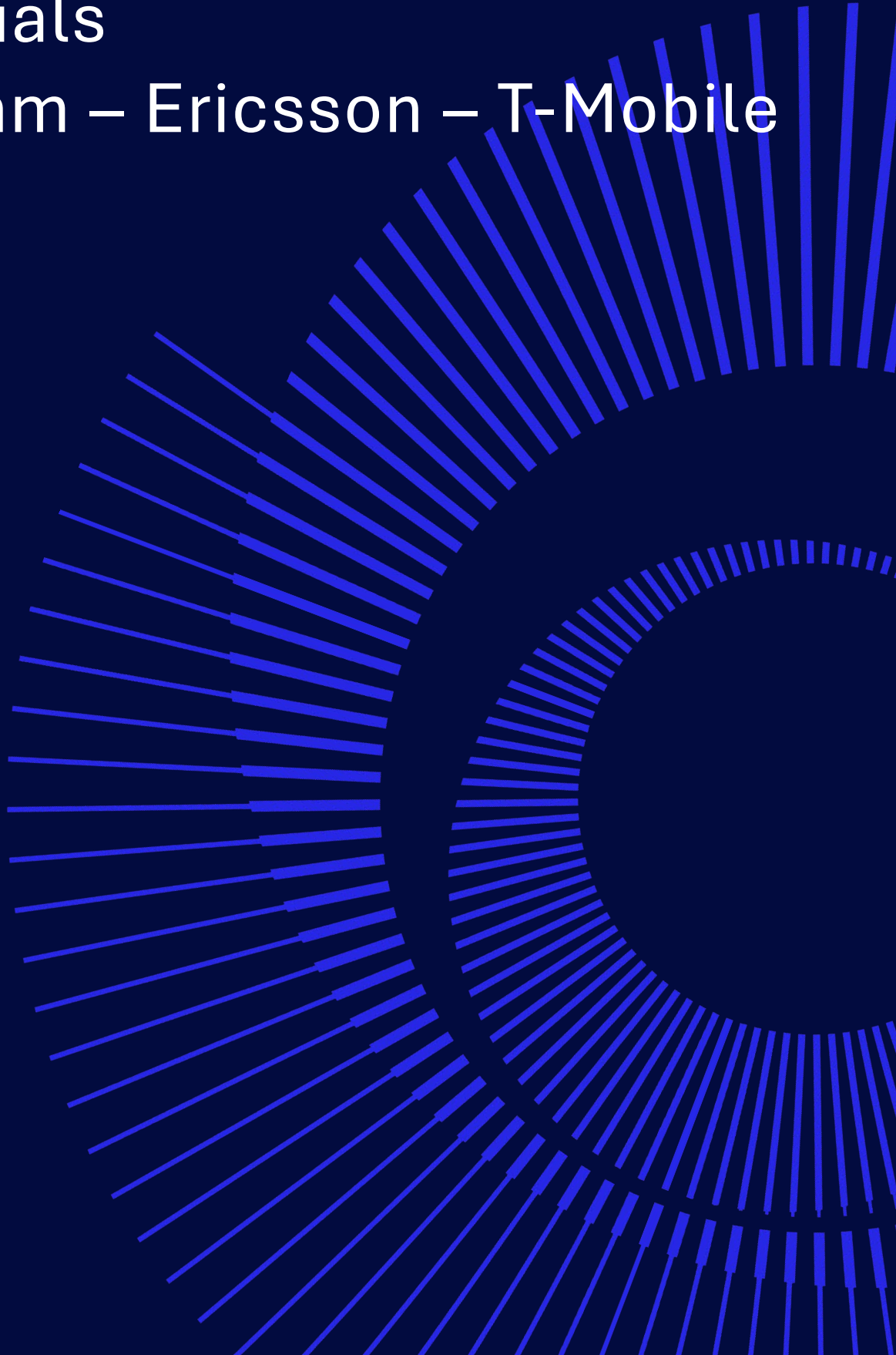


5G XR Trials

Qualcomm – Ericsson – T-Mobile

December 16, 2025



Contents

Introduction	4
5G XR Trial Overview	5
5G Advanced Features	5
Low Latency, Low Loss, Scalable Throughput	5
Uplink Configured Grant	6
On-device Features.....	7
Distributed Spatial Compute	7
5G Modem Low Latency Processing	10
5G XR Trial Details	11
Test Setup	11
Key Performance Indicators (KPIs)	12
Trial Results	13
Baseline Performance	13
Performance with L4S and DSC	13
Performance with UL Configured Grant	15
5G Modem Low Latency Processing Performance	15
Conclusion	17

Introduction

The advent of Extended Reality (XR)—encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR)—is reshaping the way humans interact with digital environments. These immersive technologies demand high amounts of processing and compute capabilities to deliver compelling, responsive experiences. As XR technologies evolve from isolated, tethered setups to wirelessly connected, on-the-go multi-user environments, the need for low latency, high-bandwidth, and reliable connectivity becomes paramount.

The rollout of 5G Standalone (SA) networks introduces transformative capabilities that directly address the latency and bandwidth needs of XR applications.

To validate XR-over-5G technology readiness, we deployed an XR distributed compute system in a commercial state-of-the-art T-Mobile 5G SA network with Ericsson infrastructure and Qualcomm 5G Mobile Platform connected to AR glasses over Wi-Fi. On the network side, T-Mobile deployed 5G Advanced features such as 3GPP Release 18 Low Latency, Low Loss, Scalable Throughput (L4S) and Release 16 Uplink Configured Grant (UL CG). Further, the Qualcomm devices implement 5G modem low latency processing and use on-device intelligence and AI to distribute the XR workloads between a network-edge compute node and the device. Putting together these network and device-side features, we showcase an end-to-end 5G XR system that is highly scalable and performant even in challenging radio and network conditions.

