GTI
5G/5G-A Key Technology and Tendency: A Chipset View

https://gtigroup.org/

GTI 5G/5G-A Key Technology and Tendency: A Chipset View

Version:	v_1.0
Deliverable Type	□ **Procedural Document** ☑ **Working Document**
Confidential Level	□ **Open to GTI Operator Members** □ **Open to GTI Partners** ☑ **Open to Public**
Program	5G Technology and Program
Working Group	
Project	
Task	
Source members	MediaTek, ASR, Qualcomm, UNISOC
Support members	
Editor	Yunqi.Wang, Di Long, Ting.Wang, Hualei.Wang
Last Edit Date	2026.01.23
Approval Date	

Document History

Date	Meeting #	Version #	Revision Contents

Table of Contents

1 Executive Summary

With growing 5G coverage and increasing SA penetration, key evolution directions include two of particular importance for devices: improving in performance and power efficiency, and diversification of device and link portfolios.

This whitepaper highlights Release 17/18 features that deliver near-term value on today's networks and devices. We focus on features with clear, observed benefits of performance, including UE power-saving features, UL Tx switching, 3TX, Dynamic Waveform Switching, 1024QAM, L1/L2-triggered mobility, L4S, UE aggregation, Redcap/eRedCap and NTN.

For each feature, we turn the spec into clear KPIs, validate readiness on commercial chipsets or prototypes, and quantify the benefits. We also give deployment and configuration guidance so operators can deliver high quality at scale and improve efficiency and we highlight the ecosystem triggers—network R17/18 SA feature rollouts, spectrum refarming, and device profiles—that will shape chipset roadmaps.

2 Abbreviations

Abbreviation	Explanation
BLER	Block Error Rate
BWP	Bandwidth Part
CA	Carrier Aggregation
CC	Component Carrier
CHO	Conditional Handover
CPE	Customer Premise Equipment
CSI	Channel State Information
CU	Centralized Unit
DCI	Downlink Control Information
DRB	(user) Data Radio Bearer
DRX	Discontinuous Reception
DU	Distributed Unit
EN-DC	E-UTRA NR Dual Connectivity with E-UTRA connected to EPC
FWA	Fixed Wireless Access
L4S	Low Latency, Low Loss, and Scalable Throughput
LEO	Low Earth Orbit
MEO	Middle Earth Orbit
NTN	Non-Terrestrial Network
PCell	Primary Cell
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PEI	Paging Early Indication
PMO	PDCCH monitoring occasion
QoE	Quality of Experience

RLC	Radio Link Control
RLF	Radio Link Failure
RTT	Round Trip Time
SA	Standalone
SAR	Specific Absorption Rate
SCell	Secondary Cell
SUL	Supplementary Uplink
TN	Terrestrial Network
TRP	Transmit/Receive Point
TX	Transmitter
U2N	UE-to-Network
UE	User Equipment
UL	Uplink
VoNR	Voice over NR
VR	Virtual Reality

3 Introduction

5G's initial phase is largely complete. Baseline coverage has been established using a mix of low-band FDD and mid-band TDD, with wide-area mid-band overlays and ongoing densification continuing to lift capacity. The ecosystem has matured across mainstream FR1 bands and device categories, and multi-band NR carrier aggregation and VoNR are now common in new models. Standalone (SA) coverage and device penetration continue to rise, underpinned by steady growth in public SA deployments and a broad FR1 device portfolio.

Building on this foundation, 3GPP continues to evolve: Release 17 expands device and access diversity, improves power efficiency, and strengthens overall network capabilities and reliability. Release 18—the first iteration of 5G-Advanced—advances 5G on multiple fronts: performance and predictability; access to new market segments; sustainability and efficiency at scale; and AI-driven network automation, while further improving mechanisms for uplink, mobility, and low-latency operation. Although 3GPP is already moving into Release 20, but most commercial networks and devices are effectively at R17/18 now. That gap is normal: chipset and device cycles and network software timelines mean features reach the field later than they reach the standard.

From the device and chipset perspective, two evolution directions are of particular importance: (1) uplifting performance and power efficiency, and (2) diversifying device and link portfolios. These directions shape chipset roadmaps and device design choices as OEMs balance RF, thermal, and form-factor constraints while pursuing a premium user experience.

Aligned with these directions, this white paper spotlights a set of near term, high impact features anchored in Release 17/18:

To uplift performance and power efficiency:

- Uplink enhancements (e.g., UL Tx switching and 3TX) to raise user and cell level capacity;
- Dynamic waveform switching to extend the UL coverage;
- 1024QAM to raise peak throughput and spectral efficiency on high-SINR link;
- L4S for consistently low queueing delay and jitter;
- L1/L2 triggered mobility to shorten interruption time and stabilize;
- And low power features to extend battery life.

