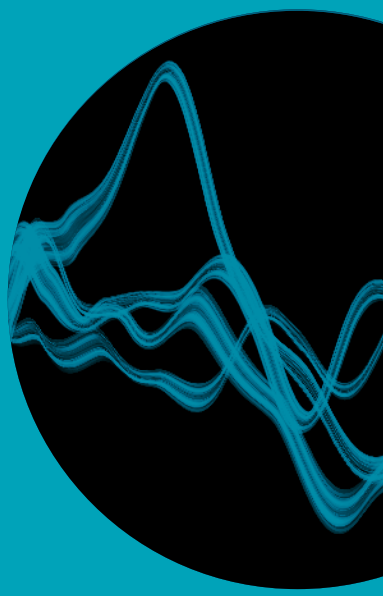
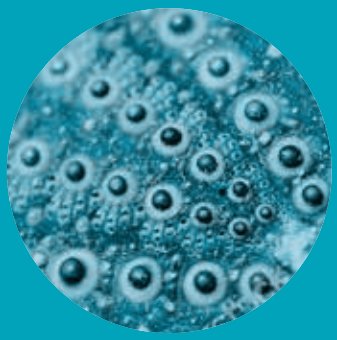
CHARTING THE FUTURE OF INNOVATION | VOLUME 96 I 2018-01

ERICSSON TECHNOLOGY Review



MACHINE INTELLIGENCE & INDUSTRIAL AUTOMATION

5G NR PHYSICAL LAYER



IOT SECURITY MANAGEMENT

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, Peter Öhman, Håkan Olofsson, Dan Fahrman, Robert Skog, Patrik Roseen, Jonas Högberg, John Fornehed and Sara Kullman

FEATURE ARTICLE
Industrial automation enable by robotics, machine intelligence and 5G by Roberto Sabella, Andreas Thuelig, Maria Chiara Carrozza and Massimo Ippolito

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: 96, 2018

5G, IoT AND THE NEXT INDUSTRIAL **REVOLUTION**

WE ARE PUBLISHING this magazine shortly after the first release of 5G from 3GPP, which is significant because the first release of a completely new standard only happens every 10 years. Fittingly, many of the articles in this issue relate to what we think is most important in 5G and how to address the new opportunities that 5G entails.

Defining New Radio (NR) was a key focus in the first release of 5G. While the development of the radio for a new generation has traditionally focused on the introduction of a new modulation and coding scheme, the main focus this time has been on flexibility to support a large range of devices and services with vastly different characteristics, different types of deployments and frequency allocations that range from below 1 GHz well up into the mmWave bands. To support the expected growth in data volumes, innovative technologies in the area of massive antenna systems, beamforming and energy efficiency have been introduced.

One of the key reasons for the flexibility provided in 5G is the desire to support industries to use connectivity, virtualization, machine intelligence and other technologies to change their processes and business models as part of the next industrial revolution, Industry 4.0. It is therefore a pleasure to be able to include an article in this issue that we have co-written with Comau and the Sant'Anna School of Advanced Studies on the topic of industrial automation.

In this issue we also cover the topic of 5G network programmability, which allows the network platform to support a wide range of applications, including the mission critical applications that are required to support industrial automation. We have also included an article on the subject of intelligent transport that describes a transport network solution that connects all the pieces of the RAN and the mobile core network, and where high levels of intelligence, flexibility and automation are used to provide optimal performance for a variety of different 5G scenarios.

We know that the protection of both networks and data is a key requirement for any enterprise or industry that uses LTE and 5G networks. We address this topic in an article that looks specifically at end-to-end (E2E) security management for the Internet of Things (IoT).

We also know that video is sure to play in important role in many 5G applications. With the increased consumption of video over mobile networks, it is crucial to ensure a high-quality viewing experience. In light of this, Ericsson has been part of standardizing a model for measuring viewing quality, which we present here in an article about video QoE (Quality of Experience).

With the first 5G standard released and commercial deployments starting this year, we are entering a new era of mobile communications with the potential to profoundly change the way that many businesses and industries operate. This change may well be similar to the way that the introduction of mobile phones changed the way people interact both with each other and with different applications. We will continue to explore the possibilities of 5G in future ETR articles.

If you would like to see the contents of this magazine in digital form, you can find all of the articles, along with those published in previous issues, at: www.ericsson.com/ericsson-technology-review

> WE ARE ENTERING A NEW ERA OF MOBILE COMMUNICATIONS WITH THE POTENTIAL TO PROFOUNDLY CHANGE THE WAY THAT MANY BUSINESSES AND INDUSTRIES OPERATE

ERIK EKUDDEN
SENIOR VICE PRESIDENT,
GROUP CTO AND
HEAD OF TECHNOLOGY & ARCHITECTURE

Video QoE

LEVERAGING STANDARDS TO MEET RISING USER EXPECTATIONS

'How happy are our users with their video experience?' has become a vital question for mobile network operators and media service providers alike. New standards for QoE testing have the potential to help them ensure that they are able to meet user expectations for the service that will account for three-quarters of mobile network traffic in five years' time.

GUNNAR HEIKKILÄ, JÖRGEN GUSTAFSSON

In 2016, Ericsson ConsumerLab found that the average person watches 90 minutes more TV and video every day than they did in 2012 [1]. While traditionally broadcast linear TV is still popular with many viewers, internet-based TV and video delivery and video on demand (VOD) services are all growing rapidly. In fact, 20 percent of video consumption occurred on handheld devices in 2016 [1].

As users become more accustomed to video streaming services, their quality expectations rise, presenting a big challenge for media service providers (MSPs) and mobile network operators (MNOs). While delivering high-quality video over a fixed connection can be difficult, doing so wirelessly is a much more demanding undertaking. Yet by 2022, 75 percent of all mobile data traffic is expected to come from video, according to the 2016 Ericsson Mobility Report [2].

At the same time, TV screens are getting larger, which requires higher video resolution. High definition (HD) is now the new baseline, and ultra high definition (UHD) is coming to both fixed and mobile devices, demanding higher bandwidth. To provide a consistently high QoE – particularly in wireless cases with significant fluctuations in available bandwidth – both MSPs and MNOs must have a clear understanding of which impairments their users are experiencing and be able to accurately assess their perception of video quality.

Adaptive HTTP-based streaming

Adaptive HTTP-based video streaming variants such as DASH and HLS are the dominant video delivery method used today. Their streaming servers provide several versions of the same video, each separately encoded to offer varying levels of quality. The streaming client in the user's phone, tablet or PC dynamically switches between the different versions during playout depending on how much network bitrate is available.

If the available network bitrate is high, the client will download the best-quality version. If the bitrate suddenly drops, the client will switch to a lower-quality version until conditions improve. The purpose of this is to avoid stalling (when the client stops playout intermittently to fill its video buffer) – which is known to annoy users.

While the ability to switch intermittently between versions of a video significantly decreases the risk of stalling, the quality variations that this leads to can also be annoying for the user. *Figure 1* provides an example of a worst-case scenario with both quality variations and a stalling event, which would result in an overall low-quality experience for the user.

Subjective quality

ITU-T P.910 [3] is the recognized standard for performing subjective video quality tests. The tests are carried out in a lab equipped with mobile phones, tablets, PCs or TVs, where a number of videos are shown to a group of individuals. Each individual then grades each video according to their subjective perception of its quality, selecting one of the following scores: 5 (excellent), 4 (very good), 3 (good), 2 (fair) or 1 (poor). Finally, the average score for each video is calculated. This number is known as the mean opinion score (MOS).

> THE TESTS ARE DESIGNED TO MAKE ANALYSIS AS ACCURATE AND STRAIGHTFORWARD AS POSSIBLE

The video sequences are typically produced in a lab environment so that all types of impairments can be included. High-quality sports, nature and news videos are used as a starting point. Impairments (codec settings, rate and resolution changes, initial buffering, stalling and so on) are then emulated by varying the bitrate over a certain range, for example, or placing stalling events of different lengths at various points in the videos.

The tests are designed to make analysis as accurate and straightforward as possible. Devising subjective tests is a time consuming and expensive process, though, and lab tests can't assess exactly what an MNO's real users are experiencing. The best way to overcome these challenges is by using objective quality algorithms.

Terms and abbreviations
APN – Access Point Name | **AVC** – advanced video coding | **DASH** – Dynamic Adaptive Streaming over HTTP | **DM** – device management | **eNB** – eNodeB (LTE base station) | **ETSI** – European Telecommunications Standards Institute | **HD** – high definition | **HEVC** – High Efficiency Video Coding | **HLS** – HTTP Live Streaming | **HTTP** – Hypertext Transfer Protocol | **ITU-T** – International Telecommunication Union Telecommunication Standardization Sector | **MNO** – mobile network operator | **MOS** – mean opinion score | **MPD** – media presentation description | **MSP** – media service provider | **OMA** – Open Mobile Alliance | **Pa** – short-term audio predictor | **Pq** – long-term quality predictor | **Pv** – short-term video predictor | **QoE** – quality of experience | **RRC** – Radio Resource Control | **SMS** – short message service | **UHD** – ultra high definition | **VOD** – video on demand | **VP9** – An open-source video format and codec

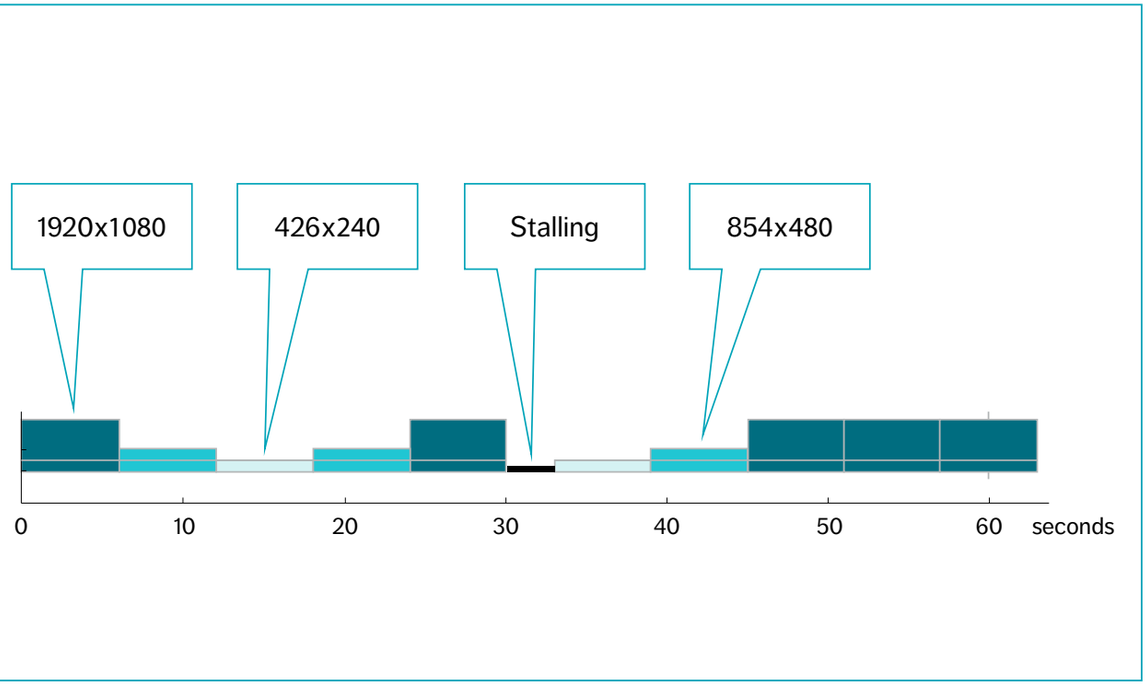


Figure 1 A 60-second video with quality variations (resolution) and a three-second stall in the middle

> IDEALLY, DIFFERENT LEVELS OF COMPLEXITY SHOULD BE USABLE BY A SINGLE MODEL

Objective quality

As the wording suggests, objective quality algorithms (also known as objective models) are designed to mimic the behavior and perception of humans. The goal is to produce the same scores as the MOS values that would result from running a subjective test on the same videos.

Many different types of objective models can be adopted, depending on the intended usage and the kind of input data employed. Those using the most limited set of input parameters base the objective quality estimation on encoding rates, video resolution, frame rates, codecs and stalling information, as these factors provide the minimum amount of information about the video playout that is required to estimate a quality score. More complex models might use the complete encoded video bitstream, or even the full received video signal, to further increase the estimation accuracy.

The objective models described above are no-reference models, where input is taken only from the receiving end of the media distribution chain. Full-reference models can also be adopted, where the video originally transmitted is compared with the one that is received. Another variant is the reduced-reference model, where the original video is not needed for reference, but certain information about it is made available to the model.

Traditionally, objective models are used to evaluate quality based on relatively short video sequences: approximately 10 seconds long. However, with adaptive video streaming, where quality can vary significantly during a given session, the model must also assess how this long-term variation affects user perception. To do this, model evaluation of much longer video sequences (up to several minutes) is required.

Ideally, different levels of complexity should be usable by a single model – that is, from only a few input parameters up to the full bitstream, depending on deployment. This is the scope that applies with the new no-reference ITU-T P.1203 standard.

Standardization of ITU-T P.1203

The P.1203 standard [4] was developed as part of an ITU-T competition between participating proponents (Ericsson and six other companies), where each one sent in its proposed quality models. The models that performed best were then used as a baseline to create the final standard. An internal model architecture was also defined to facilitate the creation of a model that would be as flexible as possible.

Architecture

The standard includes modules for estimating short-term audio and video quality, and an integration module estimating the final session quality due to adaptation and stalling, as shown in ***Figure 2***.

The short-term video and audio predictor (Pv and Pa) modules continuously estimate the short-term audio and video quality scores for one-second pieces of content. This means that for a 60-second video, there will be 60 audio scores and 60 video scores. These Pv and Pa modules are specific to each type of codec.

The Pv and Pa modules operate in up to four different modes, depending on how detailed the input is from the parameter extraction. For the least complex mode, the main inputs are related to resolution, bitrate and frame rate, while the most complex mode performs advanced analysis of the video payload.

The short-term scores from those modules are fed into the long-term quality predictor (Pq) module, together with any stalling information, and the final session quality score for the total video session is then estimated. The Pq module also produces a number of diagnostic outputs, so that the underlying causes of the score can be analyzed. The Pq module is not mode- or codec-dependent and is therefore common for all cases.

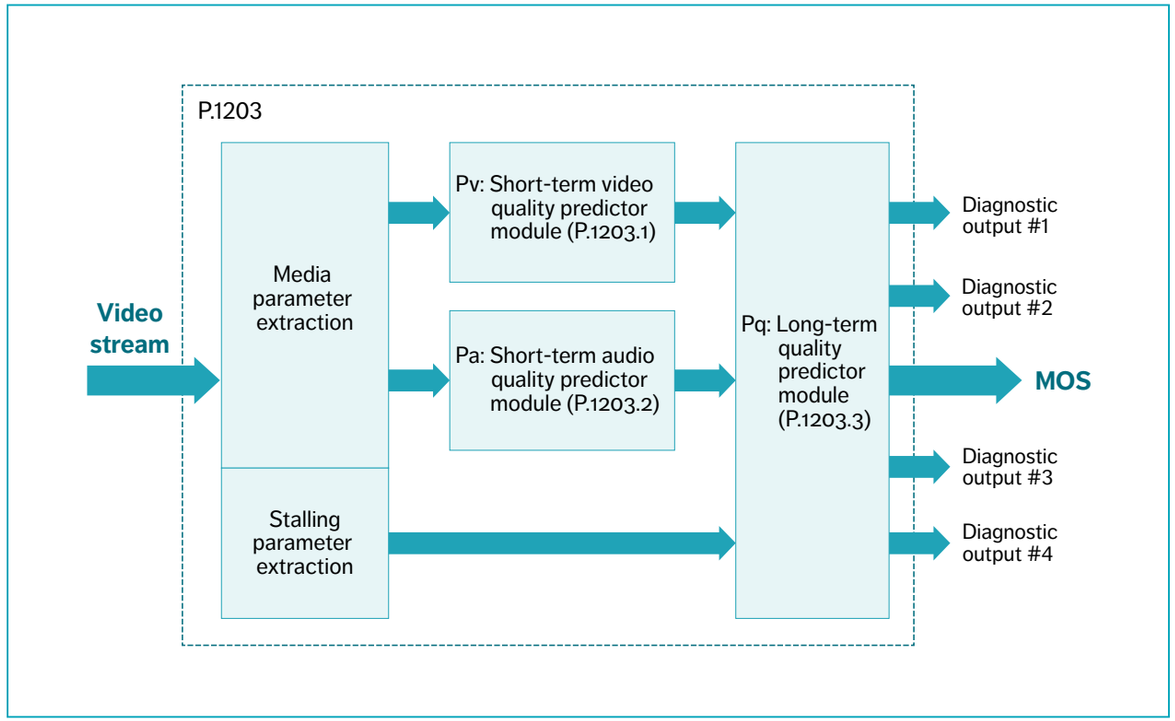


Figure 2 ITU-T P.1203 architecture

Training

In the development and standardization of P.1203, a large number of subjective test databases were created, each containing videos that were graded by at least 24 test individuals. An important goal set in the development of P.1203 was to handle the long-term perceived effects of stalling events and quality adaptations.

Thus, in contrast to traditional subjective tests in which a few 10-second videos are typically repeated, the videos used for P.1203 were all unique and between one and five minutes long. It was important to avoid repetition and continuously present new test videos that the viewers found interesting enough to pay attention to throughout the playout. The tests were done on mobile devices as well as on computer monitors and TV screens, to cover all the different use cases.

Validation

Since the development of the P.1203 standardization was run as a competition, the performance of the different proponent models required validation, and the validation could not be carried out on any existing databases. Instead, after the submission of all of the proponents' candidate models to ITU-T, all the proponents worked together to design a new set of databases, with each step distributed to different proponents so that none had complete control over any individual database.

The new set of validation databases were then used to evaluate the models, and the top-performing ones were selected to form the P.1203 standard. In cases where the respective performance of several proponent modules was so close that a statistical test could not tell them apart, the modules were merged to form a single standard without alternative implementations.

Final model

While the Pv and Pa modules were developed using traditional analytical methods and implemented as a series of mathematical functions, the Pq module is more advanced. This module is divided into two separate estimation algorithms: one using a traditional functional approach and the other based on machine-learning concepts.

The functional variant models human perception, as influenced by the effect of quality oscillations, deep quality dips, repeated quality or stalling artifacts and the ability to memorize. All of these effects are described by mathematical functions (as they are for the Pv and Pa modules), which are then combined to estimate the user's total perception of quality.

Machine learning is a method that solves a problem with the support of self-learning computer algorithms. It is well suited for problems where the relationship between the input and the output is complex, as in the Pq module. During the design and training phase, the algorithms automatically identify how various characteristics of the input data (Pv/Pa scores and stalling parameters) are reflected in changes to the output data (the test panel MOS values). The algorithm then automatically builds a black-box algorithm, which implements the final machine-trained solution and estimates the user score.

The final Pq MOS estimate is a weighted average of the output from the traditional functional algorithm and the machine-learning-based one. One of the advantages of using two different Pq algorithms is that they have statistically independent estimation errors, and when the two scores are averaged, the actual error becomes smaller.

Future standardization work

The video module currently supports H.264/AVC video codecs up to HD (1920x1080). A new work item has been started in ITU-T, which will result in a recommendation that also supports H.265/HEVC and VP9 video codecs up to UHD resolution (3840x2160). This work item is running in a similar fashion to P.1203, with a competition giving participating companies the opportunity to submit their own proposed models.

> THE USEFULNESS OF HAVING A STANDARDIZED CLIENT FEEDBACK MECHANISM HAS BEEN RECOGNIZED BY 3GPP

Implementing a quality model

Successful implementation of a quality model is dependent on access to the input data required by the model itself. The most demanding models, such as full-reference variants, are usually implemented close to the video streaming client, inside the device, so that the complete received video can be compared with the one sent. This is typically done for manual testing scenarios, where a special test phone is used in which a quality model has been implemented.

