Ericsson - Deutsche Telekom White Paper
BNEW-21:025455 Uen
May 2021

Enabling time-critical applications over 5G with rate adaptation

Abstract

To fulfill its true business promise [1], 5G must unlock the potential of innovative time-critical applications beyond mobile broadband (MBB). These applications are sensitive to delays caused by cellular network congestion and thus require reliable low latency. This is especially valid for emerging time-critical applications with high data rate requirements, such as cloud-driven artificial reality (AR)/virtual reality (VR), cloud gaming, and remote control of machines and vehicles over 5G networks. Congestion delays are unpredictable (and thus cannot be avoided entirely). In order to realize high data rate time-critical applications at a large scale with 5G networks, it is essential that these applications be able to react to network congestion delays and meet desired latency targets by adapting bitrates in real time.

To achieve global market adoption for this application type, a common feedback loop mechanism is required for fast rate adaptation. Low-Latency, Low-Loss, Scalable Throughput (L4S) is an existing method for fast congestion indication. It was originally defined for wired networks, but promising results, when applied in radio access networks (RAN), indicate that L4S can be the basis for a common rate adaptation framework in 5G. As part of this framework, using L4S-capable quality-of-service (QoS) flows can be one way to ensure that time-critical high data rate applications work well over 5G and become adopted on a large scale.

Content

Introduction

5G has the potential to enable a wide range of innovative services with time-critical communications for consumers and enterprise customers [9], [10]. A fundamental difference between the emerging time-critical communication applications and traditional MBB lies in reliable low latency (or bounded low latency) [2]. A system designed for MBB maximizes data rates without any guarantees on latency. In contrast, time-critical communication aims for data delivery within specific latency bounds. To ensure bounded low latencies, a system typically compromises on data rates, capacity, or coverage due to the fundamental tradeoffs [2]. Time-critical communication is therefore relevant for scenarios where the need for meeting latency targets is critical to the extent that other user or system characteristics can be compromised. Often, time-critical use cases are also referred to as “latency-critical” use cases.

There is an emerging interest in time-critical use cases in the areas of entertainment/ multimedia, gaming, AR/VR, real-time video conferencing, vehicle to everything (V2X), teleoperated driving, and drones operated through mobile networks. These use cases require bounded low latency in conjunction with medium to high bitrates and the ability to scale across a large number of consumer devices. To meet this demand, entire networks must be optimized. Edge computing reduces transport network latency and jitter by moving data centers from the central cloud into a communication service provider’s network [3], [12]. In addition, 5G’s new radio (NR) network interface introduces key functional and architectural enhancements to enable time-critical services. New, wider spectrum and the possibility for shorter transmission time intervals (TTI) are also two enablers reducing radio interface latency. Service-specific treatments, where (for example) time-critical services are handled differently compared to best-effort services, can be solved with a combination of network slicing [2], [6] and QoS-related features like scheduling and admission control.

Still, these enhancements alone cannot fulfill the goal, and specific functionalities are needed to improve the quality of experience (QoE) for high data rate applications requiring bounded, stable, and low end-to-end latency. This calls for methods to prevent latency spikes due to queuing delays (typically caused by application data rates that are higher than what a network can serve). To fulfill latency requirements, an

application must be able to adapt the bitrate to minimize the risk of increased delay while maximizing the service quality under this constraint.

This white paper describes a concept for such rate adaptation in 5G networks based on L4S, a mechanism being standardized in the Internet Engineering Task Force (IETF). The overall objective is to enable a greater number of time-critical, high data rate applications over 5G mobile networks. This will change 5G as well as the application ecosystem.

Background

Time-critical use cases

There are four fundamental time-critical use case families that are common across various verticals. These consist of real-time media, remote control, industrial control, and mobility automation [2], [9]. These use cases partially have high data rate traffic characteristics and can apply rate adaptation.

With respect to functionality and characteristics, time-critical use cases beyond general real time media streaming can be divided into two general categories. One category of use cases, offloading, is characterized by a shift of computationally intense executions from consumer devices to the edge cloud. The other category, synchronization, is characterized by a time-critical exchange of time synchronous data and information in a distributed system. Use case examples from these categories are illustrated in Figure 1.

