but undeployed spectrum is also common in FWA target areas, which makes radio deployment in new bands possible without the cost of acquiring new spectrum. The geographical fit for FWA is excellent, since FWA targeted areas are often suburban and rural, where unused spectrum is most prevalent.

Add radio network capabilities

In an MBB RAN, radio capabilities are continuously added to handle more traffic, more customers and better app coverage. To handle FWA as an extra service, some of these additions may have to be made sooner to achieve a combined network with sufficient capabilities.

An existing mobile operator has the significant advantage of being able to add the following radio network capabilities and co-finance them for MBB and FWA:

- Spectrum – upcoming wide spectrum bands in 3-6GHz and millimeter wave (mmWave) open up potential for providing high data rates and capacity, benefiting both MBB and FWA
- Higher-order modulation, multiple-input, multiple-output (MIMO) and beamforming – offering the potential to squeeze out the most from each spectrum band
- FWA-tailored software features – to enhance performance for FWA users and to provide adequate quality to MBB and FWA in shared deployments
- Additional sectors on existing sites
- 5G New Radio (NR) access – designed for low latency and for wide spectrum bands, creating an excellent overall network together with LTE.

> THE ABILITY TO UTILIZE EXISTING RADIO NETWORK ASSETS IS A FUNDAMENTAL ADVANTAGE

Densify the radio network grid

When the 'utilize' and 'add' tools have been used to their full potential, densification can offer further gains. In these cases, MBB enhancements tend to be necessary as well, so the upgrade needs of MBB and FWA should be considered together and the densification of the network should be co-financed.

The two options for densifying the radio network grid are macro site densification and small cell site

> QUALITY ACROSS BOTH SERVICES IS ENSURED THROUGH EXISTING SOFTWARE FEATURES

densification on poles. Macro site densification is an opportunistic approach: where new macro sites can be found, such opportunities can be taken. Small cell site densification on poles may be necessary if the macro grid is sparse and performance requirements are high.

Spectrum sharing across MBB and FWA

Sharing spectrum across FWA and MBB enables significant gains in overall spectral efficiency because higher utilization is possible with one 'bigger pipe'. This is explained by the trunking gain effect, which has been known and used in mobile systems since their infancy, all the way from voice channel capacity to LTE carrier aggregation for MBB. It is also applicable to FWA.

The logical consequence of this is that spectrum assets should be shared as one pool, employing carrier aggregation for LTE and dual connectivity for LTE/NR to ensure that all resources are utilized to the maximum, while securing good user experience for both MBB and FWA. Quality across both services is ensured through existing software features such as RAN slicing.

By contrast, any artificial split of spectrum resources for different services would result in under-utilization of the spectrum assets.

Performance differences of FWA CPE types

Using FWA to deliver broadband services requires new FWA customer premises equipment (CPE), from simple indoor nomadic devices to fixed outdoor-installed units, provisioned through standard device retail or new methods. A CPE management system is likely to be needed to manage CPE in the fixed broadband sense – enabling the operator to log in to the devices, configure them and check status remotely. Converged operators have the choice of reusing the fixed access CPE management system or deploying a separate one for FWA. Both CPE and CPE management systems are separate network entities that generally have limited integration with cellular networks, meaning that the operator can acquire best-of-breed products and expect them to work using standard protocols. The biggest difference between the CPE alternatives is the ability to achieve promised service levels, especially during busy hours.

An outdoor CPE provides the best performance, as it has a built-in directional antenna (3.5GHz, 10-14dBi) and is installed with a predictable radio link quality to the selected base station. The typical antenna configuration has two Rx antennas, but devices with four Rx antennas are also available. The normal transmission mode is rank-2 MIMO, as the modem is expected to be installed with good line-of-sight. Most outdoor LTE devices support CAT 6 and 20+20MHz carrier aggregation but more advanced devices up to CAT 16 support are also available. Inter-band carrier aggregation between FDD and TDD is especially useful, as services can be started on existing FDD bands and later expanded as FWA subscribers and traffic increase.