The rest of this whitepaper is organized as follows: in the next section we provide further details on the 5G Advanced and on-device features that are used in these 5G XR trials. In the subsequent section, we discuss the end-to-end system setup, followed by the trial performance results. Finally, we end with some concluding remarks.

5G XR Trial Overview

5G Advanced Features

Low Latency, Low Loss, Scalable Throughput

Low Latency, Low Loss, Scalable Throughput (L4S)¹, is a networking technology designed to reduce network buffer build-up through application rate adaptation using network congestion feedback and thereby achieving lower stable latency while maintaining highest possible throughput and low packet loss. It is part of a broader effort to modernize congestion control mechanisms on the Internet, particularly for real-time applications like video calling/streaming, cloud gaming and XR.

Traditional congestion control methods typically rely on packet loss and round-trip time (RTT) as signals to adapt transmission rates, making them slow to react to congestion events as these metrics are inherently only observable in a delayed fashion. With L4S, by contrast, the gNodeB is able to detect OTA congestion in real-time, and provide feedback via Explicit Congestion Notification (ECN) to an on-device L4S-compatible congestion controller that responds in a standardized way (following the Prague requirements²) to these early congestion signals, before packet loss occurs.

Further, L4S uses traffic segregation to enable prioritized low-latency service for applications that support it, without negatively impacting legacy traffic. The result is a more responsive internet experience, especially under high load conditions. L4S is being standardized by the IETF and is already seeing early adoption in some networks and operating systems.

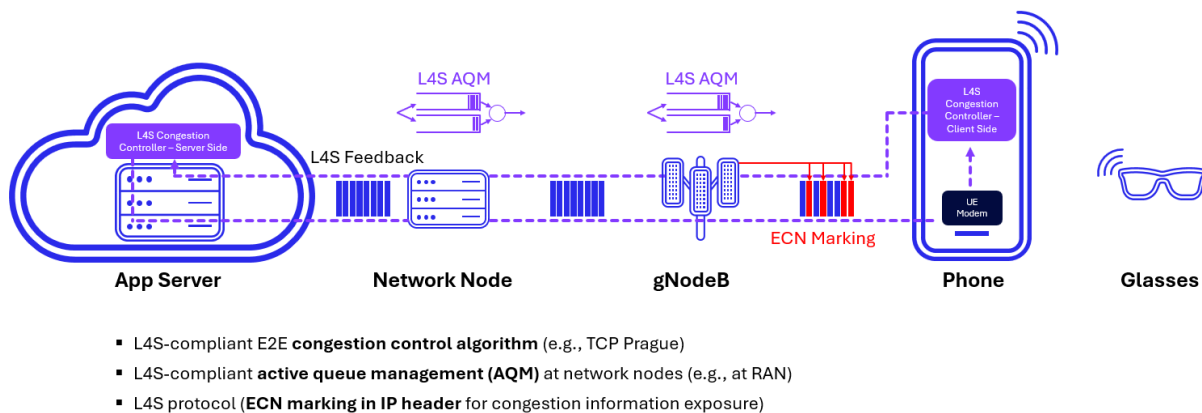


FIGURE 1: ARCHITECTURAL COMPONENTS OF L4S

¹ <https://datatracker.ietf.org/doc/rfc9330/>

² <https://www.ietf.org/archive/id/draft-briscoe-iccrp-prague-congestion-control-03.html>

Uplink Configured Grant

Uplink Configured Grant (UL CG) is a feature in 5G Advanced to reduce medium access delay and signaling overhead associated with uplink transmissions. This is particularly useful for applications that need to transmit data periodically on the UL and with low delay such as XR.

In traditional 5G medium access for uplink, the device must first send a scheduling request (SR) to the 5G base station (gNodeB) and wait for an uplink (UL) grant. In turn, the uplink grant is used by the device to transmit a buffer status report (BSR) indicating the amount of data awaiting transmission to the network. The device may also transmit data alongside BSR depending on the size of the uplink grant. The base-station can then schedule the device with a properly sized grant to transmit the remaining data.

With UL CG, the 5G base-station pre-allocates uplink resources to the device, allowing it to transmit data at scheduled intervals without needing to request resources each time. As we will see later, this is suitable for distributed compute XR applications since the pose of the user needs to be periodically transmitted on the UL to the network-edge compute node. The figure below depicts the reduction in latency of data transmission as well as lower signaling overhead associated with multiple UL grant allocations enabled by UL CG.

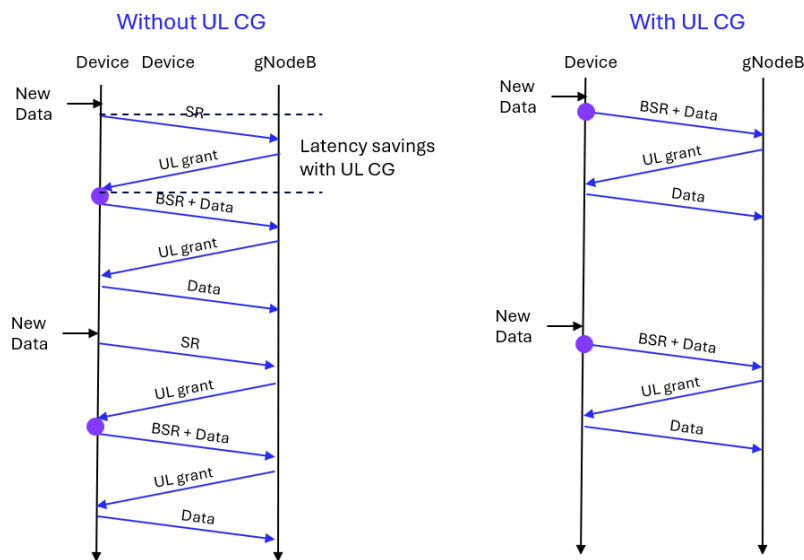


FIGURE 2: 5G MEDIUM ACCESS PROCEDURE WITH AND WITHOUT UL CG

For this XR trial, the above two features were implemented using a network slicing framework, deployed on T-Mobile's 5G SA network, so only applications such as XR can take advantage of these costly features.