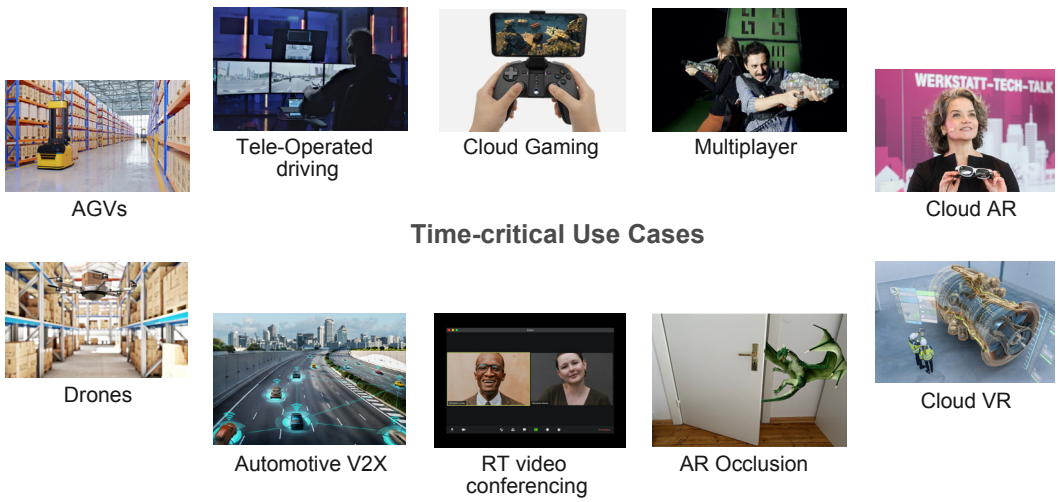

Figure 1. Time-critical use cases

Photo source: Deutsche Telekom (see detailed image sources under references)

The principle of offloading can be well understood in the context of cloud gaming [7]. Consumer devices transmit user input as data over networks to the edge cloud, which hosts the game engines that compute new game scenes. After rendering on the edge cloud GPU, encoded and compressed video streams are sent over networks for display on the consumer devices. The decisive benefit of offloading this processing in cloud gaming is that even high-quality and technically demanding games can be played on lightweight and inexpensive consumer devices. On the other hand, high latency and jitter in networks directly impact the gaming experience by causing lag.

The gain from the offloading principle is also utilized by cloud AR/VR, which enables high-quality graphics on AR/VR glasses, smartphones, and tablets [5]. Other examples that benefit from a less complex consumer device include teleoperated driving, cloud-based automatic guided vehicles (AGVs), and drone control. For these, the camera feed from the vehicle, AGV, or drone [11] is transmitted in the uplink and processed either by a human driver or by an AI system in the edge cloud, sending control information back to the receiving device.

Examples of use cases in the synchronization category include multiplayer games, formation flights of drones, and V2X use cases — such as collision avoidance warning, where a high number of mobile road users (vehicles, bicycles, e-bikes, e-scooters, and so on) on a road segment exchange their positions, speed vectors, and environment models — also require a medium bitrate and, at the same time, low and stable latencies.

Over the top–based rate adaptation

In the baseline rate adaptation method illustrated in Figure 2, the endpoints solve the rate adaptation over the top (OTT) without explicit support from a network. The latency and bitrates are measured by the receiving client and fed back to the sender in an out-of-band protocol. This model is used in (for example) general video streaming services and relies on a large buffer at the client side to give enough time to react to changes in network conditions.