A correctly installed outdoor CPE is directed to the best-serving cell, leading to a lower path loss and increasing the value of mid-band and mmWave TDD spectrum. The large gain in signal quality is a result of the 10dB difference in antenna gain and the avoidance of 10-15dB in wall/window attenuation losses suffered by indoor devices. Another contributor to signal attenuation for indoor devices is the deep indoor loss, as the device is likely to be placed in a hidden location or to provide optimum Wi-Fi coverage. This could contribute another 5dB in path loss.

Whereas an indoor CPE is comparable to a smartphone in terms of spectrum efficiency, an outdoor CPE is two to three times more efficient. To put it another way, for the same data consumption, around two to three times as many households can be served using outdoor rather than indoor units – or two to three times as much spectrum would be needed to serve indoor-only FWA households. A final advantage of outdoor CPE is that the relative performance difference between the best, median and worst five-percentile users is significantly lower.

In terms of performance, indoor CPE units normally start with CAT 6 capabilities of up to 300Mbps. More advanced devices could support CAT 16 up to 1Gbps and offer rank-4 MIMO. More advanced CPE architectures are also being discussed, such as a split design, where an outdoor window antenna is connected to an indoor unit via induction through the window glass.

Case study: the country town

The country town example represents a market within our Build with Precision segment, characterized by relatively mature LTE MBB and decent fixed broadband offerings, complemented by terrestrial or satellite broadcast services to meet households' linear TV needs. The typical monthly ARPU for MBB is around USD 20, and the predicted willingness to pay is USD 40 for a dedicated household FWA internet service with a sold rate of 50-200Mbps and unlimited data.

The operator uses the following as the basis for dimensioning the system:

- The network should be designed to be able to connect at least 30 percent of households. In contrast to the extensive upfront investments required in a fiber deployment, the ability to design and invest for a limited market share from the beginning and expand later as the subscriber base grows is a useful property of FWA.
- There is no ambition to offer IPTV over FWA, as household TV needs are assumed to be served by satellite or terrestrial access.
- The dominant use case is meeting all the households' internet needs.
- For video streaming support, households should, when needed, experience at least a minimum data rate (Rmin) of 10Mbps even during busy hours. This corresponds to one high definition TV video stream, with some margin, or a combination of multiple standard definition TV streams.
- Based on the operator's experience from similar FWA areas, the average household's consumption during busy hours is 0.9GB/h, corresponding to an average data flow of 2Mbps during busy hours. With the assumption that 10 percent of data is being consumed during busy hours, this would correspond to 270GB per month.

Network starting point

Coverage is provided by a macro network with three-sector sites and an inter-site distance of about 1km. The operator has access to six FDD bands: three bands below 1GHz (typically 700, 800 and 900MHz), and three bands in the 1-3GHz range (typically 1,800, 2,100 and 2,600MHz). The MBB traffic in this area is handled using a subset of the available bands. The majority of smartphones are LTE-capable, and there is also GSM and WCDMA coverage to handle simpler phones. A typical macro site has two LTE carriers (800 and 1,800MHz) as well as a WCDMA carrier in the 2,100MHz band, and a few GSM carriers in the 900MHz band.

High-level analysis [2] has shown that the deployed LTE capacity in western and central Europe is less than 40 percent utilized, given the LTE smartphone subscriber density in the area. This means that there is spare radio capacity that can be utilized by FWA.