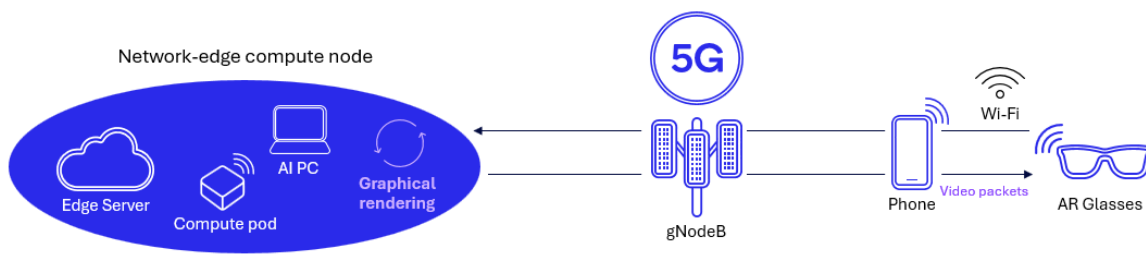
On-device Features

Distributed Spatial Compute

A typical AR application involves various compute workloads, including rendering of graphics to be augmented on the user's field of view. The graphics rendering task can be executed either on the device or at a network-edge compute node. Distributed Spatial Compute (DSC) is a framework which enables intelligent offloading of XR workloads, such as graphics rendering, between the device and network-edge compute node. We showcase DSC using Phone-to-AR Glass (P2G) architecture with a network-edge compute node as depicted in Figure 3.

In network-edge rendering, the graphics rendering is executed at a network-edge compute node. The user's uplink head pose information flows from the AR glasses to the phone over Wi-Fi, and from the phone to the network-edge compute node over 5G. The network-edge compute node processes the data and sends back the downlink encoded graphics via the phone to the glasses using a high bandwidth 5G link. Network-edge rendering can enable higher-quality graphics and lower the phone's power consumption. However, network-edge rendering adds more load on the 5G network, needing high bandwidth and low latency, which can sometime be a challenge in wireless network given the uncertainty and temporal variability of the wireless link signal quality and network load.

Network-edge Rendering



On-device Rendering

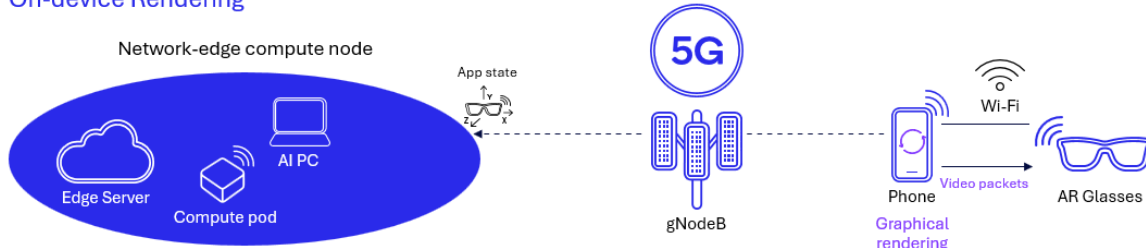


FIGURE 3: NETWORK-EDGE RENDERING VS. ON-DEVICE RENDERING

In contrast, for on-device rendering, the graphics rendering is performed locally on the phone. The user's uplink head pose information flows from the AR glasses to the phone and the rendered graphics are streamed back from the phone over Wi-Fi. The 5G link is used for low bandwidth communication of the app state information between the phone and the network edge. Compared to network-edge rendering,

on-device rendering provides higher service availability, as well as enabling seamless and ubiquitous AR experience in adverse radio and/or network conditions, at the cost of increased power consumption.

DSC uses on-device intelligence to provide a recommendation between network-edge rendering and on-device rendering to the application client via an on-device DSC API. The application client then interacts with the network-edge compute node to dynamically perform the rendering switch. The DSC recommendation is based on the observed environment and AI-based modem predictions of radio and network conditions, compute availability and device power considerations.

Table 1 lists the application traffic profile in network-edge rendering and on-device rendering mode. It is worth noting that during on-device rendering, the application also sends periodic probe to evaluate radio and network conditions.