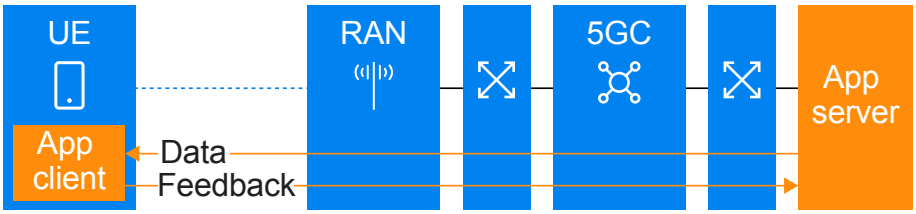

Figure 2. OTT/endpoint-based rate adaptation without explicit network support

Graphic asset: Ericsson

Mobile cloud gaming is an emerging trend [7], which streams real-time, cloud-encoded game video and depends on reliable, high-rate communication with low latency. Existing cloud gaming applications (Google Stadia, Microsoft xCloud, and Nvidia GeForce Now, for example) use end-to-end OTT-based congestion control methods to maintain satisfactory latency. These methods have difficulties performing well using best-effort QoS flows over 5G public networks, where (for example) variations in channel conditions increase the risk that RAN queueing delays will result in latency spikes. To reduce this risk, the RAN must support congestion control for this specific category of services. As the global reach of time-critical applications is an overall goal, the industry must agree on a common approach to handle queueing delays in mobile networks.

Congestion and congestion control

Network congestion occurs when an aggregate of end hosts (for example, servers and clients) transmits at a higher bitrate than what can be sustained by a network. This situation can be avoided if network hosts are prevented from sending more data into a network than what comes out, effectively realized with a transport protocol informing the sender of the amount of data successfully received. One such protocol is the Transmission Control Protocol (TCP), where TCP acknowledge (ACK) packets contain information to allow the sender to transmit more data.

A simple stop and wait procedure that waits for an ACK just to be able to transmit a small amount of data would give poor link throughput, especially when round-trip time (RTT) is high. To solve this problem, transport protocols may have a certain amount of data unacknowledged (also known as having bytes or data “in-flight”), which is controlled by the congestion window (CWND), given by the following relation:
CWND = Bandwidth * RTT

RTT is relatively easy to determine, while the bottleneck bandwidth is typically identified by repeatedly increasing the CWND until packet loss occurs, at which point the congestion window is decreased. The problem lies in that the CWND depends on both bandwidth and RTT. This means that an increased CWND when the bottleneck bandwidth is reached will lead to an increased RTT.

Network buffer overflows will cause packet loss, resulting in poor performance with long end-to-end delays and stalled application behaviors. Implementing delay-sensitive congestion control algorithms at the endpoints will solve part of the delay problem, but it is still hard for an endpoint to know if a delay increase is due to congestion or some other factor characterized by a radio access network, like scheduling jitter, discontinuous reception (DRX), handover, Radio Link Control (RLC) layer retransmission, and so on. Active queue management (AQM) is used to mitigate the problem with large delays. Based on drop thresholds and drop strategies, AQMs selectively drop packets when a network queue delay exceeds a given threshold, implicitly signaling to the endpoints to reduce the number of packets delivered to the network. This avoids the problem with

excessive packet loss and delays, resulting in more stable application behavior. Packet dropping is a minor issue for applications that can tolerate one extra round trip for the retransmission of the missing data or applications that apply forward error correction (FEC).

L4S and ECN

For applications like real-time video, AR/VR, and cloud gaming, the extra delay caused by retransmission or FEC can become noticeable. Explicit congestion notification (ECN) uses the AQM congestion detection method but with explicit signaling to indicate to the end hosts that packets normally dropped are instead marked as congested. ECN uses 2 bits in the IP header, and the information regarding ECN congestion experienced (CE)–marked packets is carried by the TCP ACKs. The end host should then treat a CE-marked packet in the same way as a lost packet.

An ECN-capable AQM marks a packet as CE instead of dropping it when congestion is detected, with a considerable packet loss reduction but less significant latency reduction compared to a packet-dropping AQM.

L4S is an evolution of ECN, where one of the ECN codepoints, ECT(1), is devoted to L4S.

Binary codepoint	Codepoint name	Meaning
00	Non-ECT	Not ECN-capable transport
01	ECT (1)	L4S-capable transport
10	ECT (0)	ECN-capable transport
11	CE	Congestion experienced

The interpretation for an L4S-capable AQM is that it should immediately mark packets with CE when queue delays are very low if the packets are marked as being L4S capable. This gives a prompt reaction to small signs of congestion, allowing the end hosts to implement scalable congestion control. This means that instead of the multiplicative decrease approach used for ECN-capable flows, a scalable congestion control reduces the congestion window (or transmission rate) proportional to the fraction of CE-marked packets. Compared to classic ECN, the L4S approach gives denser congestion events, provides lower delay jitter, and makes it possible to have very low queue delays while maintaining high link utilization.