> THE ABILITY TO DESIGN AND INVEST FOR A LIMITED MARKET SHARE... IS A USEFUL PROPERTY OF FWA

Overall solution

We recommend utilizing the existing sites, radios and baseband deployed to provide MBB, and sharing these resources across FWA and MBB users. Current deployments have spare capacity both in LTE carriers and in baseband units. In addition, we recommend utilizing the acquired

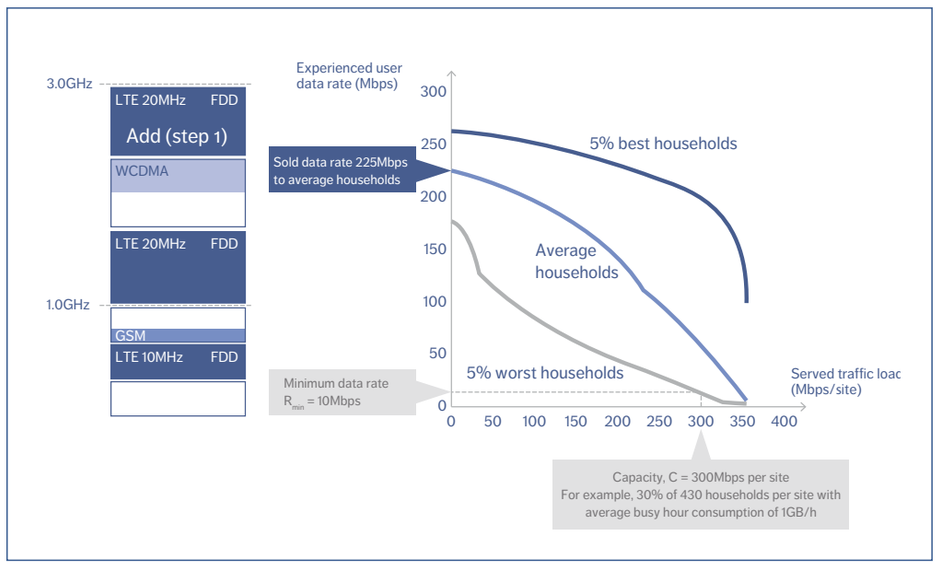


Figure 2: Performance and spectrum use of FWA deployment step 1

but undeployed band below 3GHz (such as 2,600MHz), with a new 2 Tx/Rx radio, together with the existing LTE bands by means of carrier aggregation for both FWA and MBB. Carrier aggregation improves peak speeds as well as coverage for both services. The left side of *Figure 2* shows the spectrum use of the FWA deployment at this first step deployment. A RAN slicing feature can be applied to ensure that there is no negative impact on MBB services (and vice versa) during peak loading as a result of FWA and LTE users sharing the same carriers.

There is no need to densify the network in this case. With regard to CPE choices, we suggest using high-end 4 Rx outdoor (roof-top mounted) CPE, as FWA speeds need to be high in this case to compete with xDSL services in the area. Indoor CPE may be deployed as a complement for households where their performance is acceptable.

Performance analysis

Although MBB and FWA services share spectrum in the country town case, to simplify the presentation of the performance analysis, our evaluation only shows FWA. Further, we have chosen to focus on the downlink (DL) because the FWA traffic (and broadband traffic in general) is DL-heavy and so capacity is DL-limited.

The performance is illustrated in Figure 2. The experienced DL data rate for a specific household depends on its location, as with xDSL services, and may be up to 270Mbps in this scenario. An average household would experience around 225Mbps at low system load. This could be used as the sold data rate to a typical customer.

Note however that, unlike MBB, where users move around and experience both good and bad radio environments, in this scenario the CPE is fixed and variation in the radio environment is smaller,