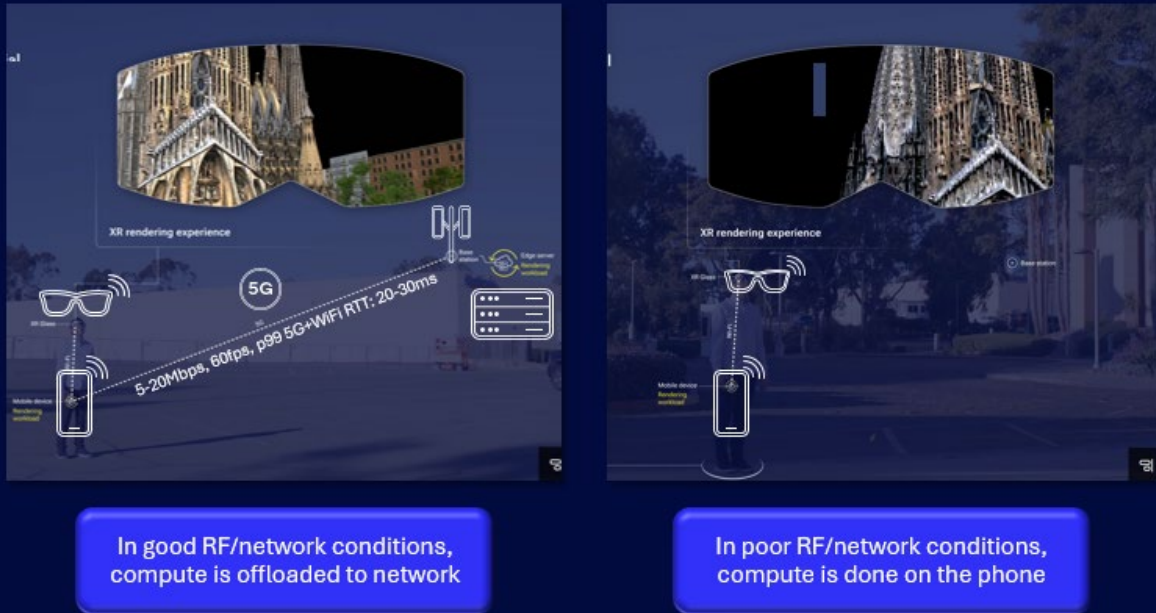
TABLE 1: AR APPLICATION TRAFFIC PROFILE

Traffic	Network-edge Rendering	On-device Rendering
Downlink	5-20 Mbps (video frames)	< 1 Mbps + probes
Uplink	< 1 Mbps (pose)	< 1 Mbps (app state)

The figure below summarizes the overall value proposition for DSC.

Distributed Spatial Compute for XR Rendering

DSC API on phone enables app to toggle between network-edge & on-device rendering



Value Propositions

Operators: Higher service availability vs. network-edge only rendering

Users: Phone power savings, improved user experience vs. on-device only rendering

FIGURE 4: DISTRIBUTED SPATIAL COMPUTE FOR 5G XR TRIALS

DSC AND L4S INTERACTION

L4S enables responsive adaptive rate control (ARC) at the application by providing early, frequent and explicit congestion signals from the network. Application ARC can use these signals to determine the appropriate bitrate for transmission to maintain low end-to-end latency which is critical for good user experience (UX). On the other hand, DSC enables toggling the rendering compute location based on the observed RF environment.

The interaction between DSC and ARC can be qualified by means of user experience vs. bitrate curve as shown in the figure below. The user experience in network-edge rendering mode is a function of the application bitrate which is controlled by the ARC algorithm. User experience improves with increasing bitrate. On the other hand, user experience for on-device rendering is, to a large extent, not dependent on the link rate. This establishes a minimum bitrate threshold T_b , above which the system should operate in network-edge rendering mode with ARC, and on-device rendering mode otherwise. The threshold T_b is typically content dependent and is known to the application.

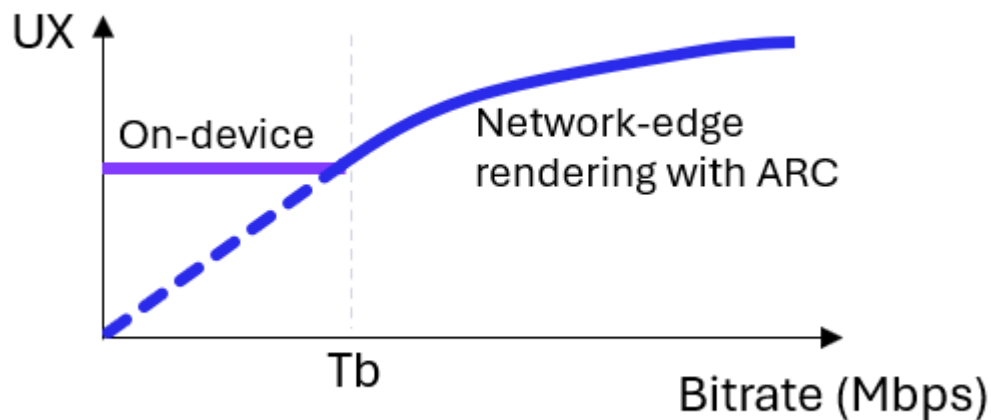


FIGURE 5: INTERACTION BETWEEN DSC AND ARC

The bitrate threshold (T_b) was set to 5 Mbps for the AR application used in our trial.

5G Modem Low Latency Processing

5G Modem Low Latency Processing is a suite of device-only features and modem APIs to support low latency application requirements. The features and APIs enable proprietary device optimizations such as traffic prioritization across application flows, low-latency and pipelined processing across the 5G protocol stack and significantly reduce latency spikes, all of which contributes to improved user experience.