Network-supported rate adaptation

The default QoS flow, established when the UE attaches to a network, is a permanent QoS flow with non-guaranteed bitrate (non-GBR) used to provide the UE with always-on IP connectivity to the packet data network. This QoS flow has become the generic MBB QoS flow, as it serves most of the mobile broadband applications.

A challenge to support L4S in general internet routers is the requirement to use a separate queue for L4S traffic compared to non-L4S traffic. A separate queue is needed to prevent latency-sensitive L4S packets from queuing up behind a burst of non-L4S packets from an application such as TCP file downloads.

In a 3GPP network, there are separate queues for QoS flows mapped to different RAN radio bearers. By using a dedicated QoS flow for time-critical L4S traffic (as illustrated in Figure 3), it is protected from potential queue buildup on the default QoS flow.

L4S requires rate-adaptive applications, adjusting their bitrates according to rate recommendations from the underlying transport. One way to ensure that time-critical high-rate applications can work well and scale over 5G is to define L4S-capable QoS flows optimized for this application type.

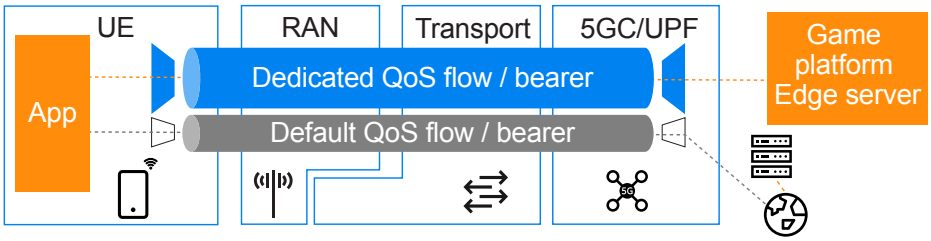

Figure 3. Principle end-to-end solution using a dedicated QoS flow/radio bearer for time-critical L4S traffic.

The establishment of the L4S QoS flow can be triggered upon detection of time-critical traffic in the User Plane Function (UPF) packet filter mechanism. This detection could be based on the L4S codepoints in the IP header, a range of IP addresses (for example, gaming edge servers), a service name indication (application ID), or a combination of those. For applications that have tighter integration with the network for special service levels, the QoS flow may also be set up using a QoS exposure API from the network. The UPF then maps these L4S-capable flows to the latency-optimized, dedicated QoS flow. For example, all L4S flows to/from a range of IP addresses corresponding to a set of edge servers for time-critical applications can be mapped to the latency-optimized QoS flow.

The L4S marking of IP packet headers is used in combination with scalable congestion control algorithms in the applications to minimize queue buildup on the latency-optimized QoS flow. As indicated in Figure 4, any node serving the L4S-capable packet flow will, upon detected congestion, mark (a portion of) the packets in the end-to-end packet flow. The receiving client uses an end-to-end protocol similar to the OTT solution to relay this in-band information to the sender, which adapts the media rate downwards in proportion to the fraction of marked packets.

Edge deployed [3] application servers cater to a small enough transport delay between application client and server, while L4S supports network vendor products to provide early RAN congestion detection and fast rate recommendation to applications. Still, more is needed to enable widespread use of these applications. The ECN bits must be accessible to applications that put demands on UE platforms in order to guarantee transparency. Additionally, the availability of common software solutions for rate adaptation algorithms will speed up the development of well-functioning applications.