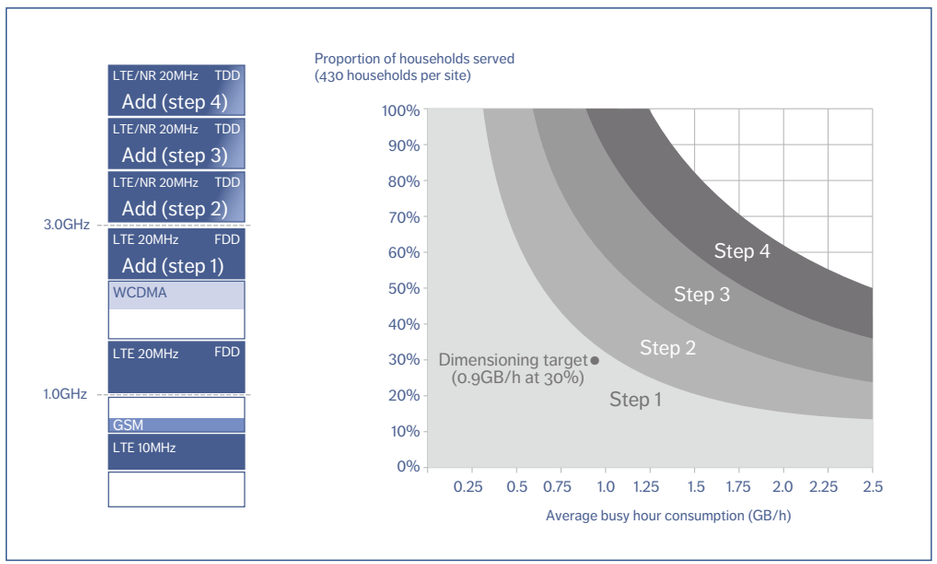


Figure 3: FWA deployment solution evolution to steps 2-4: spectrum use and performance

meaning that households with worse radio environments will likely always have worse than average data rates. In this scenario, the five percent worst-performing households experience close to 175Mbps at best. Therefore, it may be worth considering having different subscription categories; it may not be possible for all households to subscribe to the higher service level.

To dimension the system, the Rmin is set to 10Mbps. This means that the five percent worst-performing households should experience at least 10Mbps DL data rate during busy hours. This results in a capacity of 300Mbps, or 135GB/h, per site. As long as the total traffic in all three sectors does not exceed 300Mbps, the Rmin requirement will be fulfilled.

Assuming there are 500 households per square kilometer, and an inter-site distance of 1,000m, an FWA market share of 30 percent corresponds to some 130 households per site. At 135GB/h capacity, this market can be served with an average busy hour consumption of slightly above 1GB/h – that is, above the dimensioning target of 0.9GB/h (2Mbps). In addition, MBB will benefit from the additional 20MHz spectrum, for example in terms of increased peak rates.

Solution evolution

It is important that the solution is future-proof and can evolve to handle more connected households and higher demand per household over time. To provide higher capacity and cope with greater demands, operators can acquire and add a new TDD band above 3GHz (such as 3.5GHz) using 8 Tx/Rx advanced antenna system radios. The multi-user MIMO feature can be activated to provide additional capacity. *Figure 3* illustrates how additional capacity can be provided in several evolution steps.

Initially, the system is dimensioned to serve 30 percent of households with an average busy hour consumption of 1GB/h. The area of the graph in Figure 3 marked as Step 1 indicates the possible combinations of percentages of households and average busy hour consumption for this solution.

The area of the graph that is marked as Step 2 indicates the capacity provided by an additional 20MHz. This shows that the system can serve a customer base of 30 percent with an average busy-hour consumption of 1.9GB/h. Alternatively, the higher capacity can be used to serve an increased market share (up to 58 percent) with an unchanged average busy hour consumption.

Increasing the bandwidth with another 20MHz of TDD spectrum provides a system capacity represented by the area marked Step 3 in the graph. This will serve 30 percent of households in the area with an average busy-hour consumption of 3GB/h. Again, the higher capacity could instead be used to serve an increased market share with an unchanged average busy-hour consumption, or a combination of increased market share and increased average consumption.

Finally, Step 4, the darkest grey area of the graph in Figure 3, indicates what can be achieved when a total of 60MHz of TDD spectrum is added beyond Step 1. Assuming a 30 percent market share, an average busy-hour consumption of up to 4.1GB/h can be met (outside graph range).

In summary, by using the FWA toolbox and limited initial investments, and then adding TDD spectrum as needed, the chosen deployment is able to support high data rates and consumption immediately at launch. Then, through a series of smooth solution evolution steps, capacity can grow to more than four times the initial offering.