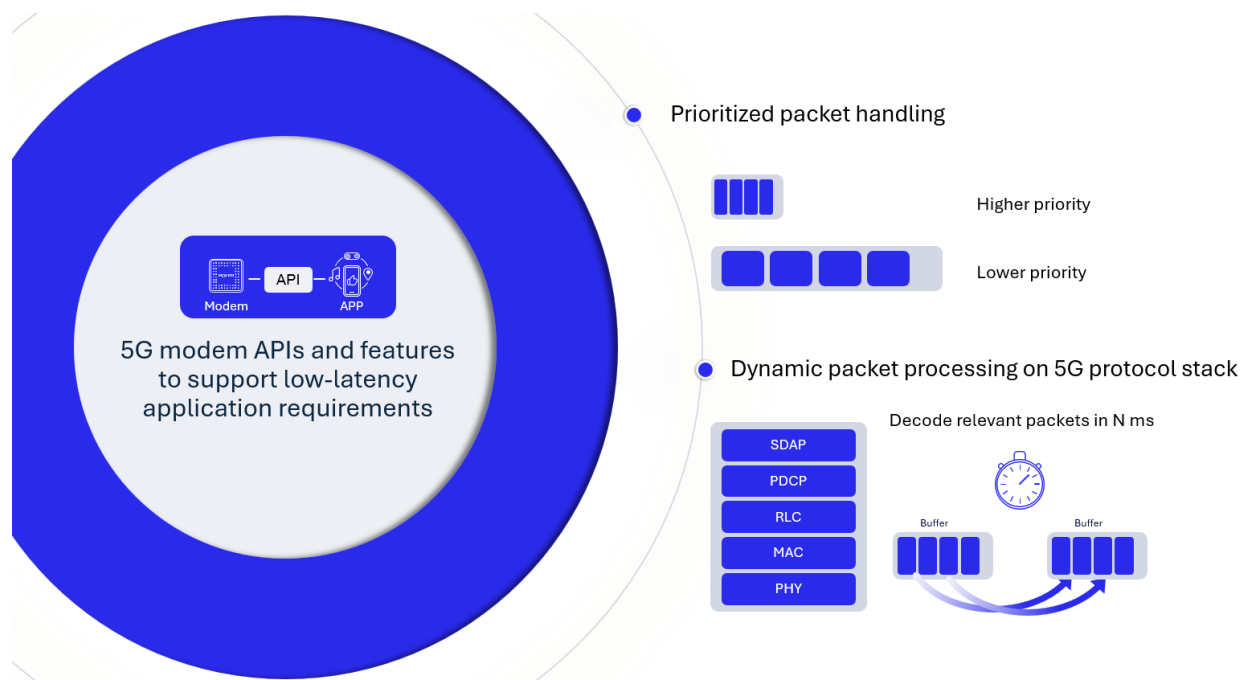


FIGURE 6: 5G MODEM LOW LATENCY PROCESSING

5G XR Trial Details

Test Setup

The end-to-end system trial was conducted in a T-Mobile 5G Standalone (SA) commercial network in the San Diego area, using Ericsson infrastructure and 5G Advanced software.

The test setup consisted of one device-under-test (DUT) user experiencing the AR application, carrying a reference phone, powered by Snapdragon Mobile Platform, and a pair of reference AR glasses powered by Snapdragon AR2 Gen1 Platform. We refer to this user interchangeably as the AR user.

In addition to the DUT user, there are users which are used to generate iPerf traffic to load the network and create adverse conditions for the DUT user. Since the tests were conducted in a live commercial network, there are also other commercial users generating data traffic on the network.

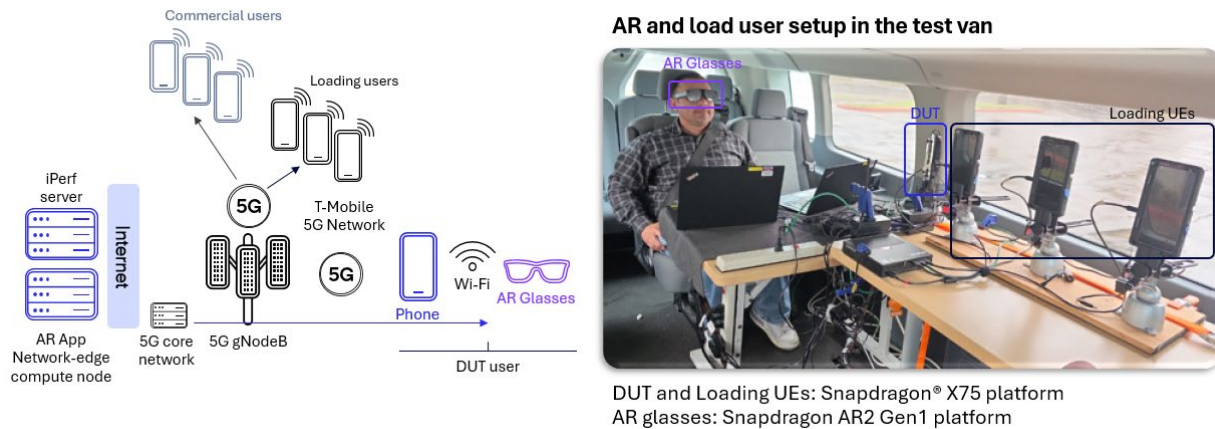


FIGURE 7: TEST SETUP FOR 5G XR TRIALS

Two test locations are chosen for carrying out the trial experiments. A near-cell location to capture performance in good network and radio conditions, and a far-cell location representing more challenging scenarios.