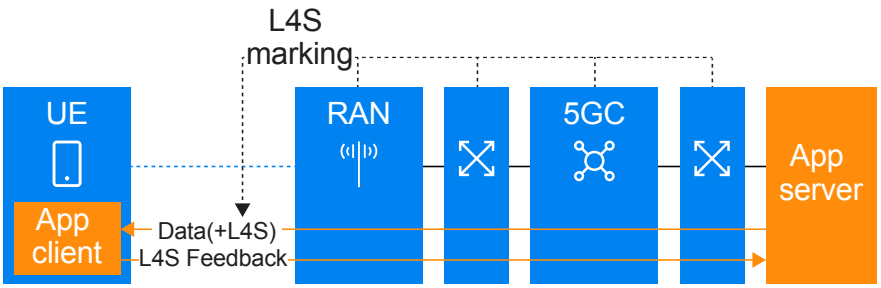

Figure 4. Rate adaptation with L4S-capable networks

Graphic asset: Ericsson

Even though L4S can be introduced in 5G networks without any changes to 3GPP protocols, new 3GPP-defined, latency-optimized 5G QoS Identifier (5QI) coupled with L4S-capable QoS flows would accelerate L4S introduction and alignment between network vendors. These new 5QIs would also enable the activation of latency-specific RAN features beyond L4S, such as latency-optimized scheduler behavior, link adaptation, and handover procedures.

L4S in RAN

Network nodes supporting L4S traffic must mark (CE) packets at early signs of congestion. The congestion marking is started when the queue delay is larger than a defined threshold, and the marking probability is increased linearly as the queue delay increases. As the queue delay target differs between L4S and classic MBB traffic, L4S-capable network nodes (including RAN) require an additional queue using a lower delay threshold (see Figure 5).

Compared to ordinary transport nodes, the RAN preconditions are rather different, as the transport media capacity may suffer from rapid changes due to variations in the radio environment. Thus, in addition to the current queue delay, the channel quality (CQI) is used to estimate the radio interface capacity, giving a more proactive reaction to congestion.

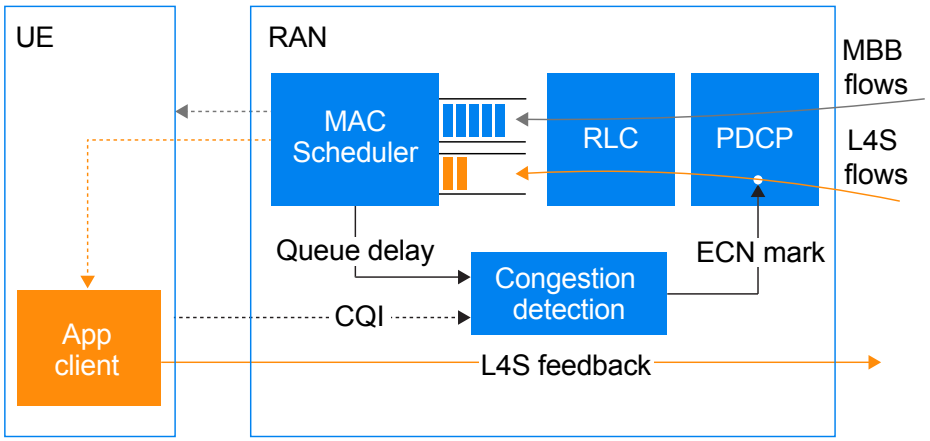

Figure 5. L4S solution in RAN

To optimize RAN for latency-critical, high-rate applications and to satisfy different use cases, rate adaptation using L4S can be complemented with other RAN features. This includes advanced scheduling, QoS, and radio resource partitioning features.

Graphic asset: Ericsson

Performance results

To assess RAN-supported L4S performance aspects, a prototype implementation of L4S in RAN was used as a test object in a lab environment as well as in field test. Both tests were jointly conducted by Ericsson and Deutsche Telekom. In the lab, the application traffic was generated from an experimental software [8] emulating a rate-adaptive, real-time video application transmitting video frames to a UE client software.

Results from the lab tests clearly illustrate the benefits of L4S. Using a fading simulator, programmable attenuators, and background traffic, the test setup can mimic different, realistic channel conditions and traffic scenarios. Figure 6 shows the results from a test where the target device used a default QoS flow and OTT rate adaptation. Background load was generated from an additional UE, emulating YouTube video streaming. The upper part of the figure shows the real-time samples for RTT, queue delay, and bitrate, while the lower part shows the cumulative distribution function for the queue delay and RTT. As shown in the diagrams, the target UE will suffer from large variations in both queue delay and RTT.