To diversify device and link portfolios:

- UE aggregation to augment uplink via inter-UE connection;
- RedCap/eRedCap to expand lightweight device classes;
- And NTN to extend direct to device reach across land, sea, and air.

For each feature, we translate standards into measurable KPI improvements, map device capabilities and readiness, and set indicative targets and validation methods. Validated test results confirm that these features are supported by current chipset capabilities, and we provide deployment and configuration guidance to help operators deliver quality at scale while improving efficiency.

The remainder of this white paper is organized as follows:

- Section 1: A high-level overview of 3GPP Release 17 and Release 18.
- Section 2: Deep dives into these high impact feature sets.

- Section 3: A synthesis of the paper, presenting chipset-validated gains on key features and calling for industry convergence.

4 Big picture of R17 and 5G-Advanced

As 5G continues to evolve, each 3GPP release marks a significant step forward in both technology and market impact. As depicted in below Figure 1,following the foundational work of Release 15 and the industry expansion in Release 16, Release 17 and Release 18 represent the second wave of 5G innovation, with timelines spanning from 2021 through 2024 and beyond. Release 17, finalized in 2022, broadens 5G's reach with new device types, expanded spectrum, and enhanced support for non-terrestrial networks. Building on this, Release 18—launched in 2024 as the first phase of 5G-Advanced—further elevates network performance, efficiency, and intelligence, setting the stage for longer-term evolution and new use cases. The following section provides a high-level overview of the key features and priorities introduced in R17 and R18, framing the context for the device-centric focus of this whitepaper.

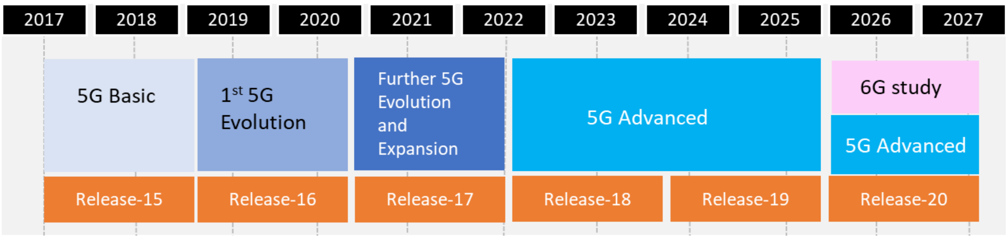

Figure 1.3GPP Timeline of 5G

4.1 Release 17 highlights

4.1.1 RedCap

Cost- and power-optimized UE categories for mid-range bandwidth/latency, enabling wearables, sensors, industrial handhelds, and CPE between NB-IoT/LTE-M and flagship NR, with reduced device complexity and longer battery across wide-area 5G deployments.

4.1.2 MIMO enhancements:

Unified TCI for tighter beam/CSI alignment; multi-TRP transmission for higher reliability and cell-edge performance; Enhancements to reciprocity-based operation include new codebooks with reduced feedback overhead and SRS upgrades.

4.1.3 Non-Terrestrial Networks

Baseline NR/NB–IoT over satellite (Doppler/propagation delay handling, beam/mobility adaptations) for land/sea/air reach and resilience.

4.1.4 Positioning and timing

Enhanced DL/UL references, multi–RTT, integrity metrics for tighter accuracy/availability in logistics, automation, and indoor scenarios.

4.1.5 Power efficiency

eDRX/PEI for deeper idle power savings; PDCCH skipping/SSSGs for sparse–data efficiency and longer battery life.

4.1.6 NR coverage enhancement

Enhancements to uplink control and data channels, utilizing techniques such as repetitions, DMRS bundling, and multi-slots transport block distribution.

4.1.7 Broadcast/Multicast Services:

Enables the broadcast/multicast services in NR. This feature efficiently delivers content such as live TV, OTA software updates and emergency notification.

4.2 Release 18 (5G-Advanced) highlights

4.2.1 Further NR RedCap UE complexity/cost reduction

Additional simplifications and coverage/power optimizations to lower device cost and broaden RedCap applicability.

4.2.2 MIMO Evolution for Downlink and Uplink

Beamforming, CSI/TCI alignment, and UL/DL MIMO including multi–TRP refinements to raise spectral efficiency, robustness, and cell–edge performance.

4.2.3 Multi-carrier enhancements for NR

Refined carrier aggregation/dual connectivity and cross–carrier scheduling to lift throughput, responsiveness, and spectrum utilization.