TABLE 2: 5G XR TRIALS TEST LOCATIONS

Location	Near cell	Far cell
Distance from cell-site	~1000 ft	~3000 ft
RSRP	-80 dBm	-105 dBm
DMRS SNR	[19, 30] dB	[3, 12] dB

Key Performance Indicators (KPIs)

To quantify the performance, we captured various application and link level metrics. For purposes of this whitepaper, the most relevant KPIs are the following:

- **Live AR experience:** The AR experience of the DUT user captured via screen recording
Note: This document does not include videos of the live AR experience due to size limitations
- **RTT latency:** Round-trip time (RTT) latency of DUT user, measured from the time of head pose transmission from the phone to network-edge compute node plus the time for video frame transmission from network-edge compute node to the phone
Note: Due to setup limitations the RTT latency measured in the tests also includes latency from packet processing in the WebRTC stack of AR application on the phone
- **5G downlink throughput:** 5G downlink physical layer throughput of DUT user

- **Target bitrate (TBR):** Bitrate output of ARC; used by network-edge compute node to encode DL video frames for AR user
- **ECN marking rate:** ECN marking rate of the downlink AR user traffic. Measured as percentage of packets marked with ECN congestion bit set, over total number of packets in the last report
- **5G uplink scheduling rate:** Percentage of 5G uplink slots used for transmission by the DUT user

Trial Results

Baseline Performance

To characterize baseline performance, we tested the AR user in near and far cell locations with loading users turned OFF, i.e., not performing any data activity. In both locations, the RTT latency was observed to be within acceptable limits and the live AR user experience was good. The system operates in network-edge rendering mode and the target bitrate does not change.

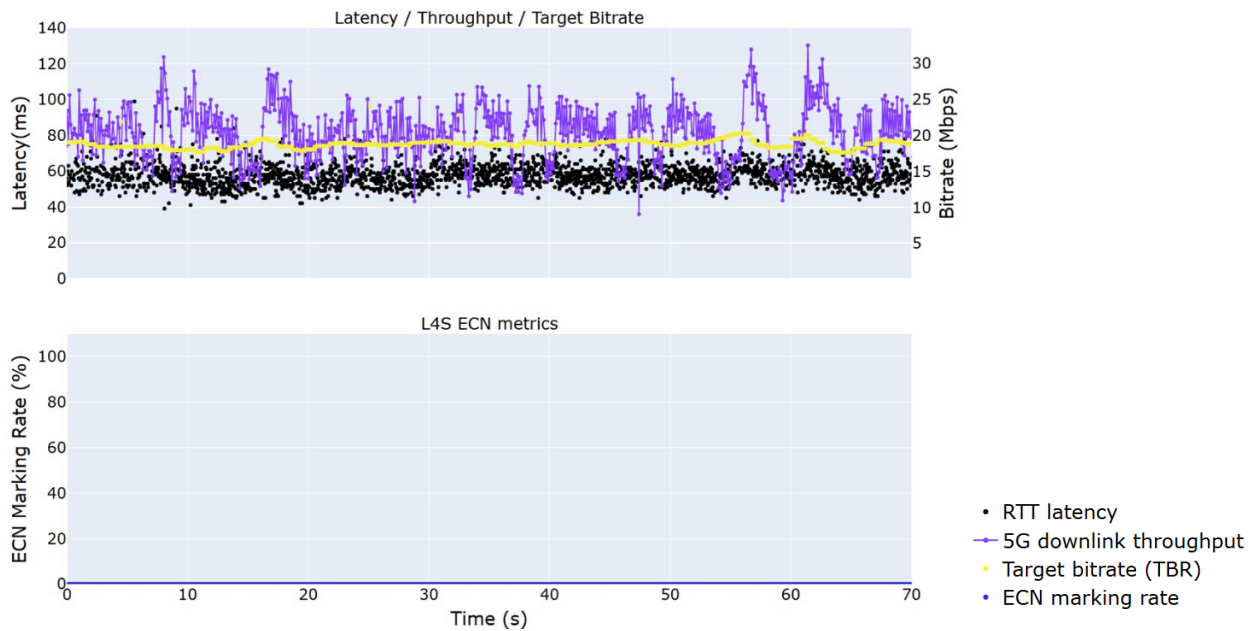


FIGURE 8: BASELINE PERFORMANCE IN NO LOAD CONDITIONS

Performance with L4S and DSC

We observe that once loading users are turned ON the application starts adapting its bitrate with L4S ARC. The 5G downlink throughput is approx. 20 Mbps before load is introduced. As successive load users are turned ON, we observe a spike in the ECN marking rate and corresponding reduction in the target bitrate. Note that the RTT latency is kept within acceptable limits indicating that L4S based ARC is

performing well as can be seen in Figure 10. On the other hand, without L4S ARC we observe latency spikes as there is no timely adjustment of application bitrate in response to congestion in the network.