This method is not feasible for passive measurements, where all or a large part of live video traffic is monitored, so model input data needs to be collected in another way. For example, an MNO that wants to gain a better understanding of its overall perceived video streaming quality would need to collect data from all streaming sessions. One way of doing this would be to intercept the traffic at certain network nodes and use the traffic content and pattern to try to infer which service is being used and the quality level being delivered. This can be difficult, however, owing to the fact that many services are now encrypted, which significantly limits access to the data required to do a detailed quality estimation.

One way to overcome this challenge is with the help of the streaming client, which has full knowledge of what is happening during the video session. A feedback link from the streaming client can be used to report selected metrics to the network (or the original streaming server) where the quality can be estimated. This technique is already used internally for many streaming services. For example, when a user clicks on a video link on the internet, the client typically continuously measures different metrics and sends them to the server. Unfortunately for MNOs, though, these feedback channels are usually encrypted and available only to the MSP. Even if an MSP were to make the metrics available to the MNO, they would still be proprietary, and it would be difficult for the MNO to compare them due to the fact that the content and level of detail typically differ between MSPs.

3GPP QoE reporting

The usefulness of having a standardized client feedback mechanism has been recognized by 3GPP, and its technical specification TS 26.247 describes how this can be implemented for a DASH streaming client [5]. The 3GPP concept is called QoE reporting, as the metrics collected and reported are related specifically to the quality of the session. Sensitive or integrity-related data such as the user's position and the content viewed cannot be reported.

The basic concept is that the streaming client can receive a QoE configuration that specifies the metrics to be collected, how often collection will occur, when reporting will be done and which entity to report to. There are three ways to send a QoE configuration to the client to facilitate different deployment cases:

1. Media presentation description

In this case, the client downloads a media presentation description (MPD) when streaming starts. The MPD specifies how the media is structured and how the client can access and download the media chunks. The MPD can also contain a QoE configuration that makes it possible to get feedback from the client. Since the MPD is usually controlled by the content or service provider, the QoE reports from the client are typically configured to go back to their servers and are not always visible to the MNO.

2. Open Mobile Alliance Device Management

Open Mobile Alliance Device Management (OMA-DM) has defined methods for how an MNO can configure certain aspects of connected devices such as APNs, SMS servers and so on. These methods also include an optional QoE

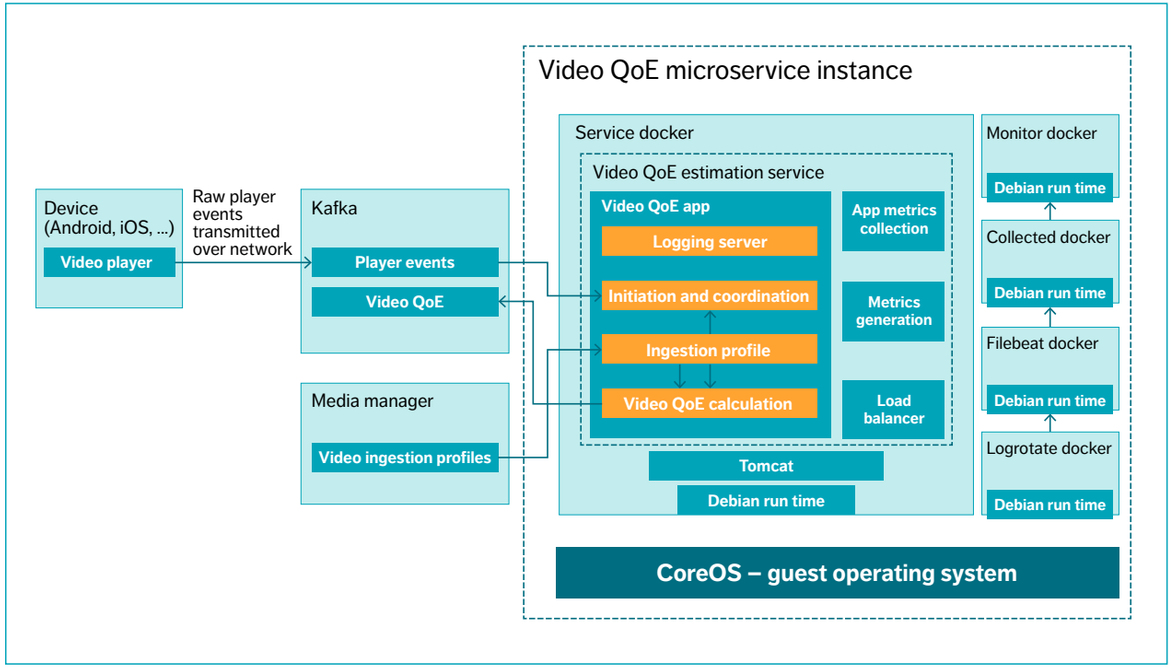


Figure 3 Example of a video QoE microservice

configuration that activates QoE reporting from the client. However, not all MNOs deploy OMA-DM in their networks.

3. Radio Resource Control

The Radio Resource Control (RRC) protocol [6] is used between the eNB and the mobile device to control the communication in the RAN. The possibility of including a QoE configuration was added in 3GPP rel-14, giving MNOs the ability to use RRC to activate QoE reporting. As a result, the QoE configuration can be handled like many other types of RAN-related configurations and measurements.

The QoE metrics reported by the client in each of these three methods are well aligned with the input requirements in P.1203. This means that, in principle, any of them would enable an MNO to gain a good overview of the video streaming quality experienced by its users. Before this can happen, though, the new standards will need to be deployed both in the network and in the clients.

Deployment of video QoE estimation

The delivery of over-the-top video today most commonly uses a cloud-based microservices platform with inbuilt video-quality estimation features. The P.1203 algorithm for estimating video quality is most suitable for implementation as a microservice that estimates video quality (in terms of MOS) for all video streams and for all individual video streaming sessions. If the estimated video quality distribution shows that there is a quality issue, a root cause analysis can be carried out and the necessary measures can be taken to improve quality.

When running a media service, and providing the video client as part of the service, the input parameters to the P.1203 algorithm are taken directly from the video client that has all the details about the playout of the stream and reported over the network to an analytics backend. *Figure 3* outlines an example of such an implementation, in which player events are reported to a stream processing system (Kafka) where the video QoE is calculated in a video QoE microservice. The output from the video QoE calculation is then posted back into the stream processing system to be used for monitoring, visualization, root cause analysis and other purposes.

> THE INPUT PARAMETERS TO THE P.1203 ALGORITHM ARE TAKEN DIRECTLY FROM THE VIDEO CLIENT

For an MNO, the standard architecture is to install a probe inside the network – in the core network, for example. The probe monitors video traffic using shallow or deep packet inspection and gathers information that can be used as input to a QoE estimation algorithm (P.1203). If a video service is not encrypted, the relevant metrics and events from the video streams can often be measured or estimated. The task becomes more challenging if the video streams are encrypted, but quality estimations of these video streams can still be done to some extent by using a combination of probes and standardized and proprietary algorithms and models. However, the ability to report quality-related metrics direct from the video clients enables much more accurate estimation of quality.

Conclusion

MNOs and MSPs stand to gain a great deal from developing a better understanding of how users experience video quality, and the ITU-T P.1203 and 3GPP TS 26.247 standards provides the framework that is necessary to help them do so. Implementation of these standards by MNOs, MSPs and device manufacturers will enable the efficient and accurate estimation of video QoE required to meet continuously rising user expectations. The standard's smart handling of the effects of stalling events and quality adaptations makes it well suited to overcome the challenges presented by VOD and by adaptive video streaming in particular.

THE AUTHORS

Gunnar Heikkilä
◆ is a senior specialist in machine intelligence at Ericsson Research. Since 1996, he has focused on user experience quality assessment and measurements, including standardization in 3GPP, ETSI and ITU-T. He joined Ericsson in 1987 and has previously worked with control system software for military defense radar systems, and with software design for synchronous digital hierarchy optical fiber transmission systems. Heikkilä holds an M.Sc. in computer science from Luleå University of Technology, Sweden.

Jörgen Gustafsson
◆ is a research manager at Ericsson Research, heading a research team in the areas of machine learning and QoE. The research is applied to a number of areas, such as media, operations support systems/business support systems, the Internet of Things and more. He joined Ericsson in 1993. He is co-rapporteur of Question 14 in ITU-T Study Group 12, where leading and global standards on parametric models and tools for multimedia quality assessment are being developed, including the latest standards on quality assessment of adaptive streaming. He holds an M.Sc. in computer science from Linköping University, Sweden.

References:

1. **Ericsson ConsumerLab report, TV and Media 2016, available at:** *https://www.ericsson.com/networked-society/trends-and-insights/consumerlab/consumer-insights/reports/tv-and-media-2016*
2. **Ericsson Mobility Report, November 2016, available at:** *https://www.ericsson.com/en/mobility-report*
3. **ITU-T P.910, April 2008, Subjective video quality assessment methods for multimedia applications, available at:** *http://www.itu.int/itu-t/recommendations/rec.aspx?rec=9317*
4. **ITU-T P.1203, November 2016, Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport, available at:** *http://www.itu.int/itu-t/recommendations/rec.aspx?rec=13158*
5. **3GPP TS 26.247, January 2015, Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH), available at:** *http://www.3gpp.org/DynaReport/26247.htm*
6. **3GPP TS 25.331, January 2016, Radio Resource Control (RRC); Protocol specification, available at:** *http://www.3gpp.org/DynaReport/25331.htm*

Designing for the future

THE 5G NR PHYSICAL LAYER

Future networks will have to provide broadband access wherever needed and support a diverse range of services including everything from robotic surgery to virtual reality classrooms and self-driving cars. 5G New Radio is designed to fit these requirements, with physical layer components that are flexible, ultra-lean and forward-compatible.

ALI A. ZAIDI, ROBERT BALDEMAIR, MATTIAS ANDERSSON, SEBASTIAN FAXÉR, VICENT MOLÉS-CASES, ZHAO WANG

Far more than an evolution of mobile broadband, 5G wireless access will be a key IoT enabler, empowering people and industries to achieve new heights in terms of efficiency and innovation. A recent survey performed across eight different industries (automotive, finance, utilities, public safety, health care, media, internet and manufacturing) revealed that 89 percent of respondents expect 5G to be a game changer in their industry [1].

5G wireless access is being developed with three broad use case families in mind: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC) [2, 3]. eMBB focuses on across-the-board enhancements to the data rate, latency, user density, capacity and coverage of mobile broadband access. mMTC is designed to enable communication between devices that are low-cost, massive in number and battery-driven, intended to support applications such as smart metering, logistics, and field and body sensors. Finally, URLLC will make it possible for devices and machines to communicate with ultra-reliability, very low latency and high availability, making it ideal for vehicular communication, industrial control, factory automation, remote surgery, smart grids and public safety applications.

To meet the complex and sometimes contradictory requirements of these diverse use cases, 5G will encompass both an evolution of today's 4G (LTE) networks and the addition of a new, globally standardized radio access technology known as New Radio (NR).

Change to 5G New Radio

5G NR will operate in the frequency range from below 1GHz to 100GHz with different deployments. There will typically be more coverage per base station (macro sites) at lower carrier frequencies, and a limited coverage area per base station (micro and pico sites) at higher carrier frequencies. To provide high service quality and optimal reliability, licensed spectrum will continue to be the backbone of the wireless network in 5G, and transmission in unlicensed spectrum will be used as a complement to provide even higher data rates and boost capacity. The overall vision for 5G in terms of use cases, operating frequencies and deployments is shown in ***Figure 1***.

The standardization of NR started in 3GPP in April 2016, with the aim of making it commercially available before 2020. 3GPP is taking a phased approach to defining the 5G specifications. A first standardization phase with limited NR functionality will be completed by 2018, followed by a second standardization phase that fulfills all the requirements of IMT-2020 (the next generation of mobile communication systems to be specified by ITU-R) by 2019. It is likely that NR will continue to evolve beyond 2020, with a sequence of releases including additional features and functionalities. Although NR does not have to be backward compatible with LTE, the future evolution of NR should be backward compatible with its initial release(s). Since NR must support a wide range of use cases – many of which are not yet defined – forward compatibility is of utmost importance.

> THE NR PHYSICAL LAYER HAS A FLEXIBLE AND SCALABLE DESIGN TO SUPPORT DIVERSE USE CASES

NR physical layer design

A physical layer forms the backbone of any wireless technology. The NR physical layer has a flexible and scalable design to support diverse use cases with extreme (and sometimes contradictory) requirements, as well as a wide range of frequencies and deployment options.

Terms and abbreviations
BPSK – binary phase shift keying | **BS** – base station | **CDM** – code division multiplexing | **CPE** – common phase error | **CP-OFDM** – cyclic prefix orthogonal frequency division multiplexing | **CSI-RS** – channel-state information reference signal | **D2D** – device-to-device | **DFT-SOFDM** – discrete Fourier transform spread orthogonal frequency division multiplexing | **DL** – downlink | **DMRS** – demodulation reference signal | **eMBB** – enhanced mobile broadband | **eMBMS** – evolved multimedia broadcast multicast service | **FDM** – frequency division multiplexing | **HARQ** – hybrid automatic repeat request | **IMT-2020** – the next generation mobile communication systems to be specified by ITU-R | **IoT** – Internet of Things | **ITU-R** – International Telecommunication Union Radiocommunication Sector | **LBT** – listen-before-talk | **LDPC** – low-density parity-check | **MBB** – mobile broadband | **MIMO** – multiple-input, multiple-output | **mMTC** – massive machine-type communications | **mmWave** – millimeter wave | **MU-MIMO** – multi-user MIMO | **NR** – New Radio | **PRB** – physical resource block | **PTRS** – phase-tracking reference signal | **QAM** – quadrature amplitude modulation | **QPSK** – quadrature phase shift keying | **SRS** – sounding reference signal | **TDM** – time division multiplexing | **UE** – user equipment | **UL** – uplink | **URLLC** – ultra-reliable low-latency communications

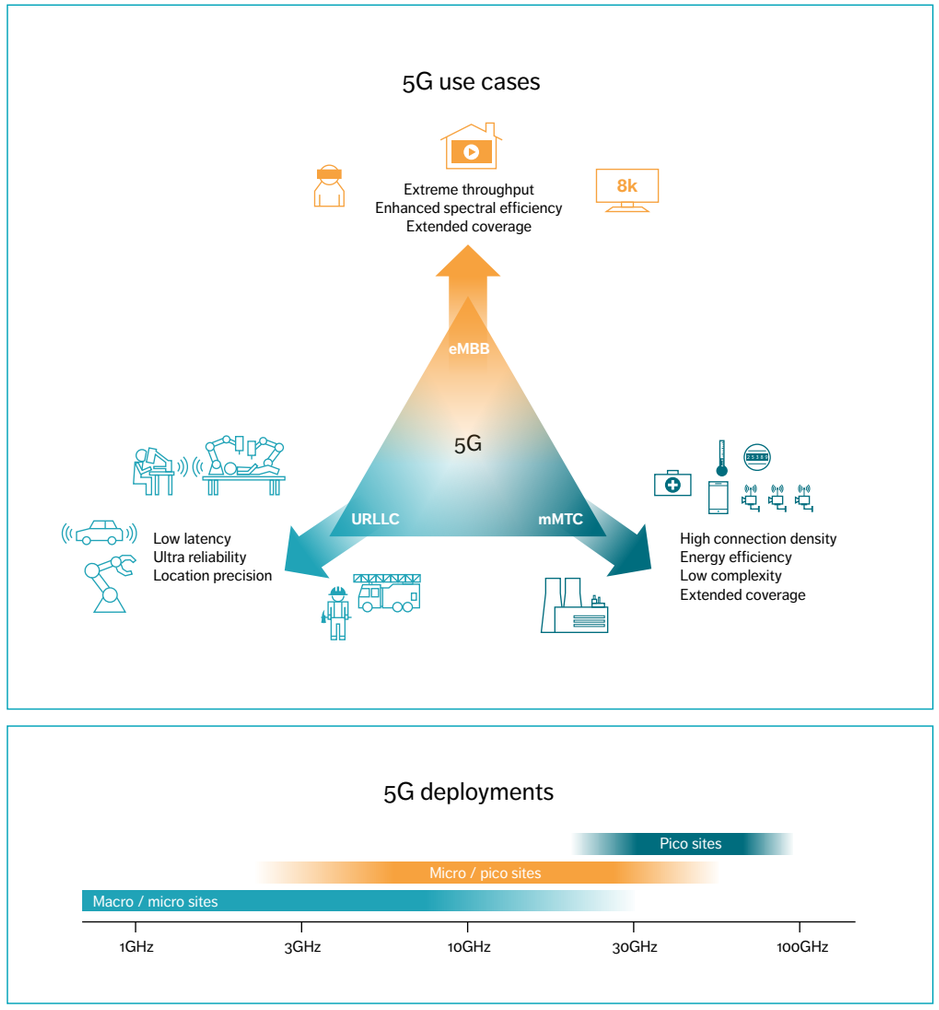


Figure 1 5G vision: use cases, spectrum and deployments

The key technology components of the NR physical layer are modulation schemes, waveform, frame structure, reference signals, multi-antenna transmission and channel coding.

Modulation schemes

LTE supports the QPSK, 16QAM, 64QAM and 256QAM modulation formats, and all of these will also be supported by NR. In addition, 3GPP has included $\pi/2$-BPSK in UL to enable a further reduced peak-to-average power ratio and enhanced power-amplifier efficiency at lower data rates, which is important for mMTC services, for example. Since NR will cover a wide range of use cases, it is likely that the set of supported modulation schemes may expand. For example, 1024QAM may become part of the NR specification, since fixed point-to-point backhaul already uses modulation orders higher than 256QAM. Different modulation schemes for different UE categories may also be included in the NR specification.

Waveform

3GPP has agreed to adopt CP-OFDM with a scalable numerology (subcarrier spacing, cyclic prefix) in both UL and DL up to at least 52.6GHz. Having the same waveform in both directions simplifies the overall design, especially with respect to wireless backhauling and device-to-device (D2D) communications. Additionally, there is support for DFT-Spread OFDM in UL for coverage-limited scenarios, with single stream transmissions (that is, without spatial multiplexing). Any operation that is transparent to a receiver can be applied on top of CP-OFDM at the transmitter side, such as windowing/filtering to improve spectrum confinement.

A scalable OFDM numerology is required to enable diverse services on a wide range of frequencies and deployments. The subcarrier spacing is scalable according to 15×2^n kHz, where n is an integer and 15kHz is the subcarrier spacing used in LTE. The scaling factor 2^n ensures that slots and symbols of different numerologies are aligned in the time domain, which is important to efficiently enable TDD networks [4]. The details related to NR OFDM numerologies are shown in *Figure 2*. The choice of parameter n depends on various factors including type of deployment, carrier frequency, service requirements (latency, reliability and throughput), hardware impairments (oscillator phase noise), mobility and implementation complexity [5]. For example, wider subcarrier spacing can be promising for latency-critical services (URLLC), small coverage areas and higher carrier frequencies. Narrower subcarrier spacing can be utilized for lower carrier frequencies, large coverage areas, narrowband devices and evolved multimedia broadcast multicast services (eMBMSs). It may also be possible to support multiple services simultaneously with different requirements on the same carrier by multiplexing two different numerologies (wider subcarrier spacing for URLLC and lower subcarrier spacing for MBB/mMTC/eMBMS, for example).

> “HAVING THE SAME WAVEFORM IN BOTH DIRECTIONS SIMPLIFIES THE OVERALL DESIGN”

The spectrum of OFDM signal decays rather slowly outside the transmission bandwidth. In order to limit out-of-band emission, the spectrum utilization for LTE is 90 percent. That is, 100 of the 111 possible physical resource blocks (PRBs) are utilized in a 20MHz bandwidth allocation. For NR, it has been agreed that the spectrum utilization will be greater than 90 percent. Windowing and filtering operations are viable ways to confine the OFDM signal in the frequency domain. It is important to note that the relationship between spectrum efficiency and spectrum confinement is not linear, since spectrum confinement techniques can induce self-interference.