4.2.4 NR Mobility enhancements

Improved measurements, beam/handover robustness for steadier performance and fewer interruptions.

4.2.5 Study on XR Enh. for NR

Scheduler, QoS, and timing studies to tighten latency/jitter and improve power efficiency for interactive XR and real–time media.

4.2.6 NR side-link relay enhancements

Enable single-hop, sidelink-based, L2 and L3 based U2U relay, enhance service continuity for single-hop Layer-2 U2N relay, and enable multi-path operation using both direct path and PC5 or N3C indirect path, providing path diversity for resilience.

4.2.7 NR Non-Terrestrial Networks enhancements

Enhanced Doppler/timing handling and access/mobility procedures to strengthen continuity over LEO/MEO and satellite–terrestrial convergence networks.

4.2.8 IoT Non-Terrestrial Networks enhancements

NTN procedures and coverage/power profiles tailored for IoT–class devices to enable efficient wide–area sensing, messaging and even voice.

4.2.9 Further NR coverage enhancements

Protocol and RF refinements to extend reach and strengthen cell–edge robustness in FR1/FR2 and challenging radio conditions.

4.2.10 Multicast and Broadcast enhancements

Enables multicast/broadcast reception in RRC_INACTIVE, which cuts RRC resumes and control–plane load, lowers UE power, and improves mobility robustness.

5 High impact device enhancements

Building on the Release 17/18 highlights, the following chapters deep-dive into a curated set of early R17–R18 UE-side features, chosen for near-term relevance and clear, field-visible benefits considering current product and demand trends.

5.1 Further enhancement of UE power

5.1.1 Introduction

While 3GPP introduced several connected-mode power-saving features in Releases 15 and 16 to optimize UE battery consumption when operating on wider channel bandwidths, one of the major contributors to modem-RF power usage remains PDCCH monitoring. This is because the UE must continuously monitor the control channel to receive scheduling information, which can be power-intensive. To address this challenge, 3GPP Release 17 introduced advanced techniques aimed at reducing the time a UE spends monitoring the PDCCH during the active portion of Connected Discontinuous Reception (CDRX) on the active bandwidth part (BWP) of the serving cell. Among these innovations, Search Space Set Group (SSSG) switching stands out as a key enhancement.

The primary goal of SSSG switching is to enable a UE to dynamically switch between different pre-configured SSSGs on the active downlink BWP of the serving cell. This flexibility allows the network to optimize PDCCH monitoring behavior based on traffic conditions and UE requirements. Although the exact implementation may vary depending on the RAN infrastructure, the high-level process typically involves the following steps:

- Configuration: The gNB configures two SSSGs (SSSG0 and SSSG1) per BWP where SSSG switching is supported, using RRC signaling.
- Monitoring and Decision: The gNB monitors traffic patterns and buffer occupancy to determine when a switch is beneficial.
- Switch Trigger: The gNB sends a Downlink Control Information (DCI) message instructing the UE to switch to a different SSSG.
- Timer Management: The gNB may also configure a timer for switching between SSSG0 and SSSG1. This timer can represent either a fixed duration or an inactivity threshold for SSSG1.
- Execution: After the switch delay, the UE stops monitoring the old SSSG and begins monitoring the new one at the start of the next slot.

In DCI-based SSSG switching, the typical expectation is that SSSG0 enables discontinuous PDCCH monitoring to conserve UE power, while SSSG1 configures continuous monitoring at every downlink slot. This design allows the gNB scheduler to dynamically adapt switching the UE to SSSG1 during periods of high throughput or low-latency traffic and reverting to SSSG0 during low activity or latency-tolerant scenarios.

Figure 2 illustrates this concept: during low data activity, the UE monitors PDCCH infrequently under SSSG0, conserving power. When a high data burst occurs, the UE switches to SSSG1 and monitors PDCCH in every slot, ensuring timely scheduling and improved performance.