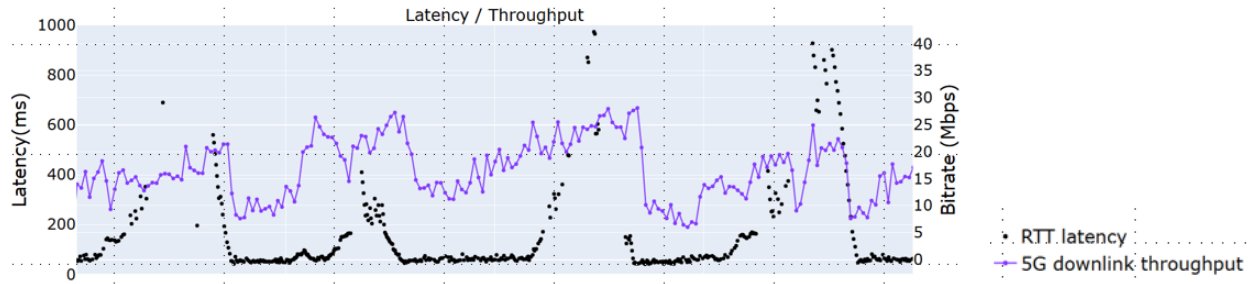


FIGURE 9: PERFORMANCE WITHOUT ARC IN FAR CELL WITH MEDIUM LOADING

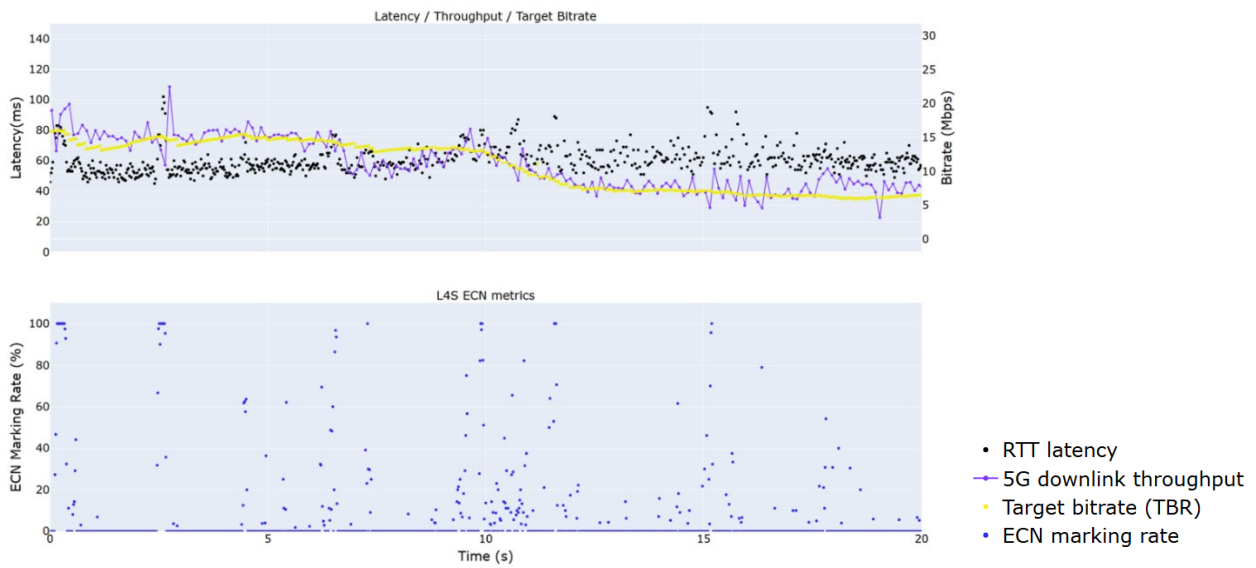


FIGURE 10: ARC PERFORMANCE IN FAR CELL WITH MEDIUM LOADING

As we increase the loading beyond three load users, the target bitrate continues to move lower due to lower available capacity. We observe ECN congestion marking spikes correlated with introduction of each of the loading users. Once the application specified minimum target bitrate T_b ($= 5\text{Mbps}$) cannot be sustained, the system switches to on-device rendering based on DSC recommendation. We observe good user experience throughout the test run with seamless switch from network-edge rendering to on-device rendering.

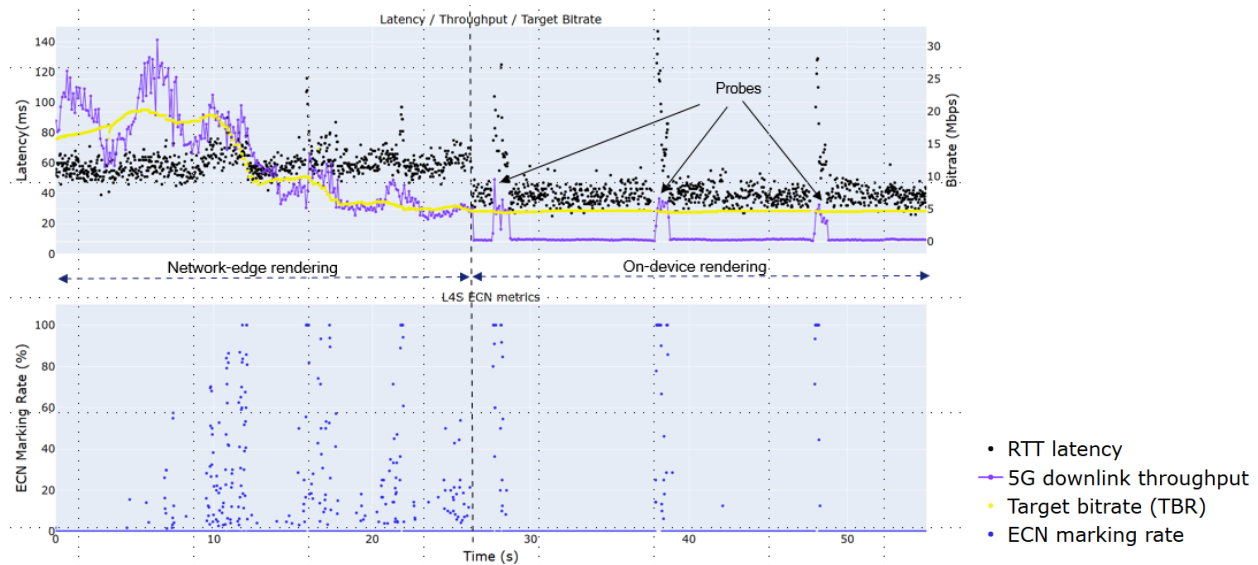


FIGURE 11: DSC PERFORMANCE IN FAR CELL WITH HIGH LOADING

When in on-device rendering mode, we can see periodic application probes being transmitted.