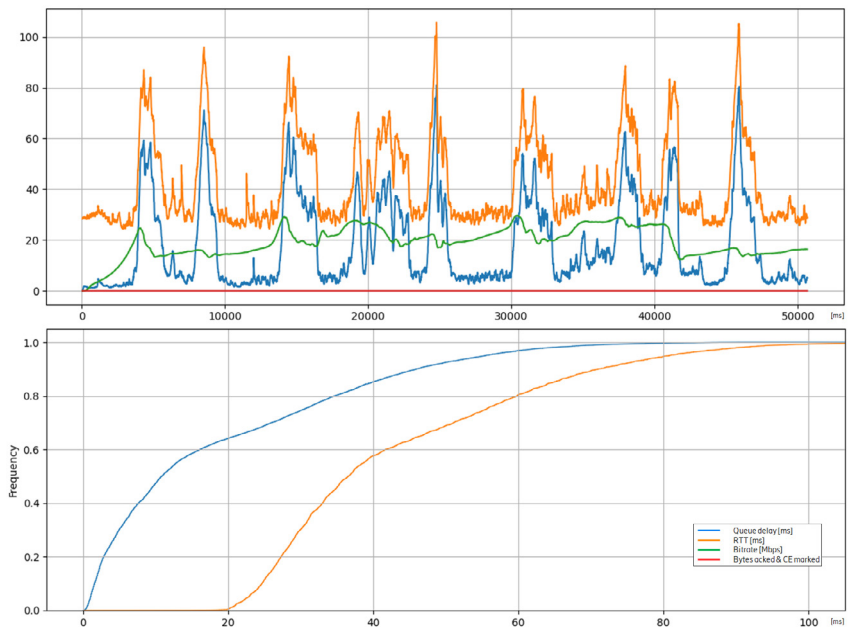

Figure 6. Static users outdoor cell edge: default QoS flow with OTT rate adaptation

Graphic asset: Ericsson

The advantages of network-supported rate adaptation can be seen in the diagrams in Figure 7 below. Here, the test scenario is the same as before, but the target UE used an L4S-enabled, dedicated QoS flow over the RAN. Compared to the previous test, the peak bitrate was lower, while the delay spikes were removed, resulting in stable, low queue delays and RTT.

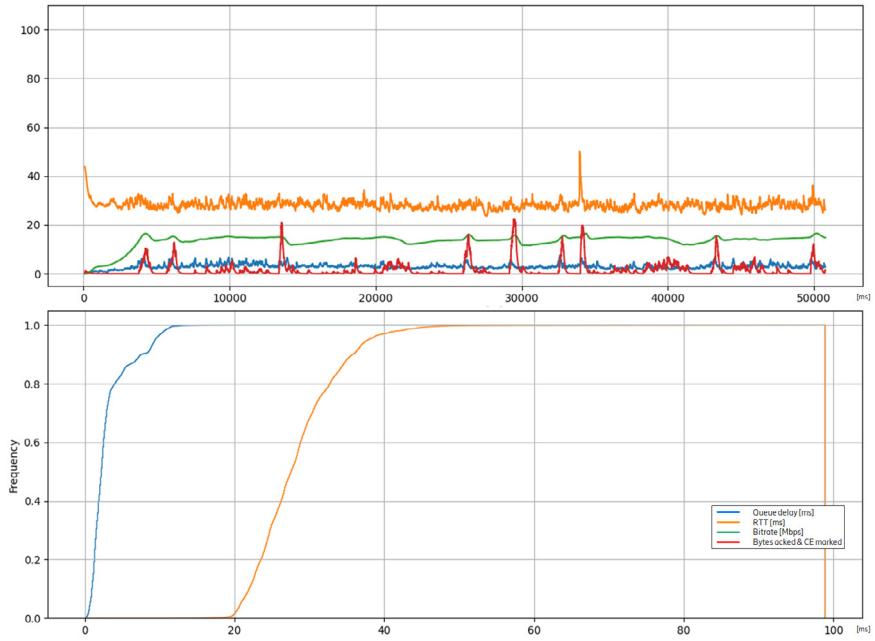

Figure 7. Static users outdoor cell edge: L4S QoS flow and network-supported rate adaptation.