Subcarrier spacing	15kHz	30kHz (2 x 15kHz)	60kHz (4 x 15kHz)	15 x 2^nkHz, (n = 3, 4, ...)
OFDM symbol duration	66.67 µs	33.33 µs	16.67 µs	66.67/2^n µs
Cyclic prefix duration	4.69 µs	2.34 µs	1.17 µs	4.69/2^n µs
OFDM symbol including CP	71.35 µs	35.68 µs	17.84 µs	71.35/2^n µs
Number of OFDM symbols per slot	7 or 14	7 or 14	7 or 14	14
Slot duration	500 µs or 1,000 µs	250 µs or 500 µs	125 µs or 250 µs	1,000/2^n µs

Figure 2 Scalable OFDM numerology for NR

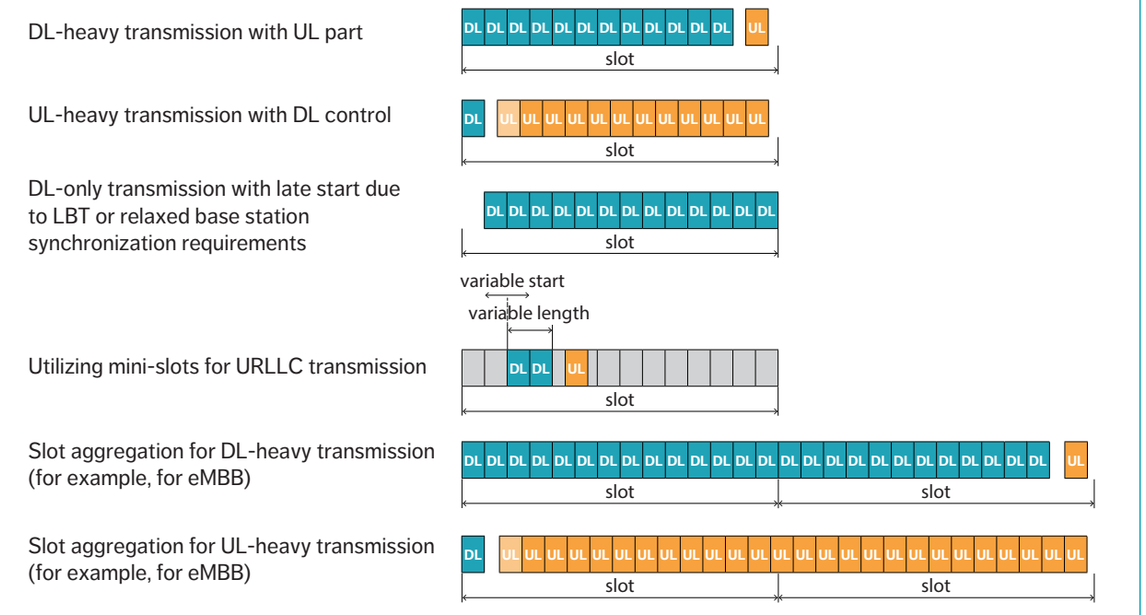


Figure 3 TDD-based frame structure examples for eMBB, URLLC and operation in unlicensed spectrum using listen-before-talk (LBT)

Frame structure

NR frame structure supports TDD and FDD transmissions and operation in both licensed and unlicensed spectrum. It enables very low latency, fast HARQ acknowledgements, dynamic TDD, coexistence with LTE and transmissions of variable length (for example, short duration for URLLC and long duration for eMBB). The frame structure follows three key design principles to enhance forward compatibility and reduce interactions between different features.

The first principle is that transmissions are self-contained. Data in a slot and in a beam is decodable on its own without dependency on other slots and beams. This implies that reference signals required for demodulation of data are included in a given slot and a given beam.

The second principle is that transmissions are well confined in time and frequency. Keeping transmissions together makes it easier to introduce new types of transmissions in parallel with legacy transmissions in the future. NR frame structure avoids the mapping of control channels across full system bandwidth.

The third principle is to avoid static and/or strict timing relations across slots and across different transmission directions. For example, asynchronous HARQ is used instead of predefined retransmission time.

As shown in Figure 2, a slot in NR comprises seven or 14 OFDM symbols for ≤ 60kHz numerologies and 14 OFDM symbols for ≥ 120kHz numerologies. A slot duration also scales with the chosen numerology since the OFDM symbol duration is inversely proportional to its subcarrier spacing. *Figure 3* provides examples for TDD, with guard periods for UL/DL switching.

A slot can be complemented by mini-slots to support transmissions with a flexible start position and a duration shorter than a regular slot duration. A mini-slot can be as short as one OFDM symbol and can start at any time. Mini-slots can be useful in various scenarios, including low-latency transmissions, transmissions in unlicensed spectrum and transmissions in the millimeter wave spectrum (mmWave band).

In low-latency scenarios, transmission needs to begin immediately without waiting for the start of a slot boundary (URLLC, for example). When transmitting in unlicensed spectrum, it is beneficial to start transmission immediately after LBT. When transmitting in mmWave band, the large amount of bandwidth available implies that the payload supported by a few OFDM symbols is large enough for many of the packets. Figure 3 provides examples of URLLC- and LBT-based transmission in unlicensed spectrum via mini-slots and illustrates that multiple slots can be aggregated for services that do not require extremely low latency (eMBB, for example). Having a longer transmission duration helps to increase coverage or reduce the overhead due to switching (in TDD), transmission of reference signals and control information.

The same frame structure can be used for FDD, by enabling simultaneous reception and transmission (that is, DL and UL can overlap in time). This frame structure is also applicable to D2D communications. In that case, the DL slot structure can be used by the device that is initiating (or scheduling) the transmission, and the UL slot structure can be used by the device responding to the transmission.

NR frame structure also allows for rapid HARQ acknowledgement, in which decoding is performed during the reception of DL data and the HARQ acknowledgement is prepared by the UE during the guard period, when switching from DL reception to UL transmission.

To obtain low latency, a slot (or a set of slots in case of slot aggregation) is front-loaded with control signals and reference signals at the beginning of the slot (or set of slots).

Reference signals

NR has an ultra-lean design that minimizes always-on transmissions to enhance network energy efficiency and ensure forward compatibility. In contrast to the setup in LTE, the reference signals in NR are transmitted only when necessary. The four main reference signals are the demodulation reference signal (DMRS), phase-tracking reference signal (PTRS), sounding reference signal (SRS) and channel-state information reference signal (CSI-RS).

“NR HAS AN ULTRA-LEAN DESIGN THAT MINIMIZES ALWAYS-ON TRANSMISSIONS”

DMRS is used to estimate the radio channel for demodulation. DMRS is UE-specific, can be beamformed, confined in a scheduled resource, and transmitted only when necessary, both in DL and UL. To support multiple-layer MIMO transmission, multiple orthogonal DMRS ports can be scheduled, one for each layer. Orthogonality is achieved by FDM (comb structure) and TDM and CDM (with cyclic shift of the base sequence or orthogonal cover codes). The basic DMRS pattern is front loaded, as the DMRS design takes into account the early decoding requirement to support low-latency applications. For low-speed scenarios, DMRS uses low density in the time domain. However, for high-speed scenarios, the time density of DMRS is increased to track fast changes in the radio channel.

PTRS is introduced in NR to enable compensation of oscillator phase noise. Typically, phase noise increases as a function of oscillator carrier frequency. PTRS can therefore be utilized at high carrier frequencies (such as mmWave) to mitigate phase noise. One of the main degradations caused by phase noise in an OFDM signal is an identical phase rotation of all the subcarriers, known as common phase error (CPE). PTRS is designed so that it has low density in the frequency domain and high density in the time domain, since the phase rotation produced by CPE is identical for all subcarriers within an OFDM symbol, but there is low correlation of phase noise across OFDM symbols. PTRS is UE-specific, confined in a scheduled resource and can be beamformed. The number of PTRS ports can be lower than the total number of ports, and orthogonality between PTRS ports is achieved by means of FDM. PTRS is configurable depending on the quality of the oscillators, carrier frequency, OFDM subcarrier spacing, and modulation and coding schemes used for transmission.

The SRS is transmitted in UL to perform CSI measurements mainly for scheduling and link adaptation. For NR, it is expected that the SRS will also be utilized for reciprocity-based precoder design for massive MIMO and UL beam management. It is likely that the SRS will have a modular and flexible design to support different procedures and UE capabilities. The approach for CSI-RS is similar.

Multi-antenna transmissions

NR will employ different antenna solutions and techniques depending on which part of the spectrum is used for its operation. For lower frequencies, a low to moderate number of active antennas (up to around 32 transmitter chains) is assumed and FDD operation is common. In this case, the acquisition of CSI requires transmission of CSI-RS in the DL and CSI reporting in the UL. The limited bandwidths available in this frequency region require high spectral efficiency enabled by multi-user MIMO (MU-MIMO) and higher order spatial multiplexing, which is achieved via higher resolution CSI reporting compared with LTE.

For higher frequencies, a larger number of antennas can be employed in a given aperture, which increases the capability for beamforming and MU-MIMO. Here, the spectrum allocations are of TDD type and reciprocity-based operation is assumed. In this case, high-resolution CSI in the form of explicit channel estimations is acquired by UL channel sounding. Such high-resolution CSI enables sophisticated precoding algorithms to be employed at the BS. This makes it possible to increase multi-user interference suppression, for example, but might require additional UE feedback of inter-cell interference or calibration information if perfect reciprocity cannot be assumed.

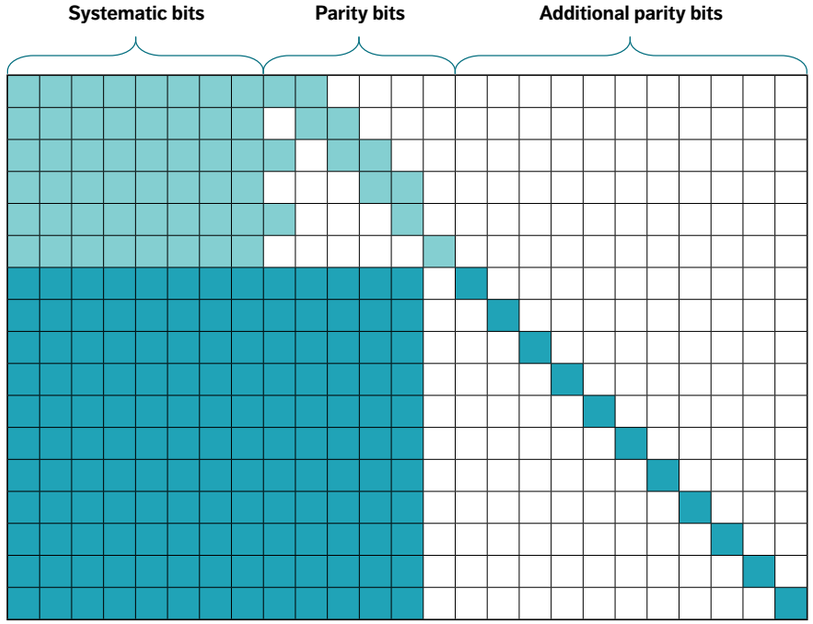


Figure 4 Structure of NR LDPC matrices

For even higher frequencies (in the mmWave range) an analog beamforming implementation is typically required currently, which limits the transmission to a single beam direction per time unit and radio chain. Since an isotropic antenna element is very small in this frequency region owing to the short carrier wavelength, a great number of antenna elements is required to maintain coverage. Beamforming needs to be applied at both the transmitter and receiver ends to combat the increased path loss, even for control channel transmission. A new type of beam management process for CSI acquisition is required, in which the BS needs to sweep radio transmitter beam candidates sequentially in time, and the UE needs to maintain a proper radio receiver beam to enable reception of the selected transmitter beam.

To support these diverse use cases, NR features a highly flexible but unified CSI framework, in which there is reduced coupling between CSI measurement, CSI reporting and the actual DL transmission in NR compared with LTE. The CSI framework can be seen as a toolbox, where different CSI reporting settings and CSI-RS resource settings for channel and interference measurements can be mixed and matched so they correspond to the antenna deployment and transmission scheme in use, and where CSI reports on different beams can be dynamically triggered. The framework also supports more advanced schemes such as multi-point transmission and coordination. The control and data transmissions, in turn, follow the self-contained principle, where all information required to decode the transmission (such as accompanying DMRS) is contained within the transmission itself. As a result, the network can seamlessly change the transmission point or beam as the UE moves in the network.

Channel coding

NR employs low-density parity-check (LDPC) codes for the data channel and polar codes for the control channel. LDPC codes are defined by their parity-check matrices, with each column representing a coded bit, and each row representing a parity-check equation. LDPC codes are decoded by exchanging messages between variables and parity checks in an iterative manner. The LDPC codes proposed for NR use a quasi-cyclic structure, where the parity-check matrix is defined by a smaller base matrix. Each entry of the base matrix represents either a ZxZ zero matrix or a shifted ZxZ identity matrix.

> THE FRAMEWORK ALSO SUPPORTS MORE ADVANCED SCHEMES SUCH AS MULTI-POINT TRANSMISSION AND COORDINATION

Unlike the LDPC codes implemented in other wireless technologies, the LDPC codes considered for NR use a rate-compatible structure, as shown in *Figure 4*. The light blue part (top left) of the base matrix defines a high rate code, at a rate of either 2/3 or 8/9. Additional parity bits can be generated by extending the base matrix and including the rows and columns marked in dark blue (bottom left). This allows for transmission at lower code rates, or for generation of additional parity bits such as those used for HARQ operation using incremental redundancy similar to LTE. Since the parity-check matrix for higher code rates is smaller, decoding latency and complexity decreases for high code rates. Along with the high degree of parallelism achievable through the quasi-cyclic structure, this allows for very high peak throughputs and low latencies. Further, the parity-check matrix can be extended to lower rates than the LTE turbo codes, which rely on repetition for code rates below 1/3. This allows the LDPC codes to achieve higher coding gains also at low coding rates, making them suitable for use cases requiring high reliability.

Polar codes will be used for layer 1 and layer 2 control signaling, except for very short messages.

Polar codes are a relatively recent invention, introduced by Arıkan in 2008 [6]. They are the first class of codes shown to achieve the Shannon capacity with reasonable decoding complexity for a wide variety of channels.

By concatenating the polar code with an outer code and keeping track of the most likely values of previously decoded bits at the decoder (the list), good performance is achieved at shorter block lengths, like those typically used for layer 1 and layer 2 control signaling. By using a larger list size, error correction performance improves, at the cost of higher complexity at the decoder.

Conclusion

Flexibility, ultra-lean design and forward compatibility are the pillars on which all the 5G NR physical layer technology components (modulation schemes, waveform, frame structure, reference signals, multi-antenna transmission and channel coding) are being designed and built. The high level of flexibility and scalability in 5G NR will enable it to meet the requirements of diverse use cases, including a wide range of carrier frequencies and deployment options. Its built-in forward compatibility will ensure that 5G NR can easily evolve to support any unforeseen requirements.

References:

1. **Ericsson Survey Report, 2016, Opportunities in 5G: the view from eight industries, available at:** *https://app-eu.clickdimensions.com/blob/ericssoncom-ar0ma/files/5g_industry_survey_report_final.pdf*

2. **The 5G Infrastructure Public Private Partnership (5G-PPP) Technical Report, February 2016, 5G empowering vertical industries, available at:** *https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf*

3. **3GPP Technical Report TR38.913, ver. 14.2.0, Release 14, March 2017, Study on scenarios and requirements for next generation access technologies, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2996*

4. **IEEE Communications Magazine, vol. 54, no. 11, pp. 90-98, A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. D. Silva, November 2016, Waveform and numerology to support 5G services and requirements, available at:** *http://ieeexplore.ieee.org/document/7744816/*

5. **IEEE Communications Standard Magazine (submitted), A. A. Zaidi, R. Baldemair, V. Molés-Cases, N. He, K. Werner, and A. Cedergren, OFDM Numerology Design for 5G New Radio to Support IoT, eMBB, and MBSFN**

6. **IEEE Transactions on Information Theory, vol. 55, pp. 3051-3073, E. Arıkan, July 2009, Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels, available at:** *http://ieeexplore.ieee.org/document/5075875/*

THE AUTHORS

Ali A. Zaidi
◆ joined Ericsson in 2014, where he is currently a senior researcher working with concept development and standardization of radio access technologies (NR and LTE-Advanced Pro). His research focuses on mmWave communications, indoor positioning, device-to-device communications, and systems for intelligent

transportation and networked control. He holds an M.Sc. and a Ph.D. in telecommunications from KTH Royal Institute of Technology in Stockholm, Sweden. Zaidi is currently serving as a member of the Technology Intelligence Group Radio and a member of the Young Advisory Board at Ericsson Research.

Robert Baldemair
◆ is a master researcher who joined Ericsson in 2000. He spent several years working with research and development of radio access technologies for LTE before shifting his focus to wireless access for 5G. He holds a Dipl. Ing. and a Ph.D. from

the Technische Universität Wien in Vienna, Austria. He received the Ericsson Inventor of the Year award in 2010, and in 2014 he and a group of colleagues were nominated for the European Inventor Award.

Mattias Andersson
◆ is an experienced researcher who joined

Ericsson in 2014. His work focuses on concept development and standardization of low-latency communications, carrier aggregation and channel coding for LTE and NR. He holds both an M.Sc. in engineering physics and a Ph.D. in telecommunications from KTH Royal Institute of Technology in Stockholm, Sweden.

Sebastian Faxér
◆ is a researcher at Ericsson Research. He received an M.Sc. in applied physics

and electrical engineering from Linköping University, Sweden, in 2014 and joined Ericsson the same year. Since then he has worked on concept development and standardization of multi-antenna technologies for LTE and NR.

Vicent Molés-Cases
◆ is a researcher at Ericsson Research. He received an M.Sc. in telecommunication

engineering from the Polytechnic University of Valencia in Spain in 2016 and joined Ericsson the same year. Since then he has worked on concept development and 3GPP standardization of reference signals for NR.

Zhao Wang
◆ is a researcher at Ericsson Research who joined the company in 2015. He holds a Ph.D. in telecommunications from KTH Royal Institute of Technology in Stockholm, Sweden. His work currently

focuses on reference signal design of NR for 3GPP standardizations.

5G network programmability

FOR MISSION-CRITICAL APPLICATIONS

5G will make it possible for mobile network operators to support enterprises in a wide range of industry segments by providing cellular connectivity to mission-critical applications. The ability to expose policy control to enterprise verticals will create new business opportunities for mobile network operators by enabling a new value chain through the integration of telecom with other industries.

RAFIA INAM, ATHANASIOS KARAPANTELAKIS, LEONID MOKRUSHIN, ELENA FERSMAN

A key differentiator of 5G systems from previous generations will be a higher degree of programmability. Instead of a one-size-fits-all mobile broadband service, 5G will provide the flexibility to tailor QoS to connectivity services to meet the demands of enterprise customers. This enables a new range of mission-critical use cases, such as those involving connected cars, manufacturing robots, remote surgery equipment, precision agriculture equipment, and so on.

Network programmability can support rapid deployment of new use cases by combining cloud-based services with mobile network infrastructure and taking advantage of new levels of flexibility. Further, network programmability will enable a greater number of enterprise customers to use such services, and consumers will benefit from a unique and personalized experience.

A number of use cases in mission-critical scenarios can benefit from QoS programmability because a cellular network's connectivity requirements – including latency, throughput, service lifetime and cost – vary widely across different use cases. To support them all, we have developed an application programming interface (API) that allows third parties to specify and request network QoS. We have also demonstrated the usefulness of this API on a test mobile network using a transport-related use case.

As part of this use case, we have been collaborating with commercial vehicle manufacturer Scania to develop the QoS requirements for teleoperation. Teleoperation is the remote operation of an autonomous vehicle by a human operator in cases where the vehicle encounters a situation that the autonomous system cannot overcome by itself (a road obstacle or malfunction, for example).

Drivers of network programmability

The key drivers behind the creation of a programmable network are the need to accelerate time to market, and the desire to reduce operational costs and take advantage of the business opportunities presented by a new mission-critical service market. In a programmable network, traditional network functions requiring specialized hardware are replaced with software functions hosted on commercial off-the-shelf infrastructure. Technologies such as software-defined networking and network functions virtualization are essential to cutting operational and capital costs in mobile networks.