Performance with UL Configured Grant

The performance with UL CG is shown in the table below. As expected, we see a reduction in RTT latency with UL CG – up to 10ms at 90th percentile and 19ms at 99th percentile. However, the reduction in latency comes at the cost of higher 5G uplink scheduling rate (fraction of Resource Blocks allocated to the UE over time and frequency). This can be attributed to uplink grants being more frequently available to the device with UL CG, thereby affording less opportunities for separate application data bursts to be accumulated in the device’s buffer and being transmitted together in a single shot. The tradeoff between latency reduction and frequent uplink transmissions can be controlled by setting the periodicity of UL CG appropriately. In these tests, UL CG periodicity was set to 10ms.

TABLE 3: PERFORMANCE WITH UL CG, NEAR CELL, NO LOADING

Scenario	RTT latency (90 th , 99 th) percentile	5G uplink scheduling rate
Without UL CG	70 ms, 84 ms	~ 15%
With UL CG	60 ms, 65 ms	~ 34%

5G Modem Low Latency Processing Performance

To evaluate the performance with 5G Modem Low Latency Processing, the DUT is placed in far-cell location, and the loading users are turned OFF. The 5G link throughput and RTT latency is per expectation and AR user experience is good.

Next, we stress the 5G uplink by introducing background uplink iPerf traffic on the DUT, which mimics file uploads in practice. For this test we assume no differential traffic treatment in the network between traffic

from different applications on the DUT. We notice large latency spikes once iPerf traffic is started. As a result, the user observes image freezes and the application is unusable.

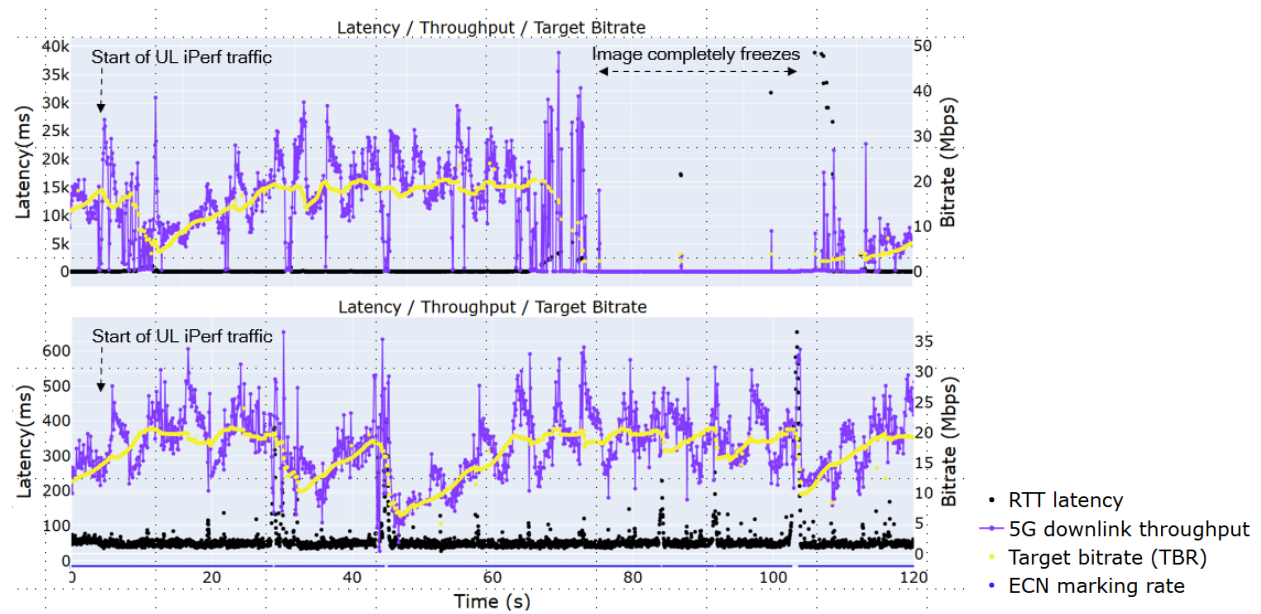


FIGURE 12: PERFORMANCE WITHOUT (TOP) AND WITH (BOTTOM) 5G MODEM LOW LATENCY PROCESSING

Under the same conditions, we see improved performance with 5G Modem Low Latency Processing turned ON. The RTT latency is maintained even after iPerf traffic is introduced, there are no image freezes, and the AR user experience is good.

Conclusion

XR technologies are evolving from isolated, tethered setup to mobile and on-the-go systems. At the same time XR experiences are becoming richer and more immersive. Supporting such experiences over a wide area cellular deployment brings forth challenges associated with variability and unpredictability of wireless radio and network conditions.

To validate network readiness of supporting XR experiences over Ericsson 5G infrastructure, we have conducted trials with an end-to-end 5G XR system over a commercial T-Mobile 5G SA network and AR smart-glass slice. We have shown how 5G Advanced features such as L4S and UL CG, and on-device features such as DSC and 5G Modem Low Latency Processing, can enable distributing the XR rendering workload to a network-edge compute node over 5G. We showcase a highly scalable and seamless distributed compute AR experience that is performant even in adverse radio and network conditions.

Looking forward, we expect low-latency XR service over 5G SA to be a foundational enabler of seamless wide area XR experiences. The applications of such service extend to GenAI and other use cases natively supported in XR devices. As the ecosystem matures, we believe this will be essential to unlocking XR's full potential and ensuring that immersive experiences are accessible, scalable, and future-ready.

2025



** This page is intentionally left blank to maintain the layout of the document. Deleting this page may cause misalignment of elements. Please ensure it remains blank when printing the final document.

Do not print this page. **