From the early test results, it's evident that an L4S-based congestion detection in the RAN outperforms a corresponding detection in the application client due to the faster discovery of RAN-originated congestions. The early congestion detection and the faster feedback loop enable maintenance of a stable, low, bounded latency for rate-adaptive, high-rate time-critical services.

Graphic asset: Ericsson

Conclusion

There is a great market opportunity for high-rate time-critical services over 5G networks. With emerging edge-enabled use cases (such as cloud gaming, cloud AR/VR, teleoperated driving, V2X, cloud AGVs and more), there is an opportunity for the telecom industry to leverage its network assets to significantly improve the performance of applications. This will result in an improved experience for consumers. The deficiencies of current OTT-based congestion control methods limit the possibility to achieve the latency required to satisfy these applications. It is therefore evident that specific measures are required to overcome the challenges of delivering time-critical services over radio access networks with constrained resources and varying channel conditions.

Results from lab and field tests with the L4S solution clearly indicate that a bounded low latency can be achieved in RAN. The tests also prove the advantage of a network-integrated feedback loop compared to the endpoint-based rate adaptations used in existing OTT solutions.

To reach application scaling gains and global access for high-rate time-critical applications, the industry needs to align the methods for congestion detection, fast feedback loop, and rate adaptation principles. Even though L4S can be implemented using current 3GPP standards, further improvements within 3GPP can help to achieve these goals, reaching widespread availability of high-rate time-critical applications and unlocking the associated use cases.

At a later stage, the implementation of a common 5G rate adaptation framework for high-rate time-critical applications combined with end-to-end network slicing can open up more use cases. Thus, together, these solutions can enable new services and business model innovations that will create new revenue opportunities for communication service providers.

References

1. Harnessing the 5G consumer potential - Ericsson
2. ETR: Critical IoT connectivity: Ideal for time-critical communications
3. ETR: Future of cloud computing: distributed & heterogeneous - Ericsson
4. Gaming: 5 things you need to know about 5G if you're a gamer
5. How 5G and edge computing can enhance virtual reality
6. Network Slicing
7. Mobile cloud gaming
8. SCReAM running code https://github.com/EricssonResearch/scream
9. Ericsson Mobility Report, Nov. 2020
10. Cellular IoT in the 5G era, Ericsson white paper, 2020
11. Edge Computing in 5G for Drone Navigation: What to Offload?
12. How Mobile Edge Computing Will Benefit Network Providers

Figure 1 Deutsche Telekom image source

Multiplayer	iStock.com/ozgurcankaya
AR Occlusion	iStock.com/Vac1
Automotive V2X	iStock.com/Blue Planet Studio
AGVs	iStock.com/Vanit Janthra
Cloud VR	iStock.com/gorodenkoff
Drones	gualtiero boffi//shutterstock.com (https://www.shutterstock.com/de/image-illustration/drone-work-classic-warehouse-3d-image-248134027)
RT Video Conferencing	Photo by Visuals on Unsplash
Cloud AR	Deutsche Telekom (https://www.telekom.com/de/medien/medieninformationen/detail/nur-technik-fuer-menschen-ist-fortschritt-595358)
Cloud Gaming	Deutsche Telekom (https://www.telekom.de/magenta-gaming/controller)
Tele-Operated Driving	Deutsche Telekom (https://www.t-systems.com/de/en/industries/automotive/connected-mobility/teleoperated-driving)