Cloud-based services and applications are enablers for programmability. Service provisioning in the cloud and managed access to the provisioned services and applications are important. This requires collaboration between telecom and other industries (IT application and content providers, and automotive original equipment manufacturers, for example). One way to simplify and accelerate the deployment of services and applications from industry verticals is the automatic translation of industrial requirements to service requirements, and then on to resource-level requirements (in other words, network requirements). Network slicing provides a dedicated, virtualized mobile network containing a set of network resources, and provides guaranteed QoS. The network slices are not only beneficial but also critical to support many applications in vertical industries.

New network communication services can also be provisioned programmatically; that is, by using a software service orchestration function instead of manual provisioning by engineers. As orchestration will also be used for provisioning connectivity services to mission-critical applications, mobile networks need to support QoS programmability.

Mission-critical Intelligent Transportation System use cases

5G will support a diverse range of use cases in different industry sectors, each putting its own QoS requirements on the mobile network [1]. It is possible to use network programmability to realize mission-critical use cases with QoS requirements by creating highly specialized services tailored to industrial needs and preferences.

Features like lower latency (reaction times that are five times faster), higher throughput (10 to 100 times higher data rate) and an enormous increase in the number of connected devices (10 to 100 times more) can support the large-scale use of massive machine-type communication (mMTC) and mission-critical MTC (MC-MTC) use cases for the first time. Further, a dedicated network slice would meet the specific requirements of each use case.

In mMTC use cases, a large number of sensors and actuators are connected using a short-range radio (capillary network) to a base station (eNodeB) using a low protocol overhead to save the battery life of the devices.

This requires a network slice with broad coverage, small data volumes from massive numbers of devices. MC-MTC use cases emphasize lower latency (down to a level of milliseconds),

Terms and abbreviations
AAR – Authentication Authorization Request | **AF** – application function | **API** – application programming interface **eNB** – eNodeB | **EPC** – Evolved Packet Core | **EPS** – Evolved Packet System | **HSS** – Home Subscriber Server | **ITS** – Intelligent Transportation System | **MME** – Mobility Management Entity | **mMTC** – massive machine-type communication | **MTC** – machine-type communication | **PCRF** – policy and charging rules function | **PDN** – packet data network | **PGW** – PDN gateway | **QCI** – QoS class identifier | **Rx** – radio receiver | **SAPC** – Service-Aware Policy Controller | **SGW** – service gateway | **UDP** – User Datagram Protocol | **UE** – user equipment

> WE ENVISION REALIZING THESE USE CASES WITH A FLEXIBLE NETWORK PROGRAMMABILITY TECHNIQUE

robust transmission and multilevel diversity due to their mission-critical nature and, consequently, need a network slice of very low latency, high reliability and availability (packet loss down to 10-9). This is possible by creating a slice of very high priority. We envision realizing these use cases with a flexible network programmability technique.

The current focus of our research is within the Intelligent Transportation System (ITS) domain and includes a few 5G use cases in mMTC and MC-MTC, including transportation and logistics, autonomous cars and teleoperated vehicles.

Transportation and logistics

The lower latency and high throughput of 5G will support multiple use cases related to connected cars, transportation and retail logistics that consist of fleets of connected/driverless vehicles transporting people and goods. The key network requirements for mission-critical automotive driving are high throughput and low latency up to 100ms. Failure is not an option in these cases. There are also many potential sub-use cases. For example, a journey from A to B in a driverless vehicle could involve vehicle-to-vehicle connections, connections between vehicles and street infrastructure for traffic management, and high-speed reliable connectivity to support cloud applications.

Autonomous vehicles

The vision of fully autonomous vehicles aims to reduce the risks associated with human error. A system to achieve this vision would need to connect the cars and the road infrastructure with 1ms latency in all areas (100 percent coverage). Unfortunately, 1ms latency is currently not possible in mobile networks. However, the bandwidth requirements to make this possible are not excessive, as only vehicle control data needs to be communicated. This capability is expected in 5G.

Teleoperation of vehicles

The ability to control a self-driving vehicle from a distance is an important use case that is needed in public transportation when an onboard, autonomous system faces a difficult situation, such as a traffic accident, an unexpected demonstration, unscheduled roadworks or flooding. These scenarios require the planning of an alternate route, and an operator needs to drive the vehicle remotely for a short time. Another case could be a mechanical malfunction or an injury on a bus that requires remote intervention to mitigate the risk of danger to others. Network requirements for remote monitoring and control include broad coverage, high data throughput and low latency to enable continuous video streaming and the ability to send commands between a remote operations center and a vehicle [2].

Why an API?

To guarantee QoS for the three ITS cases described above (and mission-critical use cases in general) mobile network operators typically go through manual network planning and configurations. Examples include configuring manual data routes via different routers, configuring Differentiated Services and allocating dedicated spectrum ranges to each use case. However, doing this is costly because it requires the configuration and deployment of network equipment. Nor is it particularly feasible, as this kind of configuration deployment cannot be done merely in parts of the transport network (such as backhaul). If, on the other hand, the resources were virtualized and there was software that could set up these routes over the same physical network link, both of the limiting factors would be eliminated: the cost of configuration and the deployment of multiple routes. As a result, it would be both financially and technically feasible to support these use cases concurrently.

It is clear that operators will not be able to support the volume and diversity of use cases with the current network management approach. They need a different means of managing the network to stay competitive. In our view, developing an API is the logical first step toward exposing a programmable network to the industry verticals. This approach will result in a solution that is more responsive than rigid commercial offerings, such as preconfigured subscription packages.

Architecture of the Ericsson-Scania project

Teleoperating a bus requires data from sensors on the bus, including a video feed from a camera at the front that is streamed to a remote operations center over LTE radio access with an evolved 5G core network. The commands to drive the bus are sent from the center to the bus using Scania's command system.

Figure 1 illustrates the data streams that need to be prioritized to meet QoS demands: sensor data, the video feed originating from the vehicle user equipment (UE) and the commands to remotely drive the bus. Sending these data streams over low-priority data traffic (like infotainment) is a critical requirement. We used QOS class identifier (QCI) bearers, as detailed in the corresponding 3GPP standard, to enforce this prioritization. We assigned QCI class 5 and 2 to video and sensor data respectively and lowest-priority QCI class 9 to infotainment. In our lab environment, we have confirmed that the high-priority streams (QCI 2 and 5) can be kept regardless of the amount of low-priority background data traffic in the network [3, 4]. The next step will be to test our testbed setup for the prioritized video stream in the presence of the network load due to infotainment-type background traffic.

How it works

A cloud-hosted application function (AF) dynamically sets up virtual connections between vehicles and the 5G Evolved Packet Core (EPC) network, with specific QoS attributes, such as designated latency levels and guaranteed throughput. *Figure 2* illustrates the architecture of the system on which we have implemented the API. In addition to deploying a standard EPC and LTE band-40 RAN, an AF is deployed on an OpenStack-managed cloud. This application functionality allows third parties to set up QoS for their UEs through an API.

The AF consists of the following components: a knowledge base module, an API endpoint module and a transformer.

Knowledge base module

This module maps domain-specific concepts to the generic concepts. The knowledge base is implemented as a graph database, has a schema of general concepts and can be extended with additional domain concept documents that instantiate the general concept schema. The schema includes a basic vocabulary of general concepts that model QoS requests. These concepts can be instantiated in domain concepts for a specific enterprise. In our case, the enterprise is automotive.

Within the knowledge base module, an "agent" is a string that is semantically related to the mobile device for whic h QoS is requested. In our case, the agent is instantiated with the "vehicle" domain concept. QoS class identifiers (QCIs) are indicators of network QoS for a given agent. The QCI concept was introduced in 3GPP TS 23.203 Release 8, with additional classes being introduced in Release 12 and Release 14.

Every QCI class has an integer identifier, for example QCI1 or QCI2, and is mapped to a set of QoS metrics such as an indicator of priority of data traffic, an upper ceiling for network latency and, in some cases, guaranteed bit rate. In our case, QCIs are instantiated with domain concepts for real-time vehicle traffic domain concepts. For example, QCI3 is instantiated as "vehicle_control_traffic" and QCI4 is instantiated as "vehicle_video_traffic." For users browsing their mobile devices in the vehicles, we instantiate a low-priority class QCI9 as vehicle_web_browsing."

Data traffic descriptors are generic contents that can configure each QCI. The configuration

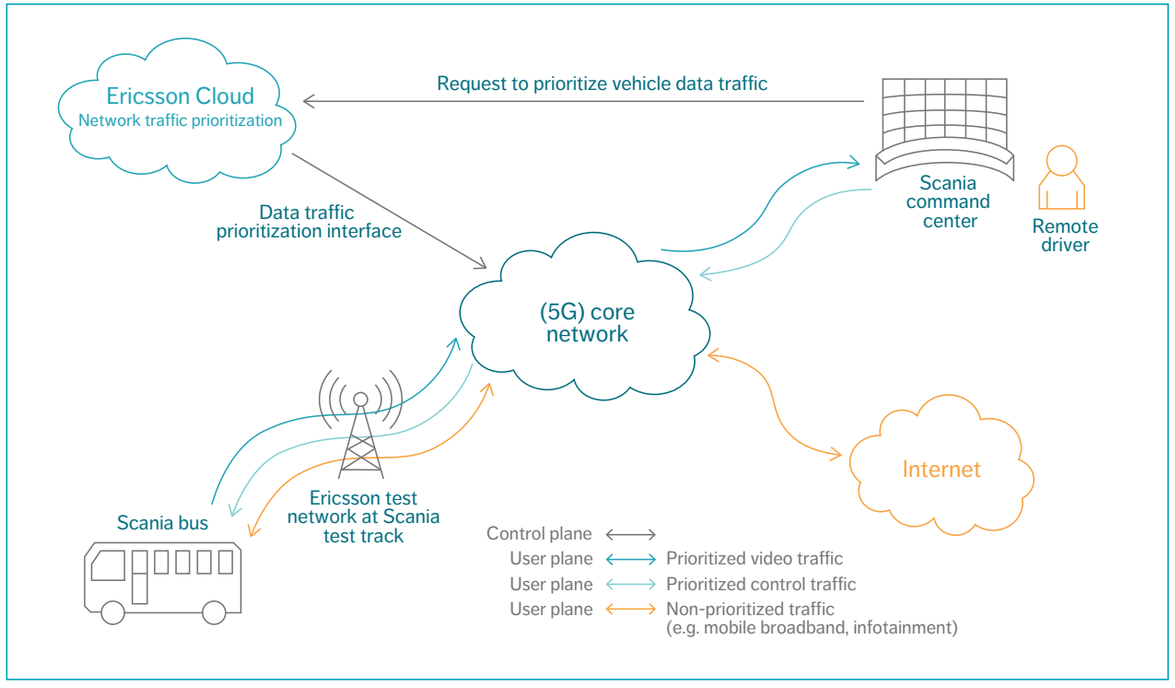


Figure 1 Prioritized data streams to meet QoS demands

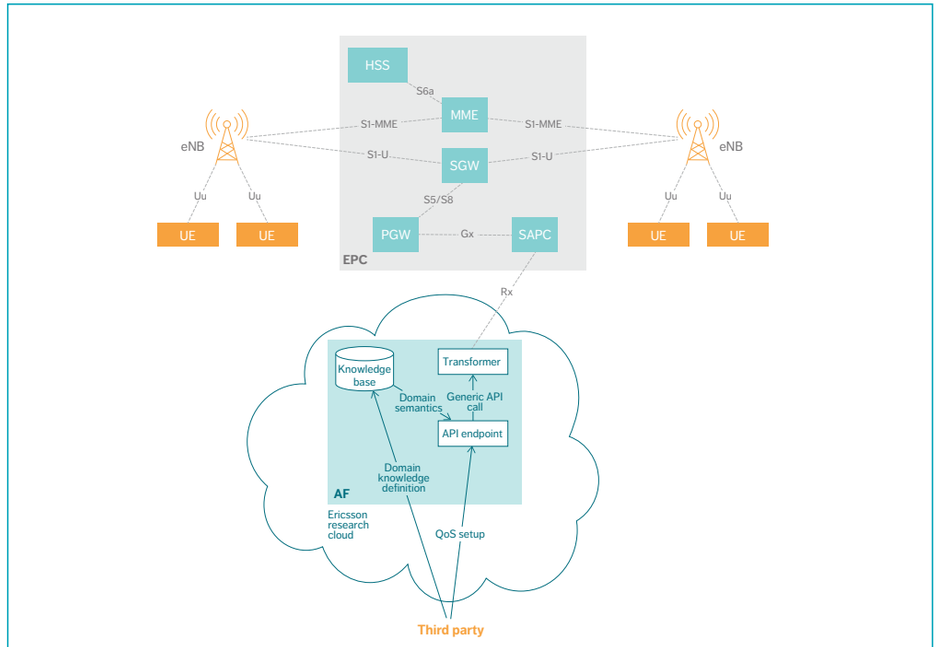


Figure 2 Architecture for programmable QoS in existing an LTE EPC network

pertains to a characterization of the traffic in terms of required throughput for both "uplink" and "downlink," the former being data traffic transmitted from the agent and the latter being the opposite. Optionally, descriptors may also contain the type of data packets exchanged (for example, UDP/IP or TCP/IP), as well as potentially the port or port range. For example, in the case of "vehicle_control_traffic," the data traffic descriptor identifies an uplink bandwidth of 1Mbps and a downlink bandwidth of 1Kbps.

A combination of agents, QCIs and their associated data traffic descriptors are stored in the knowledge base as a domain concept document. Every use case has its own domain concept document, while each specific enterprise has more than one document. For example, in our case, there is an "automotive/teleoperation" document. However, other documents for automotive can also exist, such as "automotive/autonomous drive" or "automotive/remote fleet management." Because the data is stored as linked data, concepts from one domain concept document can be reused in another.

API endpoint module

This module composes API specifications from every domain concept document in the knowledge base. This API specification is RESTful, uses symmetric encryption (HTTPS) and can be called from any third party. These API calls get translated into generic concept calls that are subsequently sent to the transformer module. Note that, in addition to an API call for setup of specialized QoS, there is another API call for teardown of this QoS. For example, when a vehicle is decommissioned or does not need to be teleoperated, there can be a call to tear down the QoS tunnel, so network resources can be allocated to UEs in other vehicles or devices. *Figure 3* provides an overview of domain-specific and generic requests.

Transformer module

The transformer module translates generic requests for QoS to Rx AAR requests, as these requests are specified in 3GPP TS 29.214. The Rx requests are sent directly to the PCRF node in order to set up the "EPS bearer" (in other words, the data tunnel with the requested QoS). As is the case with the endpoint module, the transformer module can also translate a teardown request to an Rx request to revert to the lowest-priority default bearer (in most cases, QCI9).

> CONCEPTS FROM ONE DOMAIN CONCEPT DOCUMENT CAN BE REUSED IN ANOTHER

Domain specific request	Generic request	Description of the request
GET /vehicle	GET /agent	Retrieve QoS information for all UEs
GET /vehicle/<IP>	GET /agent/<IP>	Retrieve QoS information for one UE, based on its IP address
GET /qos	GET /qos	Retrieve QoS for all UEs
GET /qos/vehicle_control_traffic	GET /qos/QCI3	Retrieve all UEs with QCI3 bearer setup

POST /qos/
{
“source_IP”:<src_IP>,
“source_port”:<src_port>,
“destination_IP”:<dst_IP>,
“destination_port”:<dst_port>,
“qos_class”:”vehicle_control_traffic”,
“protocol”:”TCP”,
“type”:”vehicle_video_stream”
}
	POST /qos
	{
	“source_IP”:<src_IP>,
	“source_port”:<src_port>,
	“destination_IP”:<dst_IP>,
	“destination_port”:<dst_port>
	“qos_class”:”QCI3”,
	“protocol”:”TCP”,
	“max-requested-bandwidth-UL”:”1024,
	“max-requested-bandwidth-DL”:”100”
	}
		Set QoS for UE with IP src_IP and port src_port toward destination with IP dst_IP and port dst_port. Protocol in this example is TCP but it can also be UDP.

DELETE /vehicle
{
“source_IP”:<src_IP>,
“source_port”:<src_port>,
“destination_IP”:<dst_IP>,
“destination_port”:<dst_port>,
“protocol”:”TCP”
}
	DELETE /agent
	{
	“source_IP”:<src_IP>,
	“source_port”:<src_port>,
	“destination_IP”:<dst_IP>,
	“destination_port”:<dst_port>,
	“protocol”:”TCP”
	}
		Remove QoS for UE with given source and destination IP and port

Figure 3 Overview of requests

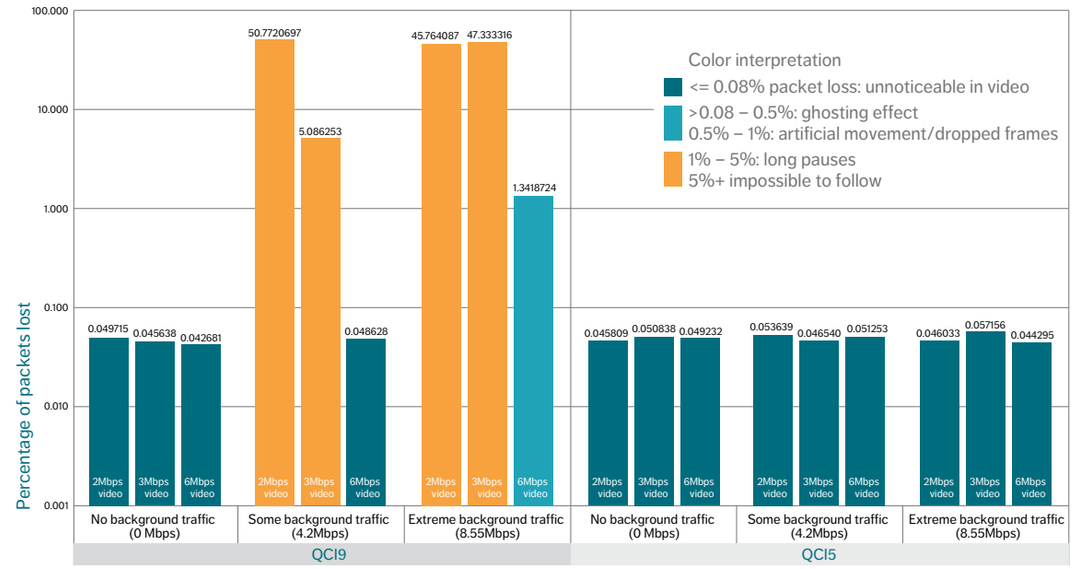


Figure 4 Packet loss (in percentage of total packets) in best effort (QCIŁ9) and prioritized (QCI5) bearers

Testbed results

To assess QoS, we performed experiments on the uplink prioritized video stream using QCI5 in the presence of the network load due to infotainment-type background traffic using QCI9. The total measured available bandwidth on the network was approximately 8.55Mbps. We tested several network load scenarios and measured the results against three background traffic conditions:

- none (0Mbps)
- some (4.2Mbps or 49 percent of the available bandwidth)
- extreme (8.55Mbps or 100 percent of the available bandwidth)

We measured both throughput and one-way network delay under these traffic conditions. We also measured the ratio of packets lost versus packets sent to test the throughput quality of the network for three different qualities of video streams:

- excellent (6Mbps or 70 percent of the available bandwidth)
- good (3Mbps or 35 percent of the available bandwidth)
- borderline drivable (2Mbps or 23 percent of the available bandwidth)

Borderline drivable is the minimum requirement to perform teleoperation. We obtained the packet

drop requirements from empirical observations during test driving. We took a total of 160 measurements for each experiment and plotted the graphs based on the respective average value.

Measurements from the 5G-network testbed show that resource prioritization can assure predefined QoS levels for mission-critical applications, regardless of background traffic. *Figure 4* illustrates guaranteed uplink packet loss for a critical application, in which the acceptable packet loss of less than or equal to 0.08 percent is unnoticeable in the video stream. This is true even with extreme background traffic when the system is congested – the critical traffic is still served with no performance degradation.

However, as Figure 4 also illustrates, for the non-prioritized infotainment traffic (QCI9) the packet loss increases heavily with the increase in the background traffic, introducing long pauses and making teleoperation impossible even for a lower level of congestion.