Conclusion

The large number of underserved households around the world represents a profitable FWA growth opportunity for current 3GPP operators. Mobile-only operators can explore a new business opportunity with FWA, and converged operators can add FWA as a complement to their fixed broadband strategy for certain locations as a more cost-efficient solution with faster time to market. Segmented solutions are needed, with subscriptions and dimensioning based on fixed and mobile paradigms. We believe that the best way to deliver future-proof broadband solutions is based on the evolution of LTE and 5G NR, and that the most promising approach is shared investment using the same ecosystem, assets and spectrum bands for both MBB and FWA.

References

1. **Ericsson, Fixed Wireless Access Handbook (extracted version), available at:** *https://www.ericsson.com/assets/local/narratives/networks/documents/fwa-handbook.pdf*
2. **Ericsson Mobility Report, November 2017, available at:** *https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-november-2017.pdf*

Further reading

- **Ericsson Technology Review, Fixed wireless access on a massive scale with 5G, 2017, Furuskär A; Laraqui, K; Nazari, A; Skubic, B; Tombaz, S; Trojer, E, available at:** *https://www.ericsson.com/assets/local/publications/ericsson-technology-review/docs/2017/2017-01-volume-94-etr-magazine.pdf*
- **Ericsson ConsumerLab, Connected homes, June 2015, available at:** *https://www.ericsson.com/assets/local/news/2015/6/ericsson-consumerlab-connected-homes.pdf*

The authors would like to thank the following people for their contribution to this article: Tomas Dahlberg, Hani Elmalky, Bo Göransson, Henrik Johansson, George Jöngren, Michael Kühner, Per Lindberg, Staffan Lindholm, Reiner Ludwig, Claes Martinsson, Björn Möller, Richard Möller, Per Arne Nilsson, Christoph Schrimpl-Rother, Sibel Tombaz, Henrik Voigt, David Waite and John Yazlle.

THE AUTHORS

Håkan Olofsson

◆ has worked in the mobile industry for 25 years, with a particular focus on its RAN aspects. He joined Ericsson in 1994 and has served in a variety of capacities, mostly dealing with strategic technology development and the evolution from 2G all the way to 5G. He is currently head of the System Concept program at Development Unit Networks. Olofsson holds an M.Sc. in physics engineering from Uppsala University, Sweden.