Abbreviations

5QI	5G QoS Identifier
AGV	Automated guided vehicle
AI	Artificial intelligence
AQM	Active queue management
AR	Augmented reality
CE	Congestion experienced
CQI	Channel quality indicator
CWND	Congestion window
DRX	Discontinuous reception
ECN	Explicit congestion notification
ECT	ECN-capable transport
FEC	Forward error correction
GBR	Guaranteed bitrate
IETF	Internet Engineering Task Force
L4S	Low-Latency, Low-Loss, Scalable Throughput
MAC	Medium Access Control
MBB	Mobile broadband
NR	New radio
OTT	Over the top
PDCP	Packet Data Convergence Protocol
QoE	Quality of experience
QoS	Quality of service
RAN	Radio access network
RLC	Radio link control
RTT	Round-trip time
SCReAM	Self-Clocked Rate Adaptation for Multimedia
TTI	Transmission time interval
UE	User equipment
UPF	User Plane Function
VR	Virtual reality

Authors

Per Willars is a Senior Expert in radio network functionality and end-to-end architecture at Business Area Networks. He joined Ericsson in 1991 and has worked intensively with RAN issues ever since. This includes 3G RAN, QoS, indoor solutions, and 5G eco-system alignment. He has also worked with service layer research and explored new business models. In his current role, he analyzes the requirements on 5G RAN architecture and functionality. Per holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology.

Emma Wittenmark joined Ericsson in 1998 and has worked with research and development of the 2G, 3G, 4G, and 5G mobile network generations, first on the UE modem side and since 2014 as part of Business Area Networks. Her current focus is on architecture and deployment aspects for the 5G RAN and 3GPP RAN1 support work. Emma holds an M.Sc. in electrical engineering and a Ph.D. in Information Theory from the Faculty of Engineering at Lund University, Sweden.

Henrik Ronkainen joined Ericsson in 1989 to work with software development and later became a software and system architect for the 2G and 3G RAN systems. With the introduction of high speed downlink packet access, he worked as a system architect for 3G and 4G UE modems. Ronkainen currently serves as a system researcher at Business Area Networks, where his work focuses on analysis and solutions related to the architecture, deployment, and functionality required by 5G RAN. He holds a B.Sc. in electrical engineering from Lund University in Sweden.

Christer Östberg is an expert in the physical layer of radio access at Business Area Networks, where he is currently focusing on analysis and solutions related to the architecture, deployment, and functionality required by 5G RAN. After first joining Ericsson in 1997 to work with algorithm development, he later became a system architect responsible for the modem parts of 3G and 4G UE platforms. Östberg holds an M.Sc. in electrical engineering from Lund University.

Ingemar Johansson joined Ericsson in 1992 to work on speech compression algorithms for cellular access. In the current role, he works in the areas of transport protocol development, the interaction between cellular access and applications, congestion control, and video streaming. His special interest is in low latency congestion control for interactive applications such as gaming, XR, and video-assisted remote control. He is an active contributor to the work on L4S in IETF. Ingemar holds an M.Sc. in computer engineering from the Luleå University of Technology.

Johan Strand started at Ericsson in 2001, pioneering the implementation and systemization of multimedia functionality on mobile platforms. Later on, Johan focused on mobile platform systemization and has held several technical and leading positions within the mobile platform and 3G/4G UE modem projects. Currently, Johan is a System Developer at Business Area Networks focusing on E2E aspects, system design, and performance evaluation. Johan holds an M.Sc. in electrical engineering from Lund University.

Petr Lédl is VP of Network Trials and Integration Lab and chief architect of the Access Disaggregation program. His main responsibility is identification, prototyping, and technical evaluation of future mobile network technologies and related concepts (mainly 5G, LTE-A/Pro, Managed connectivity, etc.) and defining and implementing DTAG disaggregated RAN architecture based on O-RAN principles.

He studied Electrical Engineering at Czech Technical University in Prague between 1995 and 2004. Since 2003 he has been working in different positions in Technology organizations of Deutsche Telekom and T-Mobile Czech Republic.

Dominik Schnieders is the Head of Edge Computing and Innovation Lead at Deutsche Telekom focusing on the enablement of latency-critical applications in telecommunication networks.

Previously, he was the program manager for the development and rollout of narrowband-IoT and headed the product development of an IoT platform. Before joining Deutsche Telekom in 2014, Dominik worked 16 years in the automotive industry for BMW where he held different management positions. Dominik holds an M.Sc. in electrical engineering as well as in industrial engineering from the Technical University of Munich.