When we measured the uplink delay, we found that it is preserved (remaining at less than 34ms) for the critical video traffic even when the system exhibits congestion. For the non-prioritized traffic, the delay reaches up to 600ms during congestion.

The next step is to develop the concept for a self-service portal where network customers could specify QoS requirements on their own terms; for example, to prioritize 4K video traffic for 40 buses in an urban scenario. The software would then translate this specification into instructions for network resource prioritization.

Conclusion

5G attributes such as network slicing and low latency will soon make mission-critical use cases such as safe, autonomous public transport a reality. Automated network resource prioritization via a programmable API can support network QoS for diverse use cases with different connectivity requirements on the cellular network. By developing an API that allows a third party to request network resources and implementing it on a test mobile network, we have demonstrated how the technology works in an urban transport-related use case with Scania. The initial results show that throughput and latency are maintained for high-priority streams regardless of the network load. ✪

Further reading

- **YouTube, Remote bus driving over 5G, November 2016:** *https://www.youtube.com/watch?v=lPyzGTD5FtM*
- **Ericsson Research blog, 5G teleoperated vehicles for future public transport, June 8, 2017, Berggren, V; Fersman, E; Inam, R; Karapantelakis, A; Mokrushin, L; Schrammar, N; Vulgarakis, A; Wang, K:** *https://www.ericsson.com/research-blog/5g/5g-teleoperated-vehicles-future-public-transport/*
- **Ericsson Mobility Report, June 2017:** *https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-june-2017.pdf*

THE AUTHORS

The authors would like to acknowledge the work of Keven (Qi) Wang on this project during his stay at Ericsson.

Rafia Inam
◆ joined Ericsson Research in 2015. She works as a senior researcher in the area of machine intelligence and automation, and her research interests include 5G management, service modeling, virtualization of resources, reusability of real-time software and ITSs. Inam received her M.S. from Chalmers University of Technology in Gothenburg, Sweden, in 2010. She received her Licentiate and doctoral degrees from Mälardalen University in Västerås, Sweden in 2012

and 2014 respectively. Her paper, "Towards automated service-oriented lifecycle management for 5G networks", won a best paper award in 2015.

Athanasios Karapantelakis
◆ joined Ericsson in 2007 and currently works as a master research engineer in the area of machine intelligence and automation. He holds a B.Sc. in computer science from the University of Crete in Greece and an M.Sc. and Licentiate of Engineering in communication systems from KTH Royal Institute of Technology in Stockholm, Sweden. His background is in software engineering.

Leonid Mokrushin
◆ is a senior specialist in the area of cognitive technologies. His current focus is on investigating what technological opportunities artificial intelligence may bring to Ericsson by creating and prototyping innovative concepts in the context of new industrial and telco use cases. He joined Ericsson Research in 2007 after postgraduate studies at Uppsala University in Sweden. He holds an M.Sc. in software engineering from Peter the Great St.

Petersburg Polytechnic University.

Elena Fersman
◆ is head of machine intelligence and automation at Ericsson Research and an adjunct professor in cyber-physical systems at KTH Royal Institute of Technology in Stockholm. She holds a Ph.D. in computer science from Uppsala University in Sweden and did post-doctoral research at École normale supérieure Paris-Saclay, France, before starting her industrial career. Her current research interests are in the areas of modeling and analysis and of software- and knowledge-intensive intelligent systems applied to 5G and IoT.

References

1. **IEEE, International Conference on Intelligent Transportation Systems (November 2016) – Feasibility Assessment to Realise Vehicle Teleoperation using Cellular Networks – Rafia Inam, Nicolas Schrammar, Keven Wang, Athanasios Karapantelakis, Leonid Mokrushin, Aneta Vulgarakis Feljan and Elena Fersman, available at:** *http://ieeexplore.ieee.org/document/7795920/*
2. **IEEE, International Conference on Future Internet of Things and Cloud (August 2016) – DevOps for IoT Applications Using Cellular Networks and Cloud – Athanasios Karapantelakis, Hongxin Liang, Keven Wang, Konstantinos Vandikas, Rafia Inam, Elena Fersman, Ignacio Mulas-Viela, Nicolas Seyvet and Vasileios Giannokostas, available at:** *http://ieeexplore.ieee.org/document/7575883/*
3. **Ericsson Mobility Report, Improving Public Transport with 5G, November 2015, available at:** *https://www.ericsson.com/res/docs/2015/mobility-report/emr-nov-2015-improving-public-transport-with-5g.pdf*
4. **IEEE, Conference on Emerging Technologies and Factory Automation (September 2015) – Towards automated service-oriented lifecycle management for 5G networks (Best Paper) – Rafia Inam, Athanasios Karapantelakis, Konstantinos Vandikas, Leonid Mokrushin, Aneta Vulgarakis Feljan, and Elena Fersman, available at:** *http://ieeexplore.ieee.org/document/7301660/*

Industrial automation enabled by

ROBOTICS, MACHINE INTELLIGENCE AND 5G

The emergent "fourth industrial revolution" will have a profound impact on both industry and society in the years ahead. Robotics, machine intelligence and 5G networks in particular will play major roles in this revolution by enabling ever higher levels of automation for production processes.

ROBERTO SABELLA (ERICSSON), ANDREAS THUELIG (ERICSSON), MARIA CHIARA CARROZZA (SANT'ANNA SCHOOL OF ADVANCED STUDIES), MASSIMO IPPOLITO (COMAU)

The combination of robotics, machine intelligence and 5G networks will provide a wealth of opportunities for cooperation between robots and humans that can improve productivity and speed up the delivery of services for citizens.

Most analysts agree that smart manufacturing is likely to represent the biggest portion of market revenues for the Internet of Things (IoT) in the near future. Smart manufacturing is dependent on industrial automation, which relies heavily on the use of robots and machine intelligence. The factory of the future will be realized through the digitization of the manufacturing process and plants, which will be enabled by 5G networks and all their building blocks. As a leader in 5G infrastructure – including cloud technologies, big data analytics and IT capabilities – Ericsson is well placed to take a leading role in this transformation and partner with industries to develop solutions that are tailored to fit their needs. Our fruitful collaboration with Comau, a world leader in industrial automation, and the Sant'Anna School of Advanced Studies, a highly regarded academic center of excellence in robotics, is the first step.

Understanding Industry 4.0

The German government launched the Industry 4.0 initiative in 2011 to foster the competitiveness of its industries for the decades to come. Seven years later, we can see that leading industries are implementing breathtakingly innovative concepts as a result. For example, the 2017 Hannover Industry Fair showcased several innovations such as collaborative robots, smart materials, adaptive production, and self-learning systems, which were on their way out of labs and onto real production shop floors. German industry is committed to making Industry 4.0 the new benchmark in production efficiency, with plans to invest EUR 40 billion annually until 2020.

The strength of the Industry 4.0 initiative lies in the fact that it has been a joint effort of government, universities and industries since day one. Together they foster growth, using fewer resources, reducing risk and boosting productivity and flexibility. The initiative has developed many groundbreaking concepts that have resulted in a quantum leap in the networking of humans, machines, robots and products. Manufacturing leaders are combining information technology and operations technology to create value in entirely new ways. These cyber-physical production lines are cutting-edge today, but will be the standard of tomorrow. By the end of 2017, Industry 4.0 had grown into a truly international initiative with contributors and users all over the globe, changing the basis of competition in production automation for good.

Guaranteed real-time communication between humans, robots, factory logistics and products is a fundamental prerequisite of the Industry 4.0 concept. Real-time data will generate transparency and actionable insights, while edge analytics will help reap maximum machine value and optimize production. All of the above concepts clearly require standardization and (data) security. With its standardized networking capabilities, built-in security, guaranteed grades of service, as well as distributed cloud and network slicing concepts, 5G is a perfect tool for advanced industries that want to take advantage of digital transformation.

Terms and abbreviations
IoT – Internet of Things | **OEM** – original equipment manufacturer | **PLC** – programmable logic controller | **PoC** – proof of concept

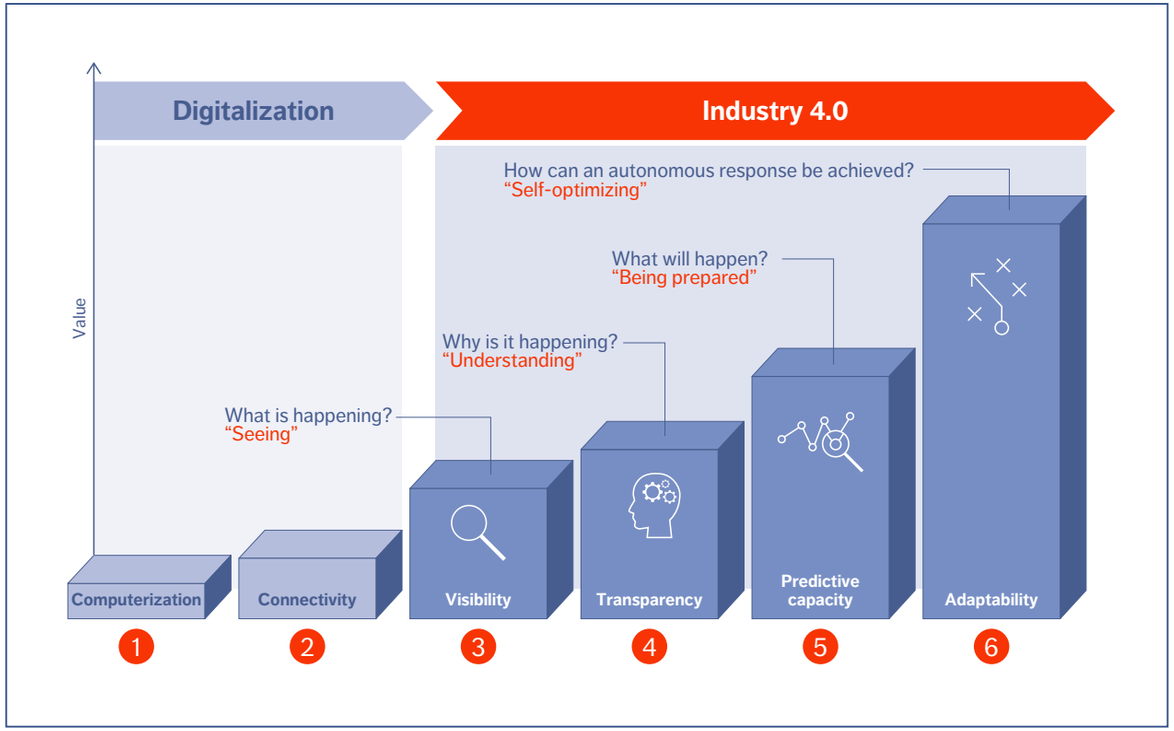


Figure 1: The Industry 4.0 maturity index

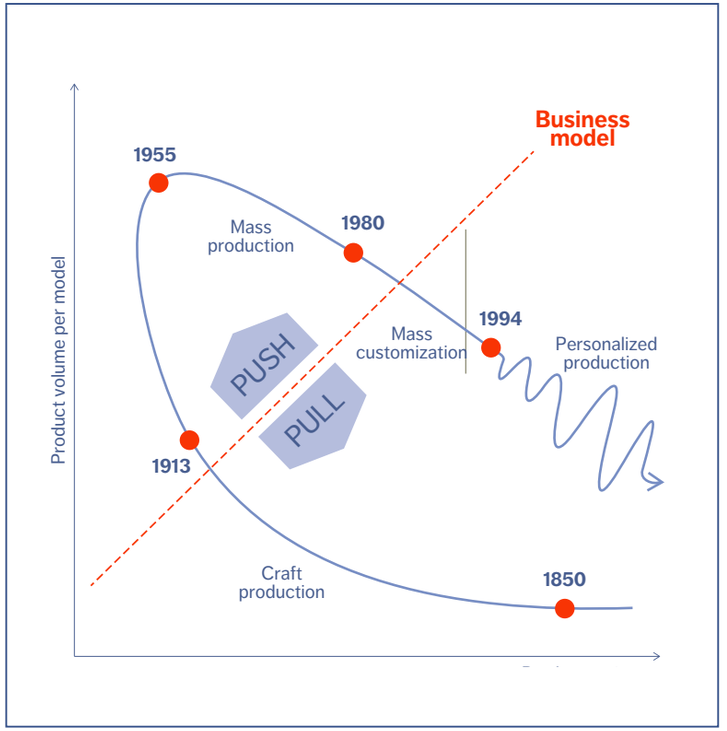


Figure 2: Evolution of the production paradigm toward Industry 4.0

Figure 1 shows the development path that allows industries to evolve step-by-step toward the full Industry 4.0 transformation process [1]. It consists of six stages with each building on the previous one. Each stage includes a description of the necessary capabilities and the consequent benefit for companies. Connectivity is the second stage, immediately after computerization, and it enables the ones that follow.

The transformation of manufacturing

The transformation of the manufacturing industry [2] from mass production to mass customization through digitized factory operations is illustrated in *Figure 2*. Although the industrial revolution at the end of the 18th century led to the advent of mechanization, industrial production remained at an artisanal level until the 20th century, when true mass production began with automotive manufacturing. Assembly-line production became a paradigm for mass production and had far-reaching impacts on society. In fact, what is called "Fordism" in social science describes an economic and social system based on industrialized, standardized mass production and mass consumption. The key concept is the manufacturing of standardized products in huge volumes, using special-purpose machinery and, at that time, unskilled labor. Although Fordism was a method used to improve productivity in the automotive industry, the principle applied to any kind of manufacturing process.

In recent decades, there has been an increasing need for customization to allow manufacturers to differentiate from competitors and broaden their product offerings. In a way, the product variety stimulates the consumer market, provided that manufacturing costs are kept low enough to have sustainable margins. The final step of the trend is "personalized production." Based on the Industry 4.0 paradigm and leveraging on new technologies across the complete value chain from suppliers to customers, it is possible to significantly increase the flexibility of the production line, and shorten production lead times. This leads to more affordable and scalable customization.

The trend for "personal" product customization is growing, along with a preference for online purchasing. Therefore, current processes need to be adapted to be more flexible and customizable, while still protecting initial investments in the production line. High speed wireless infrastructure such as 5G networks can facilitate the modification (required by customized products) of OEM machines with minimal impact.

The digitization of factory operations enabled by IoT technologies promises to make that happen. Digital tools will be able to monitor and control all tools of production, collecting data from thousands of sensors to create a digital image of the product

being realized, usually referred to as a "digital shadow." Once a digital shadow has been created for a physical product and bears its specific DNA, it is possible to manufacture that product more efficiently and with a higher degree of quality in the digitized production facility. In this way, it is possible to optimize the manufacturing process, detect quality issues early to prevent defects at the end of the production line and make continuous improvements. It is also possible to carry out predictive and preventive maintenance.

The combination of wireless sensors and high-capacity communication networks such as 4G and 5G plays a key role in this context, by enabling data collection from shop-floor level (production lines) and data transfer to cloud systems for continuous monitoring and control.

Virtual controllers that combine control, data logging and alarms into a cloud platform also help the process of digitization and save costs, panel space and maintenance activities compared to traditional systems. They can control a wide variety of production tools and are also a solution for remotely located machines and portable systems that can run standalone.

Key automation trends

Making good products is important for the success of a manufacturer, but it is not enough to be profitable and to sustain business. Production costs must be low enough for a suitable margin. This can be achieved by increasingly improving the efficiency of a manufacturing system. Automation is vital for that. Manufacturing systems require heavy investments and must be designed so that they remain profitable for the long term. If manufacturers are to remain competitive in an ever-changing marketplace, they must continuously improve both products and the production systems. Virtual commissioning is therefore necessary to continuously upgrade a production system with reasonable incremental investments. This requires a virtual (computer-based) environment that can simulate a manufacturing plant.

Virtual commissioning involves a virtual plant and a real controller. The simulated plant model has to be fully defined at the level of sensors and actuators. A major benefit of this is that it replaces the need for real commissioning with real plants and controllers, which is very expensive and time consuming. Instead, virtual commissioning allows for the identification of possible design defects and operational mistakes before investments are made in physical plant infrastructure. The digitization of the manufacturing plant allows its designers to enhance the efficiency of the production process, increase the automation density and optimize the handling of materials necessary to realize the products.

Digitization allows the introduction of the "just in sequence" concept, where components and parts arrive at a production line according to schedule, right in time for assembly. In addition, it can enable a truly lean enterprise, allowing for a much richer understanding of the customer demand and the immediate sharing of the demand data throughout complex supply chains and networks. Smart factories can produce at a faster rate with less waste. Industry 4.0 enables a much quicker flow of customized products. It has the potential to radically reduce inventories throughout the supply chain.

Finally, it is important to highlight the "zero-defect" concept. In some cases, relevant percentages of production can end up as scrap because of manufacturing defects. A "zero-defect" process requires automatic monitoring of the entire manufacturing process, from the quality of raw materials entering the production line to variances in tools and processes during each production run. As a closed-loop system, controllers are immediately alerted to any defects, and changes can be made immediately to eliminate the source of the problem. The approach has the potential to dramatically reduce scrap by detecting production errors instantly, eliminating the propagation of defects along the process stages. The manufacturing system could include knowledge-based loops, providing information and feedback to other levels of the manufacturing chain, to minimize failures via continuous optimization of the production process and the manufacturing system.

The factory of the future could consist of flexible production islands, able to realize different types of building blocks, without the rigidity of conveyors and with truly standard robotized working stations. As a result, agile shuttling robots are needed to transfer assembled blocks from one production island to another without the need for physical or virtual rails.

Digital factory elements and the role of 5G

The virtual plant concept makes it possible to carry out global system design, simulation, verification and physical mapping at a much lower cost than what is possible with a physical plant. To do so, however, the virtual plant requires new kinds of robots with the ability to increase the flexibility of the global production system. These robots need to be multipurpose and intelligent enough to adapt, communicate and interact with each other and with humans, based on a remote control that can globally manage a complete set of robot systems.

High-quality wireless connectivity is essential to the virtual plant concept. Wired connectivity, with its complex cabling, would not be feasible in this type of ever-changing environment due to the fact that cable upgrading entails high operational expenditure. The wireless connectivity must connect all physical elements of a production plant with machine (computing) elements that are able to collect and process huge amounts of data and/or with a cloud that is responsible for those operations. Communications among all these elements must work in a challenging environment characterized by electromagnetic interferences, and distributed over a large area that could span several buildings. While LTE connectivity is robust and capable enough to cope with that environment today, stringent latency requirements will soon demand 5G connectivity.

Once a huge amount of data has been collected through the wireless connectivity, it is necessary to use new methods to handle, process and transform it into a format that can be used by humans or machines or both, and to tap into the potential of cloud computing.

Big data and analytics systems are therefore essential in the digital factory. Eventually, a cyber-physical system will be needed to handle the complex production process, consisting of IT systems built around machines, storage systems and supplies. All the above leads to further efficiencies in the factory, by allowing preventive and predictive maintenance that reduces the time the plant is off-service, minimizing production delays and avoiding faults. One interesting effect of the efficiency boost is the reduced energy consumption. Process reengineering is something that will be done at a lower cost than today.

The challenge of connecting robots

A previous article in Ericsson Technology Review [3] explained the benefit of cloud robotics for various areas of industry (manufacturing, agriculture and transportation) in different logistics contexts such as harbors and hospitals. The rising level of intelligence in robots allows them to adapt to changing conditions, which is positive for development but significantly increases their complexity. Connecting robots and placing this complexity (intelligence) in a cloud will make it possible for affordable, minimal-infrastructure smart robot systems with unlimited computing capacity to evolve. As partners in the "5G for Italy" program, Ericsson, the Sant'Anna School of Advanced Studies, Comau and Zucchetti Centro Sistemi have been carrying out research together in these areas for the past three years.

The current industrial control systems architecture has provided a stable, secure, and robust platform for the past 25 years. But these legacy systems are reaching end of life in many industrial applications such as robotized cells, and ongoing maintenance, and updates are becoming complicated and expensive. In addition, legacy equipment was not built to ensure effective data access within these systems, which severely limits their potential for remote monitoring and control.