Anders Ericsson

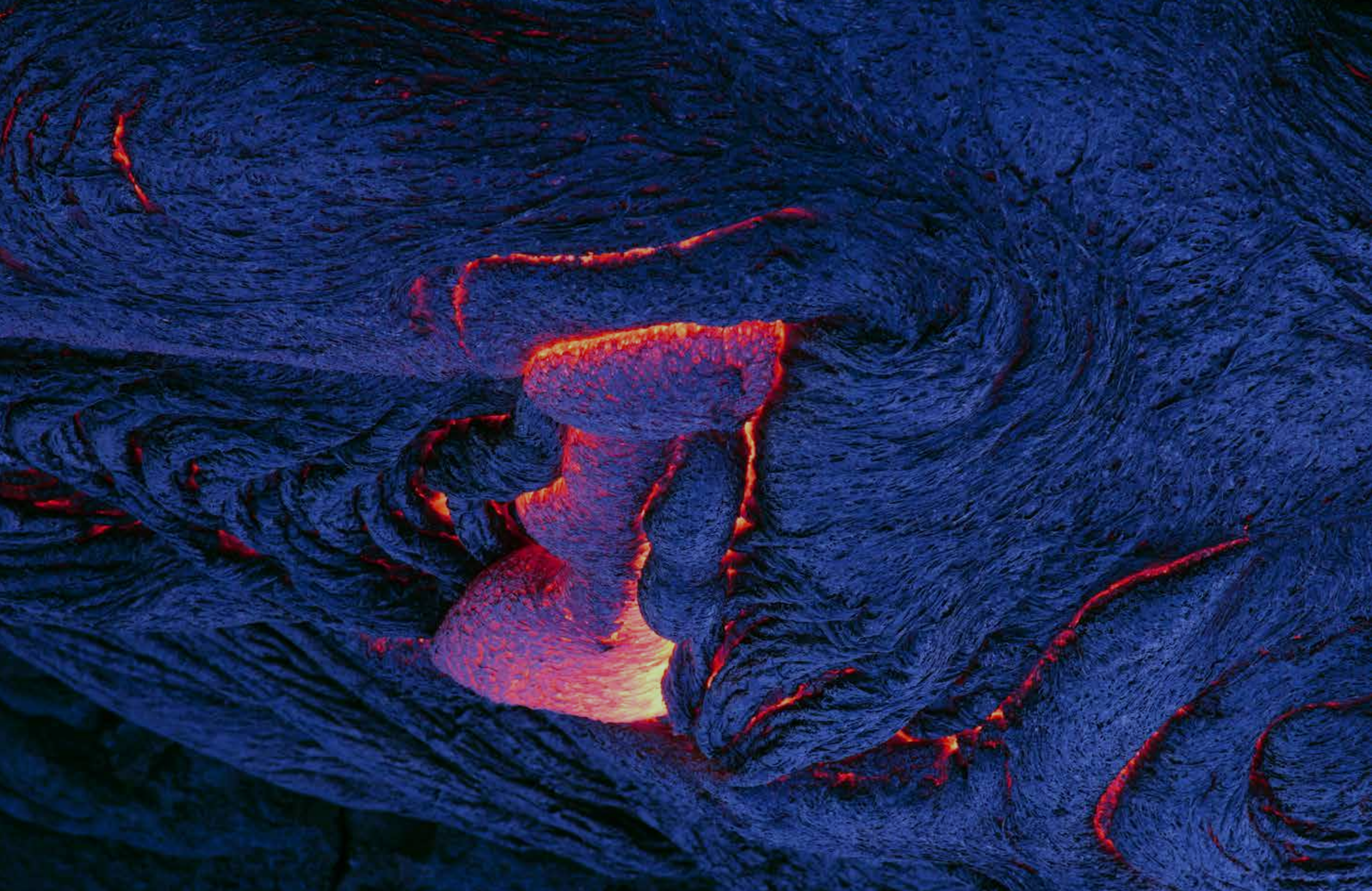
◆ joined Ericsson in 1999 and is currently working as a system designer in Development Unit Networks. During his time at Ericsson, he has worked at Ericsson Research and in system management, as well as heading up the Algorithm and Simulations department at Ericsson Mobile Platforms/ST-Ericsson. He previously worked at the Swedish National Defense Research Establishment (FOI). Ericsson holds a Licentiate Eng. in automatic control and an M.Sc. in applied physics and electrical engineering from Linköping University, Sweden.

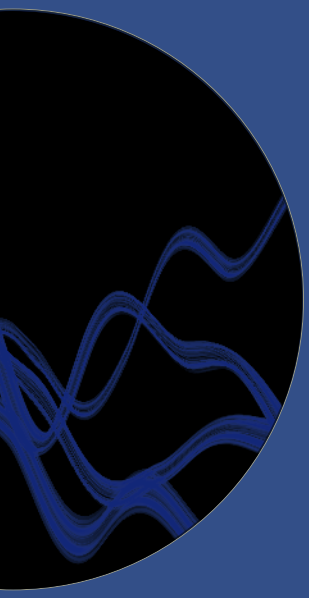
Fredric Kronestedt

◆ joined Ericsson in 1993 to work on RAN research. Since then, he has taken on many different roles, including system design and system management. He is currently working at Development Unit Networks, where he is focusing on radio network deployment and evolution aspects for 4G and 5G. Kronestedt holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology, Stockholm, Sweden.

Sven Hellsten

◆ joined Ericsson in 1993 and over the years he has worked with radio technologies ranging from analogue AMPS, GSM, WCDMA, to LTE and 5G/NR. His main focus has been on product management of base stations, but he has also worked with signal processing design and systems management. Hellsten holds an M.Sc. in physics engineering from Uppsala University.





ISSN 0014-0171
284 23-3322 | Uen

© Ericsson AB 2018
Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000