High-speed communication networks, wireless

Figure 3: Experimenting with cloud robotics in Comau's industrial automation lab in Turin, Italy

infrastructure and cloud computing technologies make it possible to enrich robotized plants with new relevant capabilities, while reducing costs through cloud technologies. As a result, it is possible to create smarter robots with "brains" (virtual controllers) in the cloud. The "brain" consists of a knowledge base, program path, models, communication support and so on, effectively transferring the intelligence of the controller into a remote virtual controller. This approach offers many benefits, including:

- on-premises cloud capability that can reside alongside legacy critical infrastructure, allowing for an evolution of legacy services and a platform for new services
- lower operating expenses as a result of maximizing the performance and capacity of a virtualization platform, which provides high reliability and performance
- fault tolerance to single and multiple software and hardware faults, with minimal loss of service
- comprehensive fault management, isolation and recovery
- high scalability and performance.

At the outset, 4G systems and Wi-Fi will be used to provide cloud robotics with the necessary connectivity, but 5G is the target technology truly capable of delivering the performance needed to support the applications of the future.

The proof of concept (PoC) that Ericsson, Comau and TIM (Telecom Italia S.p.A.) have realized together is a relevant example. Illustrated in *Figure 3*, the PoC consists of a working cell with two robots and a conveyor. The first robot is a manipulator that picks an object and places it in front of the second robot, which emulates soldering using a welding gun. The final object is then placed by the manipulator on the conveyor to send it to the next work cell. In this PoC, we have moved the control logic that drives the working station responsible for the concurrent actions of the two robots and the conveyor. It would normally reside in a control cabinet referred to as a station PLC (programmable logic controller), but in our PoC we have moved it to a cloud platform. Moving a relevant part of the control to the cloud enables the virtualization of those functionalities that can run as virtual machines on general purpose hardware.

We used an indoor LTE network installation to connect the elements of the working cell with the cloud server, which was possible because the LTE connectivity ensures a few tens of milliseconds of round trip latency. As shown in *Figure 4*, the next step will be to move the control functionalities that reside in each robot controller – the task planner, trajectory planner and inverse kinematics – to the cloud as well. That step, requiring a latency of the order of 5ms, requires 5G technology. The last step would be to move the control loop of the robot into the cloud as well.

This PoC represents a key element of the factory of the future, allowing easier implementation of new control features, avoiding the need for new PLC hardware when new actuators are deployed, and permitting the deployment of a different level of control functions on the same platform: factory, cell, actuator level. Cooperative device control is also possible.

Cooperation between humans and robots

One of the most interesting aspects of the fourth industrial revolution is the emergence of human-robot cooperation. The present and future challenge is to develop robots that can support human workers in a meaningful way to perform manipulation and assembly tasks according to a

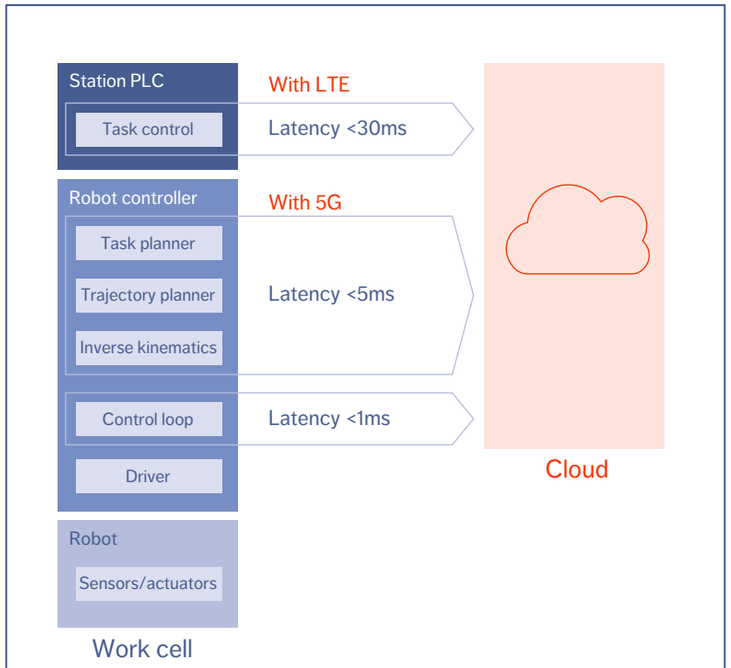


Figure 4: Virtualization of control in the cloud: steps toward true cloud robotics enabled by 5G

FEATURE ARTICLE ✱

> ROBOT-HUMAN COOPERATION IN THE FACTORY OF THE FUTURE WILL MAKE COMPLEX DECISION-MAKING ADVANCEMENTS POSSIBLE

production program. A robot that works close to a human worker must interact safely and be able to 'understand' and interpret direct user commands and support the worker in executing different actions.

Collaborative robotics is a novel paradigm of human-robot cooperation that is based on lightweight and flexible robots that are safe, smart, and easy to program, and are intended to operate in close symbiosis with human workers. Compared with the previous generation of robots, collaborative robots require sensory systems to detect and prevent collisions and impacts, as well as human-robot interfaces to understand and interpret human intentions. For these reasons, massive efforts in robotics and automation research are dedicated to the development of sensory skins and proximity sensors, and to the design of novel interfaces that enable different kinds of commands from the user to the robot.

Robot-human cooperation in the factory of the future will make complex decision-making advancements possible. For example, collaborative robots can help humans by:

- evaluating complex realities and providing synthesized and understandable representations that enable better decision making
- helping them to understand risks in advance and thereby reduce the probability of faults or fatalities
- making time-sensitive decisions when a human is not available to do so.

Conclusion

Cooperation between humans and intelligent machines is a new reality that will have a profound effect on both industry and society in the years ahead. Already today, it is possible to leverage a combination of human wisdom and intuition together with the strong elaboration capabilities of artificial intelligence, artificial learning and thinking, to create solutions that provide a high level of industrial automation. The next step, the factory of the future, will be realized through the digitization of the manufacturing process and plants, which will be enabled by 5G networks and all their building blocks.

As a leader in 5G infrastructure, including cloud technologies, big data analytics and IT capabilities, Ericsson is committed to playing a leading role in this transformation. The first results from our cooperation with Comau and the Sant'Anna School of Advanced Studies are very encouraging, and suggest that we are in a good position to be a key partner for many industries in the years to come.

References

1. **Günter Schuh et al., Industrie 4.0 Maturity Index, Acatech study, available at:** *www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Publikationen/Projektberichte/acatech_STUDIE_Maturity_Index_eng_WEB.pdf*
2. **Yoram Koren, The Global Manufacturing Revolution: Product-Process-Business Integration and Reconfigurable Systems, John Wiley & Sons, Inc., 2010, available at:** *http://onlinelibrary.wiley.com/book/10.1002/9780470618813*
3. **Ericsson Technology Review, June 2016, Cloud robotics: 5G paves the way for mass-market automation, available at:** *https://www.ericsson.com/en/publications/ericsson-technology-review/archive/2016/cloud-robotics-5g-paves-the-way-for-mass-market-automation*

Further reading

- **Ericsson white paper, 5G Systems, January 2017, available at:** *https://www.ericsson.com/assets/local/publications/white-papers/wp-5g-systems.pdf*
- **Ericsson white paper, 5G Radio Access, April 2016, available at:** *https://www.ericsson.com/assets/local/publications/white-papers/wp-5g.pdf*

THE AUTHORS

Roberto Sabella

◆ joined Ericsson in 1988 after having graduated in Electronics Engineering at the University of Rome "La Sapienza" the year before. He is manager of the Italian branch of Ericsson Research and leader of the Innovation with Industries in Tuscany initiative related to the 5G for Italy program in cooperation with TIM. His expertise covers several areas of telecom networks, such as

packet-optical transport networks, transport solutions for mobile backhaul and fronthaul, and photonics technologies for radio and data centers. He has authored more than 150 papers for international journals, magazines and conferences, as well as two books on optical communications, and holds more than 30 patents.

Andreas Thuelig

◆ joined Ericsson in 1991. He is currently responsible for the 5G for Europe program, which focuses on the relevance of 5G for industries and Industry 4.0. During his time at Ericsson, he has held leading positions in systems and technology organizations and in product management and solution management.

He holds an M.Sc. in electrical engineering from the RWTH Aachen University in Germany.

Maria Chiara Carrozza

◆ contributed to this article in her role as coordinator of the neuro-robotics area in the BioRobotics Institute at the Sant'Anna School of Advanced Studies in Pisa, Italy. She is also a partner in IUVO, a startup in wearable robotics that was founded in 2015 as a spinoff from the institute. She served as rector of Sant'Anna from

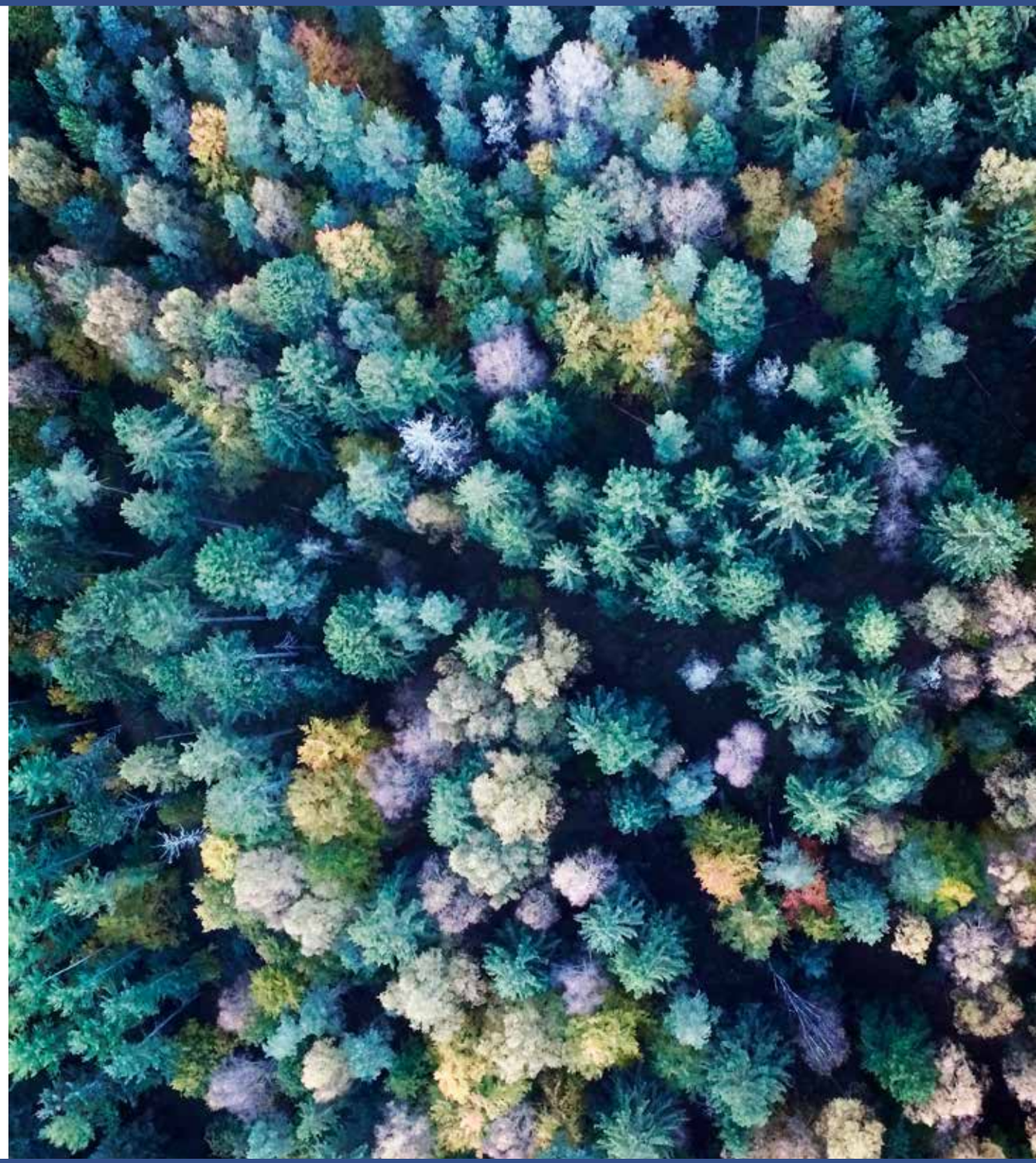
2007 to 2013, when she was elected to the Italian national parliament and was appointed to the Foreign and European Affairs Committee. She has been president of the Italian National Bioengineering Group since 2016 and became a member of the Agency for Digital Italy's task force on Artificial Intelligence in 2017.

Massimo Ippolito

◆ contributed to this article in his role as innovation manager at Comau, a multinational company specializing in industrial automation, where he has worked since 2012. He holds a Ph.D. in industrial production engineering from Parma University and an M.Sc. in computer science from the University of Milan. He has extensive experience

in methodologies and tools for product and production system design. Since 2000, he has been involved in various international research projects in the product design and manufacturing area, ranging from methodologies for product design for manufacturing to process design for energy efficiency. From 2007 to 2012, he was responsible for a Fiat Group innovation research program on manufacturing purpose.

The authors wish to acknowledge Marzio Puleri and Giulio Bottari at Ericsson Research and Guido Rumiano and Pietro Cultrona at Comau for their valuable contributions to this article.



ENABLING intelligent transport IN 5G NETWORKS

STEFAN DAHLFORT, ANTONIO DE GREGORIO, GIOVANNI FIASCHI, SHAHRYAR KHAN, JONAS ROSENBERG, TOMAS THYNI

The evolution toward 5G mobile networks is driven by the diverse requirements of a multitude of new use cases in the areas of enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC) and massive machine-type communications. Along with a demand for lower costs, these drivers have led to the development of split architectures for the RAN to support multiple deployment models. Since the transport network's role is to connect all the pieces of the RAN and the mobile core network, optimal performance in 5G scenarios will require high levels of intelligence, flexibility and automation in the transport network.

Terms and abbreviations
BBU – baseband unit | **BPF** – baseband processing functions | **BSC** – base station controller | **CO** – central office | **CPRI** – Common Public Radio Interface | **CU** – centralized unit | **DU** – distributed unit | **DWDM** – dense wavelength division multiplexing | **eCPRI** – evolved CPRI | **eMBB** – enhanced mobile broadband | **ERAN** – Elastic RAN | **GW** – gateway | **Obs** – observability | **O&M** – operations and maintenance | **ONAP** – Open Network Automation Platform | **PE** – provider edge (router) | **QoE** – quality of experience | **RBS** – radio base station | **RESTCONF** – Representational State Transfer Configuration | **RF** – radio functions | **RNC** – radio network controller | **RRU** – remote radio unit | **RRU-BF** – remote radio unit beam forming | **RU** – radio unit | **SBH** – self backhauling | **SDN** – software-defined networking | **SDNc** – SDN controller | **SLA** – Service Level Agreement | **SW** – switch | **SW/R** – switch/router | **TCO** – total cost of ownership | **TIF** – transport intelligent function | **URLLC** – ultra-reliable low-latency communication | **µW** – microwave | **VPN** – Virtual Private Network | **VRAN** – virtualized RAN

At a high level, it is clear that network operators need to continuously look into new revenue streams, faster deployment of required transport connectivity and new consumer services to lower total cost of ownership (TCO). Together with RAN and mobile core networks, transport networks need to evolve to provide the desired level of flexibility in service offerings, simple and agile service configurations, and support for new operations models and cross-domain orchestration that are expected in 5G.

New revenue streams are always difficult to predict, but greater automation can reduce opex, particularly if it enables more intelligent interaction between the RAN, the transport network and the mobile core. Network slices that are supported in a well-coordinated way across RAN, mobile core and transport networks will be able to provide improved life cycle management of services on an end-to-end basis [1]. This is particularly relevant for 5G, but valuable for 3G and 4G networks as well.

The latest demands on the transport network come from areas such as increasing RAN and mobile broadband service capacity, new 5G-enabled services (such as those in the use cases presented below), and the dynamic deployment flexibility of the 5G RAN split architecture, with its tight transport characteristics. These characteristics are especially manifested in the fronthaul portion of RAN transport (in other words, from the DU to other DUs and to the centralized unit (CU) control and packet processing parts in *Figure 1*), where the latency and synchronization requirements are very challenging. Enhanced automation capabilities in the operations and management domain represent a key requirement to meet these challenges.

> ENHANCED AUTOMATION CAPABILITIES IN THE OPERATIONS AND MANAGEMENT DOMAIN REPRESENT A KEY REQUIREMENT

Transport network solution for 5G: architectures and protocols

5G RAN's deployment flexibility and service requirements demand much greater control and knowledge of the transport network resources and characteristics compared to previous generations. The increasingly flexible and dynamic behavior in Elastic and virtualized RAN also requires the ability to dynamically add or remove transport connectivity. This is important in the mobile-centric access aggregation part of the network to maintain and improve RAN performance. It is also important for user services, where the positioning of the mobile core becomes critical, or when a new service is launched using mobile core resources at a different site.

The need for flexibility and the above requirements will be a major opex driver in the transport network unless it is fully automated.

This is especially true in the mobile-centric part of the network. To overcome this challenge, we have used software-defined networking (SDN) along with an intelligent application, the transport intelligent function (TIF), to design what we consider to be an optimal 5G transport network architecture.

SDN is one of several possible technical solutions to achieve a higher level of automation. However it is achieved, more automation will be required in the near future. Today many network (RAN, transport) operations, such as port and traffic flow configurations, are done manually – that is, by means of a network operations person inserting commands or using graphical user interfaces. While this has been doable (though not always so efficient) in earlier generations, the increased ability to configure and optimize in 5G will make manual operations inviable going forward. The use cases below further illustrate this point.

Our optimal 5G transport network is built as a self-contained infrastructure underlay with an SDN-controlled overlay for a variety of RAN and user services. The distributed control plane in the underlay maintains the basic infrastructure and handles redundancy and quick restoration in case of network failures. The service and characteristics-aware overlay is handled by the SDN controller with the TIF application, and this creates a dynamically controlled and orchestrated transport network that requires minimum manual interaction. In other words, the underlay network described here handles the infrastructure connectivity and the network overlay handles the services running on top of the underlay.

The automation focus of the overlay relates to the establishment of initial and dynamic connectivity and is based on requests from the TIF, which is a RAN-aware application [2]. It is also the responsibility of the TIF application to collect and ensure appropriate transport characteristics for different RAN connections as well as user services.

Automation application framework

Network automation in the context of transport is a broad area that includes concepts like zero-touch and self-optimizing networks, so there is a need to clarify what exactly we mean when we use the term. Ericsson is keen to define a framework with regard to automation. While our definition is in line with ongoing industrial initiatives, it can also provide unique building blocks from the RAN and transport interaction perspective [3]. Defining a transport automation framework is an important step.

> THE APPLICATION FRAMEWORK INHERITS A HIGHLY MODULAR MICROSERVICE ARCHITECTURE

An automation application framework provides the required flexibility of connectivity in the various RAN deployment scenarios. This framework defines the platform for lightweight applications to be developed on top. The following key components are used:

- cross-domain communication for real-time notifications/events and service request
- network and service observability function
- performance-aware topology manager
- path computation and optimization function
- network and service logic
- interfaces to control layer (SDN controller, for example) and to management and orchestration system
- interface to network observability aggregator function

It should be noted that some of these components are not exclusively used by the TIF. While the architectural details are outside the scope of this article, it is worth mentioning that the automation framework outlined here has been designed with the Open Network Automation Platform (ONAP) in mind.

Transport intelligent function

While the control layer provides an abstract view of the transport network and enables its programmability, the intelligence to decide how to optimize/reconfigure the network is provided by the TIF. The TIF utilizes the network programmability features offered by the

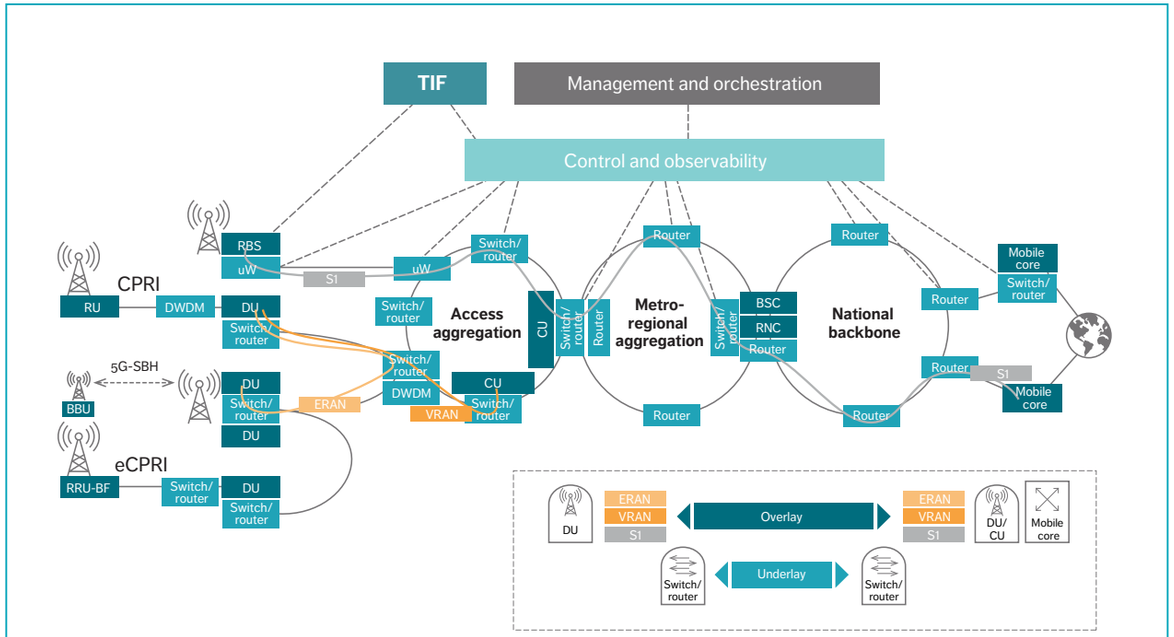


Figure 1 Transport network architecture. The color coding indicates the four different domains: RAN, Transport, Mobile core, Control and observability, and Management and orchestration. TIF represents a cross-domain function between RAN and Transport (the latter via the Transport control and observability).

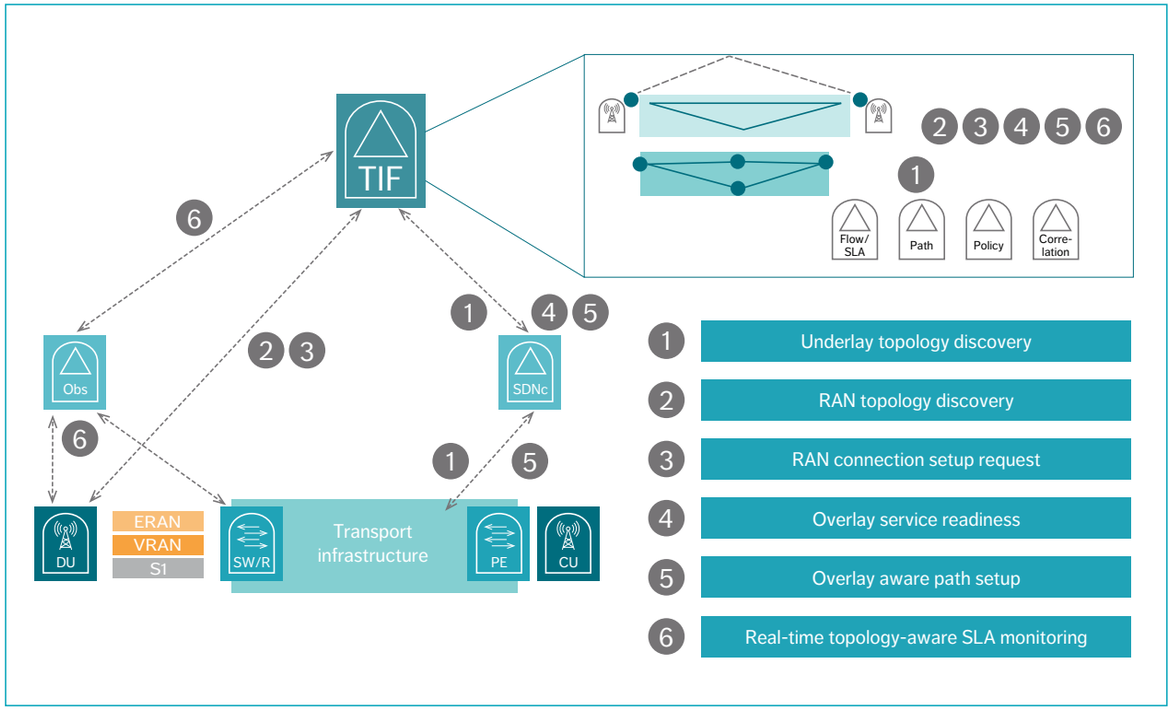


Figure 2 ERAN/VRAN auto-configuration

control layer and automation application framework, and enables the automation and optimization of the various required inter-5G-network connectivity cases.

A cross-domain communication bus is used to autonomously receive connectivity requests from the RAN. The TIF receives these requests via this bus and processes them together with performance-aware topology data and a pre-configured set of policies (retrieved from management and orchestration systems). To ensure the optimal actions are taken, the TIF continuously requests insights from the observability function, which, in turn, retrieves its status and performance information from the network via the control layer. This information enables the TIF application to decide, at any time, how best to automate, optimize and configure the transport network.

Once all the data described above is available, the path computation and optimization function processes the data and provides the configuration actions to be taken by the transport control layer.

Figure 1 shows the control architecture, including the control layer interfaces to the network and the TIF interfaces to the RAN functions and control layer for the transport network and services. The geographical area a TIF function operates in would typically be around a portion of a metro area, or even just a few sites. As illustrated in *Figure 2*, with an area of that size, the signaling load on the TIF would not be technically challenging.

The transport automation framework is applicable to three very distinct stages of the life cycle management process. In auto-integration, the framework supports auto-installation of transport network infrastructure, establishes basic configuration with O&M connectivity to the controller and registers with control and management functions. In auto-configuration, the framework automatically configures the transport connectivity service for the desired RAN interfaces (such as virtualized RAN or VRAN), and interacts with the RAN from a connectivity requirement perspective. Finally, in auto-optimization, it automatically optimizes transport resources during the run time phase to achieve overall optimal QoE for consumers and interacts with the RAN domain to achieve overall optimization.

While coordinated RAN transport automation is useful for auto-integration, its benefits are much more significant for auto-configuration and auto-optimization, the focus areas of the two TIF use cases presented below.

Use case 1: auto-configuration in ERAN and VRAN scenarios

When Elastic RAN (ERAN) is deployed, inter-site baseband connectivity is needed to give the RAN enough freedom to arrange radio coordination patterns. However, manually setting up the connectivity can be complex. On one hand, multiple paths are desirable to exploit all the resources of the transport network, distribute the traffic load and reduce the probability of packet collisions and consequent delays, which would dramatically reduce the advantages of the ERAN feature itself. On the other hand, unless very specific latency constraints are respected, the connectivity will be unusable for the application. Finally, to minimize the collision probability on the whole network, the traffic distribution should be globally optimized.

In VRAN deployments (also known as split architecture [4]), the connectivity accuracy is not as critical as in ERAN. The complexity in these cases

> THE AUTO-CONFIGURATION PROCESS FOR ERAN AND VRAN SCENARIOS REQUIRES COMPLETE AUTOMATION

is due to the larger scale and possibly heterogeneous nature of the network. Portions of the underlay network may be self-built by the operator and may consist of third-party equipment and technology. The data center architectural elements must also be orchestrated in the VRAN, which adds further complexity. The required connectivity between the baseband functions in hub sites and the virtualized functions in central offices (COs) and aggregation sites is many-to-many, and adaptive transport is needed to ensure resilience, continuous traffic optimization and reconfiguration after the addition and removal of the packet processing function.

In both ERAN and VRAN, constant monitoring of the transport network status (to detect temporary partial outage, for example) is also required to be forwarded to the RAN domain to allow for consequent actions.

Figure 2 illustrates the operation sequence, and one aspect worth highlighting is the communication between the RAN and transport that takes place through the TIF. This communication includes the specification of RAN flow setup requirements to the transport network in a dynamic fashion. The TIF is responsible for taking these specific requirements and translating them into computation of requisite transport paths and optimal distribution of RAN flows across these paths to ensure the desired Service Level Agreement (SLA) requirements.

Figure 2 shows the six main TIF operations in relation to the other components of the network: (1) underlay topology discovery, (2) RAN topology discovery, (3) RAN connection setup request, (4) overlay service readiness, (5) overlay aware path setup and (6) real-time topology-aware SLA monitoring. The auto-configuration process for

> CONGESTION IN A CERTAIN PART OF THE TRANSPORT NETWORK WOULD RESULT IN DETECTION OF ALL IMPACTED OVERLAY SERVICES

ERAN and VRAN scenarios requires complete automation of the whole sequence.

During underlay topology discovery (1), the TIF acquires the transport network underlay topology from the SDN controller via a standard RESTCONF interface. In RAN topology discovery (2), the TIF acquires the RAN topology directly from the RAN. It then integrates it with the transport topology to build the entire network topology view. During RAN connection setup request (3), RAN sends connectivity requests to the TIF when needed, including endpoint information along with constraints such as maximum allowed latency and expected bandwidth usage.

During overlay service readiness (4), the TIF triggers the process of the overlay service setup based on the connectivity setup request. This includes the VPN service configuration parameters and the policies required to be configured to match RAN flows to the desired transport paths, which are configured as part of the next step.

During overlay aware path setup (5), the TIF computes all possible paths according to the overlay service requirements and then requests the SDN controller to provide the desired paths on the transport edge nodes. One important consideration in the case of ERAN is the need for optimal load-balancing on RAN flows on all feasible transport paths. The idea is to ensure the optimal usage of all available transport resources to accommodate the inherently bursty traffic. This requires the implementation of a customized load-balancing algorithm in the TIF, as it is responsible for handling the overlay service part.

During real-time topology-aware SLA monitoring (6), the monitoring systems continuously supervise both the overlay RAN flows and the underlying transport network and update the TIF in case of any changes. This is necessary because enhanced observation levels are key to providing a desired SLA for the RAN flows.

This auto-configuration process for ERAN and VRAN scenarios is designed to enable complex and large-scale deployments in an agile manner, simplifying operations significantly and reducing TCO for the operator.

As previously mentioned, the TIF provides the mapping and correlation between the VRAN or ERAN traffic flows and the underlay transport. Such an enhanced level of observability into the usage of various transport resources and overlay services can be used to provide much more advanced service level guarantees. For example, congestion in a certain part of the transport network would result in detection of all impacted overlay services. This could then trigger optimization of various transport resources to protect SLAs for the impacted overlay services. The process of detecting SLA violation and doing a correlation and an impact analysis makes an excellent case for a type of automation called auto-optimization. This is a natural next step to the auto-integration and auto-configuration processes.

Use case 2: auto-optimization

Figure 3 illustrates an example of end-to-end auto-optimization. In this scenario, if congestion is detected in the transport path to the CO that serves the URLLC services, the transport path and the RAN and core functions can be moved to another CO that provides the same level of reliability and latency constraints. This is only possible because the TIF can detect the congestion in the transport network, identify the impacted URLLC flows and take the necessary corrective action in conjunction with other domains such as the RAN and core. This kind of flexibility, agility and cross-domain orchestration in an automated fashion is the key to offering highly demanding new 5G services such as URLLC.

Meanwhile, concurrently existing services such as eMBB must not be adversely impacted by this flexibility. In other words, a single transport network must be able to support an extremely diverse set of

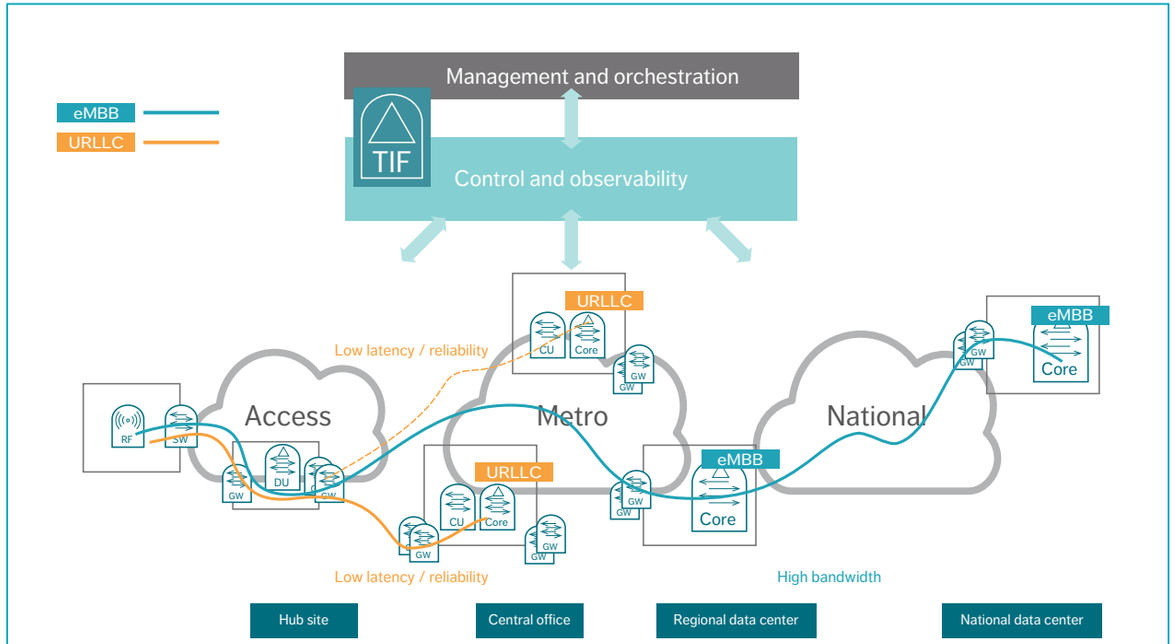


Figure 3 Example of end-to-end auto-optimization

services. The TIF plays an important role in providing the capability to track the SLAs of all the overlay services. Furthermore, with the TIF in place, run time impact analysis of changes in the underlay network characteristics like bandwidth, latency or packet loss is carried out in conjunction with the overlay SLA requirements, and the necessary corrective action is taken. One such example of corrective action might be dynamically moving the flows of an overlay service to an alternate path computed with minimal or no impact on the other overlay services.

Global orchestration and optimization of all network level resources is another important consideration in the auto-optimization process. We are currently studying this aspect, as we believe it will take network automation to a whole new level. Domain-specific intelligent functions such as TIF for transport are set to play a critical role in the global and/or cross-domain optimization owing to the level of visibility and inherent inter-domain communication capabilities.

Conclusion

The transport network tends to get less attention than RAN and mobile core networks in discussions about 5G, but as the vital link between all the pieces, it too requires significant enhancement to support the diverse set of services and deployment models expected in 5G. Intelligent, automated coordination between RAN, transport and mobile core networks will undoubtedly be a key part of a robust 5G solution, because without automation it will not be possible to achieve the required levels of flexibility and observability. The TIF solution provides the requisite intelligence and acts as a catalyst for automation, enabling operators to meet the 5G requirements of multiple use cases while simultaneously reducing opex.

References

1. **Ericsson Technology Review, February 2016, A vision of the 5G core: flexibility for new business opportunities, available at**: *https://www.ericsson.com/en/ericsson-technology-review/archive/2016/a-vision-of-the-5g-core-flexibility-for-new-business-opportunities*
2. **Ericsson Review, July 2015, Radio access and transport network interaction – a concept for improving QoE and resource utilization, available at:** *https://www.ericsson.com/en/news/2015/7/radio-access-and-transport-network-interaction--a-concept-for-improving-qoe-and-resource-utilization*
3. **Ericsson Technology Review, December 2016, Fixed wireless access on a massive scale with 5G, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2016/fixed-wireless-access-on-a-massive-scale-with-5g*
4. **Ericsson Technology Review, July 2016, 4G/5G RAN architecture: how a split can make the difference, available at:** *https://www.ericsson.com/en/ericsson-technology-review/archive/2016/4g5g-ran-architecture-how-a-split-can-make-the-difference*

Further reading

- *https://www.ericsson.com/en/networks/trending/hot-topics*
- *https://www.ericsson.com/en/networks/offerings/network-intelligence-and-automation*

THE AUTHORS

Stefan Dahlfort

◆ has worked in the telecom industry for 22 years. He founded a start-up before joining Ericsson in 2007 as a manager for fiber to the x research. Having held different management positions in Research and Systems & Technology in the transport area, he is now based in Santa Clara in Silicon Valley as a product development leader. He holds an M.Sc.in electrical engineering and a Ph.D. in optical networking from KTH Royal Institute of Technology in Stockholm, Sweden.

Antonio De Gregorio

◆ has 23 years of experience in the telecom industry. He joined Ericsson in 1997, initially working in the Operations and Maintenance area. Since then he has been responsible for the strategy and technology units of different product development organizations in Europe and in Silicon Valley in the US. He currently heads the Technical Management organization in Genoa, Italy within Development Unit Network Management. He holds an M.Sc. in electronic engineering from the University of Salerno, Italy.

Giovanni Fiaschi

◆ joined the company in 2005 when Marconi, his employer since 1990, was acquired by Ericsson. He has been working in telecommunications for more than 25 years, specializing mostly in control functions for transport networks. Fiaschi currently works at Development Unit Networks, Systems and Technology, focusing on mobile transport network simulations and control solutions. He is also active in the IPR and security areas. He holds an M.Sc. in computer science from the University of Pisa, Italy.

Shahryar Khan

◆ has nearly two decades of experience in architecture design and integration for multiservice IP and transport networks for telecom operators and large enterprises. Khan held several roles within Ericsson during the period 2005-2017, most recently serving as an expert and chief architect in multiservice IP and transport networks in Development Unit Networks in Kista, Sweden. He holds a B.Sc. in electrical engineering from the University of Engineering and Technology in Lahore, Pakistan.

Jonas Rosenberg

◆ is a senior specialist in network architecture and solutions whose main area of expertise is within technologies and software for cross-domain automation and assurance solutions in mobile networks. He works as a leader for the mobile transport technology area and is responsible for Transport O&M within Development Unit Networks. Rosenberg joined Ericsson in 2000 and holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden.

Tomas Thyni

◆ is an expert in the area of IP and transport networks. A telecommunication and network engineer, he joined Ericsson in 2000 and has worked within the IP, broadband and optical networks areas. Today, Thyni works as a chief architect on new concepts for transport in RAN at Development Unit Networks; one such concept area is RAN and transport interaction. Prior to joining Ericsson, he accumulated 15 years of experience as an IP and transport network designer with various network operators.

END-TO-END

Security Management

FOR THE IoT

Industries everywhere are digitizing, which is creating a multitude of new security requirements for the Internet of Things (IoT). End-to-end (E2E) security management will be essential to ensuring security and privacy in the IoT, while simultaneously building strong identities and maintaining trust.

KEIJO MONONEN, PATRIK TEPPO, TIMO SUIHKO

As the diversity of IoT services and the number of connected devices continue to increase, the threats to IoT systems are changing and growing even faster.

To cope with these threats, the ICT industry needs a comprehensive IoT security and identity management solution that is able to manage and orchestrate the IoT components horizontally (from device to service and service user) and vertically (from hardware to application). In addition to this, the ability to address both security and identity from the IoT device all the way across the complete service life cycle will also be essential.

Figure 1 illustrates an E2E approach to security and identity that highlights three key aspects: security and identity management, security and identity functions, and trust anchoring.

IoT actors and trust

IoT systems support new business models that involve new actors in conjunction with traditional telecommunication services. Aside from consumers and mobile network operators, enterprises, verticals, partnerships, infrastructure, and services play increasingly vital roles. All of these actors affect trust.

Figure 2 presents the main and supporting IoT actors and their trust relationships. The three main actors in an IoT solution are the IoT service user, the IoT service provider, and the devices that enable the provision of the IoT service. The supporting actors are the IoT platform service provider, whose role is to provide the IoT platform for the IoT service provider, and the connectivity service provider, whose role is to provide connectivity for the IoT devices and service.

The trustworthiness of services and service use depends on how the actors govern identities and data, security and privacy, and the degree to which they comply with the agreed policies and regulations. The combination of the security and identity functions is important for defining the trust level. For example, hardware-based trust does not help if the application does not make use of it. A fully trusted application does not help if the communication cannot be trusted. An E2E approach is therefore essential to ensure trust among all actors across the system.

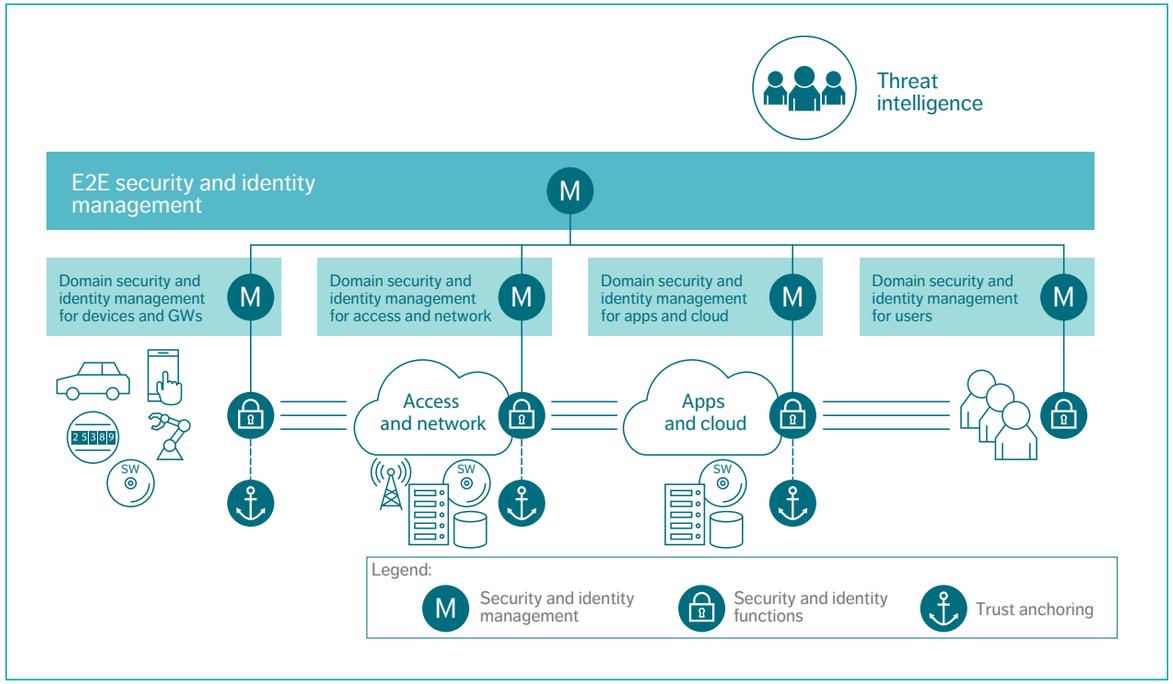


Figure 1 E2E approach to security and identity

E2E IoT security architecture

The purpose of an E2E IoT security architecture is to ensure the security and privacy of IoT services, protect the IoT system itself and prevent IoT devices from becoming a source of attacks – a Distributed Denial of Service (DDoS) attack, for example – against other systems.

Figure 3 illustrates Ericsson's view of how security can be managed and deployed in an E2E manner throughout IoT domains to monitor

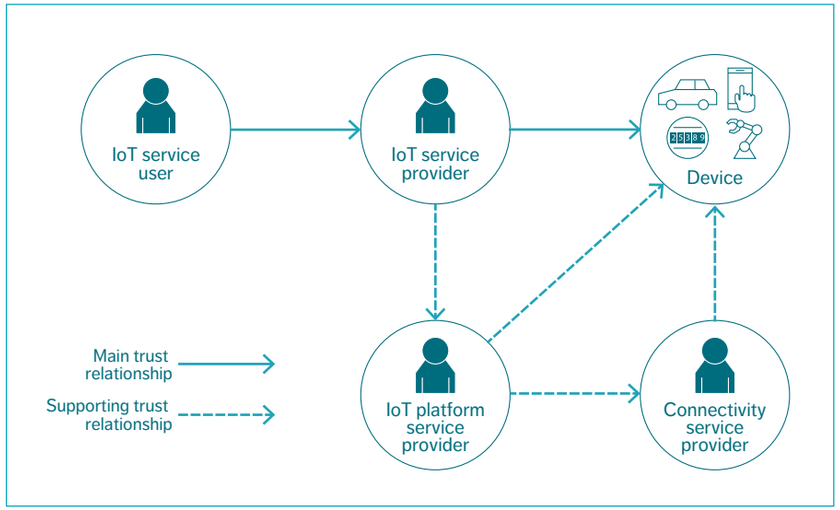


Figure 2 The main and supporting IoT actors and their trust relationships

and protect system resources and assets. The architecture consists of an E2E security and identity management layer, domain (device, gateway, access, platform and application) specific management layers, and security and identity functions in each domain component.

An IoT system spans from the device via different network interfaces to the cloud that hosts the platform and applications that provide services that are consumed by IoT service users. Each element of the chain must be considered when designing an E2E approach to security and identity in the IoT.

This approach leverages advanced security analytics and machine learning to provide threat, risk and fraud management at both E2E and domain management layers. To meet industry security and privacy standards, an E2E security management solution must also be in charge of overall security and privacy policies and compliance and be able to coordinate across a multitude of domain management systems through the establishment of cross-domain identities and relevant policies.

Domain management of security and identity functions within domains ensures that security and identities are properly managed, configured and monitored within the domain according to policies, regulations, and agreements. Vulnerability and security baseline management also occurs at the domain management layer based on E2E level policies.

According to this approach, the IoT service provider is responsible for managing IoT service security and identities E2E, whereas domain-level management can be delegated to the IoT platform service provider and connectivity service provider.

Figure 3 shows how the IoT domains are managed both horizontally and vertically. Horizontal (cross-domain) security is required at two levels: connectivity and application. Depending on connectivity type, security controls such as mutual authentication and encryption of data in transit are provided at the connectivity level. On top of connectivity, security is provided at the application level from device to cloud, based on identification and access management functions

and application security policies. Application level security can be independent of or dependent on (federated with) the connectivity level security.

Vertical security from hardware to application can be used in every domain to provide hardware-based root of trust, ensuring the integrity of the domain. The domains are built on trusted hardware and software. When required by the industry and the use case, trust is anchored to hardware.

The domains include security and privacy functions to handle identity and access management, data protection and right to privacy, network security, logging, key and certificate management, and platform/infrastructure security (including virtualization security and hardware-based root of trust).

For critical IoT services, the level of security functions must be set high in accordance with the risk management results and service provider security policies. For less critical IoT services, a lower level may be sufficient.

Security policy and compliance management

Business-optimal and trust-centric IoT security is dependent on continuous risk management that balances criticality, cost, usability and effectiveness to fulfill different types of security Service Level Agreements in multi-tenant IoT systems. Since the current management of IoT security is spotty at best, it must be transformed into unified security management with adaptive protection, detection, response and compliance driven by security policies. Only in this environment can service providers and their customers leverage E2E network and application knowledge to secure assets across all contexts.

Our vision of security policy and compliance management defines security policies using industry standards, regulation and organizational policies. This approach helps to automate security and privacy controls, maintain them at a desired level even in a changing threat landscape, and shorten the reaction time in response to potential breaches. Real-time visibility regarding general and industry-specific security standards and regulations makes it possible for IoT service providers to remediate policy

> A HIGH DEGREE OF AUTOMATION IS NECESSARY TO ENSURE A SWIFT RESPONSE TO ANY IDENTIFIED THREATS AND ANOMALIES

violations quickly and demonstrate compliance to security frameworks, including ISO, NIST, CSA, GDPR and CIS benchmarks, as well as an enterprise's own security and privacy policies. Having the security baseline configuration and compliance function at domain level ensures the automated hardening of the protected assets and supports continuous compliance monitoring in the defined security baseline.

Domain level security management requires an accurate asset inventory including all the assets that must be protected in the managed domain, such as authorized IoT devices and software. Automation of asset discovery and continuous monitoring is essential to keep the asset inventory updated. The vulnerability information is also correlated with the asset inventory to monitor and remediate the vulnerabilities of protected assets.

Rapid detection of attacks is crucial. Security monitoring and analytics functionalities must have the ability to analyze logs, events and data from IoT domain components combined with external data about threats and vulnerabilities. Machine learning technology makes it possible to learn from and make predictions based on data. Coupling a machine learning analytics engine with central threat intelligence improves the detection of zero day attacks and reduces the response time for known threats.

On top of a monitoring and analytics engine, solutions relating to vulnerability, threat, fraud and risk management, along with security policy and orchestration components, are also required to automate security controls and maintain them at desired levels in a changing threat landscape.

Combining the information feeds for vulnerability, threat and fraud management results in timely

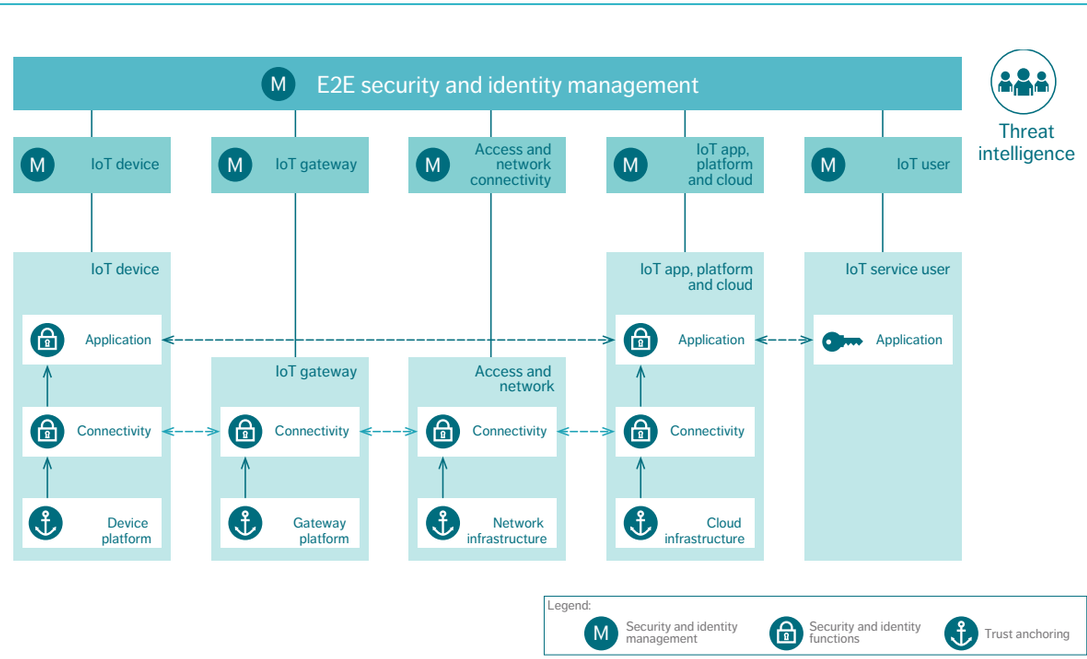


Figure 3 E2E approach to security and identity

and accurate information for evaluating potential risks and helps to direct efforts in protecting the most exposed critical assets. A high degree of automation is necessary to ensure a swift response to any identified threats and anomalies.

Since not all security breaches and attacks can be prevented, it is crucial to have an efficient security incident management process that ensures rapid response and recovery. Real-time insights and audit trails from tools such as security monitoring, analytics and log management help to find the root cause of an incident. The same information can be also used as the evidence in digital forensic investigations.

Identity management

The main purpose of identity management is to manage the life cycle of identities and provide identification, authentication and access control services for identities. There are various identities that serve different purposes in the IoT approach, but the main ones are for device and user identification. The others are used for management of devices, functions and services. Identifiers and keys are also used to sign data, including software and firmware. These different device identities are needed to identify the devices for connectivity within the access and network domains, and to identify device applications in the IoT platform and cloud domain.

The level of trust in the device identity depends on the strength of authentication both at the connectivity (for example, 3GPP, Wi-Fi and fixed) and application layers. For device identity to be trusted, strong authentication and follow-up of the device integrity – with the help of hardware-based root of trust in the device, for example – would be needed.

A device will have different identifiers depending on where it is in its life cycle. Life cycle management of device identities is part of the security management layer. More than one security management domain is involved when provisioning identities. Connectivity and IoT service provider could be different players where each player takes care of its own identity life cycle management.

When a device is manufactured, the vendor will give it an identifier that could have different trust levels. Vendor credentials could be protected in hardware (preferred) or they could be nothing more than a serial number printed on the device. The device has to be authenticated by the IoT system, and newly given identifiers and credentials (bootstrap process) will be used for connectivity and application accesses.

Identifiers and credentials can be changed during the device life cycle depending on different triggers such as expiration of credentials, change of service provider and so on. Connectivity identities are dependent on the connectivity type and have different life cycle management processes. For example, 3GPP access is based on SIM identities (IMSI and AKA credentials). SIMs are either physically removable ones or SIMs (i.e. eUICC) that can be remotely provisioned [1].

The user identities are needed to identify the users of the services within the applications and cloud domain. There may be several different ways to verify (authenticate) the user identities such as single- or multi-factor authentication, federated authentication, or authentication tokens. Each of these provides a certain level of authentication strength.

Due to layered security management architecture and the involvement of several actors (including industries) in the IoT, any identity and access management solution must be able to cooperate with and adapt to external identity and access management systems. On top of identification and authentication, there must also be access control for users so that only the permitted services are authorized.

Threat intelligence

Threat intelligence is built and shared in communities. Therefore, a centralized threat intelligence solution must be able to interface with different threat intelligence sources to learn about existing and new threats. Consolidation and correlation of security audit feeds from different domains are necessary to provide a clear view of threat insights across all IoT domains.

Automation and machine learning can be used to great advantage in threat intelligence, to create and share indicators of compromise that are actionable, timely, accurate and relevant to support strategic decision-making and to understand business risks in detail. Targeted threat intelligence feeds are a great way to generate customer-specific threat intelligence.

Two IoT use cases

Two concrete examples of how an E2E security management solution can help address IoT challenges are provided below.

DDoS detection and prevention

In October 2016, the Mirai botnet exploited a vulnerability in IoT devices to launch a DDoS attack against a critical DNS server that disrupted a number of the internet's biggest websites, including PayPal, Spotify and Twitter.

Mirai was designed to exploit the security weaknesses of many IoT devices. It continuously scans for IoT devices that are accessible over the internet and are protected by factory default or hardcoded user names and passwords. When it finds them, Mirai infects the devices with malware that forces them to report to a central control server, turning them into bots that can be used in DDoS attacks.

Strong detection and prevention mechanisms are needed against DDoS attacks that attempt

to saturate the network by exhausting the bandwidth capacity of the attacked site, the server resources or service availability. In our view, an optimal outbound DDoS (botnet) detection and mitigation solution includes remote attestation to verify device trustworthiness and detect malware, monitoring of outbound traffic, anomaly detection, infected entities isolation or blocking and setting of traffic limit policies. Optimal inbound DDoS detection and mitigation includes monitoring of inbound traffic, anomaly detection, setting of traffic limit policies and redirecting malicious traffic to a botnet sinkhole.

The security management layer plays a critical role in detecting and mitigating DDoS attacks. In our framework, DDoS attacks are detected by the security monitoring and analytics functions through the observation of device and network behavior and identification of anomalies. Once an anomaly is detected, immediate mitigation actions can be triggered.

GDPR compliance

There is a legitimate expectation in society that IoT solutions will be designed with privacy in mind. This is becoming especially evident in certain jurisdictions: for example, in the European Union with the new General Data Protection Regulation (GDPR) [2].

Data integrity, data confidentiality, accountability and privacy by design are all fundamental to the protection of sensitive personal data. Such data can be protected via appropriate privacy controls. These controls include personal data identification and classification, personal data management and fair data processing practices. When actual personal data might be exposed, additional privacy protective measures will be applied such as data encryption and data anonymization.

Another focus area in the IoT security domain is the privacy breach response. Dedicated privacy logging and audit trail functionality can be used to improve the ability to prevent, detect and respond to privacy breaches in a more prompt and flexible way. Such capabilities will be essential to respond to privacy breaches swiftly (within 72 hours, as prescribed by the GDPR).

Implementing a GDPR compliance tool in the security management layer makes it easier to meet GDPR requirements. To do its job right, it must be able to provide identification and classification of personal data, enforcement of data privacy policies according to the GDPR, demonstration of compliance to the GDPR, and detection, response and recovery from privacy incidents.

Conclusion

The IoT offers a wealth of new opportunities for service providers. Those who want to capitalize on them without taking undue risks need a security solution that provides continuous monitoring of threats, vulnerabilities, risks and compliance, along with automated remediation. Ericsson's E2E IoT security and identity management architecture is designed with this in mind, managing and orchestrating the IoT domains both horizontally and vertically, and addressing both security and identity from the IoT device throughout the service life cycle. ❂

Terms and abbreviations
AKA – Authentication and Key Agreement | **CIS** – Center for Internet Security | **CSA** – Cloud Security Alliance | **DDoS** – Distributed Denial of Service | **DNS** – Domain Name System | **E2E** – end-to-end | **eUICC** – embedded Universal Integrated Circuit Card | **GDPR** – General Data Protection Regulation | **GW** – gateway | **IMSI** – International Mobile Subscriber Identity | **IoT** – Internet of Things | **ISO** – International Organization for Standardization | **NIST** – National Institute of Standards and Technology | **SIM** – Subscriber Identity Module | **SW** – software

THE AUTHORS

Keijo Mononen
◆ is general manager of Security Solutions at Ericsson. In this role he is responsible for end-to-end security management solutions including security automation and analytics. Mononen joined Ericsson in 1990 and for the past 15 years he has held leading positions in professional security services and in security

technology development. He holds an M.Sc. in computer science and engineering from Chalmers University of Technology in Gothenburg, Sweden.

Patrik Teppo
◆ joined Ericsson in 1995 and is currently working as a security architect with the CTO Office, Architecture and Portfolio team. He is responsible for the security part of the Ericsson architecture and leads Ericsson's IoT security architecture work. He holds a B.Sc. in software engineering from Blekinge Institute of Technology, Sweden.

Timo Suihko
◆ joined Ericsson in 1992 and is currently working as a senior security specialist in the Ericsson Network Security, Security Technologies team, which belongs to Group

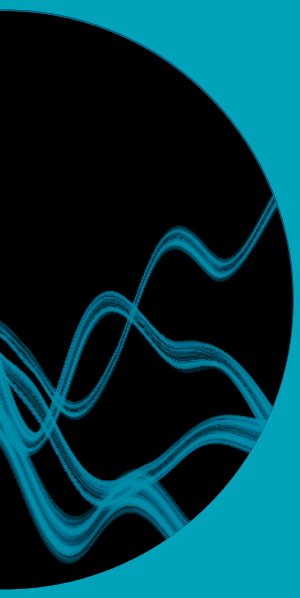
Function Technology and Emerging Business. He holds an M.Sc. from Helsinki University of Technology.

References

1. **GSMA Remote SIM Provisioning Specifications, available at:** *https://www.gsma.com/rsp/*
2. **Official Journal of the European Union, May 2016, Regulation (EU) 2016/679, General Data Protection Regulation (GDPR), available at:** *http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1490179745294&from=en*

Further reading

» **Ericsson white paper, February 2017, IoT Security – Protecting the Networked Society, available at:** *https://www.ericsson.com/en/publications/white-papers/iot-security-protecting-the-networked-society*
» **Ericsson, Security Management, available at:** *https://www.ericsson.com/en/in-focus/security/security-management*
» **Ericsson, Identity Management, available at:** *https://www.ericsson.com/en/in-focus/security/identity-management*
» **ETSI GS NFV-SEC 013, V3.1.1, February 2017, Network Functions Virtualisation (NFV) Release 3; Security; Security Management and Monitoring specification, available at:** *http://www.etsi.org/deliver/etsi_gs/NFV-SEC/001_099/013/03.01.01_60/gs_NFV-SEC013v030101p.pdf*



ISSN 0014-0171
284 23-3321 | Uen

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