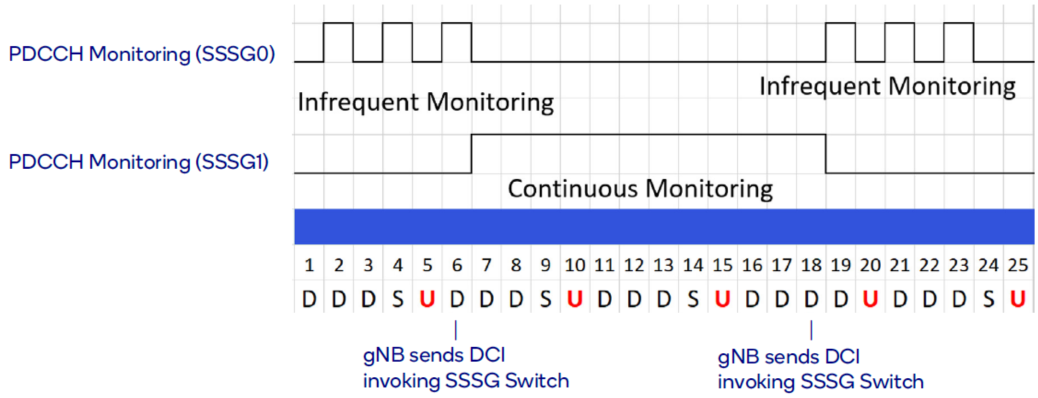

Figure 2.3GPP Rel.17 SSSG Switching

5.1.2 Summary and deployment suggestions

In addition to SSSG switching, 3GPP Release 17 also introduced PDCCH skipping as another power-saving feature. The concept behind PDCCH Skip is straightforward: it allows a UE to reduce the time spent monitoring the PDCCH, thereby lowering overall power consumption. To enable this, the gNB can transmit a DCI message to the UE specifying the duration for which PDCCH monitoring should be skipped. During this skip period, the UE does not monitor the control channel, conserving energy.

However, maintaining PDCCH Skip in the absence of actual data traffic introduces a challenge. A continuous flow of dummy DCIs would be required to keep the skip mechanism active, which can negatively impact PDCCH capacity and scheduling efficiency. This trade-off must be carefully managed by the network.

While both SSSG switching and PDCCH skipping share the same overarching goal—reducing UE power consumption—and rely on similar DCI-based signaling for activation, there are important differences between the two approaches:

Description	**PDCCH Skipping**	**SSSG Switching**
Procedural delay	Immediately after the DCI or after last data scheduling	After the switching delay (typically 2 slots)
Timer trigger	Each skip duration is pre-determined (i.e., based on a timer)	Fallback from one SSSG to another may be based on timer
Early termination of CDRX active time	Trigger PDCCH skipping for rest of the CDRX inactivity timer in the last newTx/reTx/dummy DCI	Switch to an SSSG with large periodicity for CDRX inactivity timer to expire before next PDCCH monitoring occasion (PMO). UE may still have one extra PMO before CDRX inactivity timer expires as PMOs are periodic and fixed in time
Continued skipping /	Requires a constant flow of DCIs	UE remains in current SSSG (assuming

sparse PDCCH monitoring	(dummy-DCIs)	fallback timer is not configured)
Impact to network PDCCH capacity	High	Low

Table 1.PDCCH Skipping vs. SSSG switching comparison

Based on the analysis above, the ecosystem can prioritize accelerated commercialization of SSSG switching features, given their superior performance in power savings, reduced signaling overhead, and minimal impact on service latency. Compared to the baseline Release 16 implementation, Rel.17 SSSG-switching based PDCCH monitoring is expected to deliver significant modem-RF power reductions. For FR1-TDD, estimated power savings vary by application type and can range from 10–20% when using a 10ms periodicity or skip duration.

Given the benefits, it is also evident that such implementation should be expedited for FR1-TDD. For FR1-FDD however, considering that baseline modem-RF power-consumption is not very high, implementation of such feature could be considered optionally.

As for the skip-duration, higher values tend to save more power, but also come at a cost of increased latency. Given this, ideal strategy would be for the gNB to apply different values on an application-specific basis. One example of such implementation could be as under.

- Create multiple categories of applications as a function of their latency requirement viz. low-latency, medium-latency, and latency-tolerant.
- Create individual group of QCI or network-slices. As for example, create slice-1 for low-latency-app, slice-2 for medium-latency apps, while using default network configuration for latency-tolerant apps.
- Create a mapping table of skip-duration pertaining to each slice. As for example, avoid configuring any skip-duration for low-latency slice, but configure 10ms for medium-latency slice, and 40ms for all other application types.
- Such approach would help maximize power-saving benefit, without sacrificing user-experience from latency perspective.
- If such implementation is not feasible at the network-side, 10ms could be considered as the static setting for initial implementation.

Once deployed, it is important to monitor the performance of the network to identify any potential issues.

These features are most effective in low-traffic areas. In high-traffic area therefore, adjustment to the parameters may be needed to make sure user-experience is not impacted.

5.2 UL Tx Switching and 3Tx

5.2.1 Introduction

5G SA is expanding globally, and uplink (UL) demand is accelerating across live sharing and video uploads, enterprise telemetry and machine vision, and latency-sensitive interactive applications.

In current deployments, mid-band TDD carriers provide substantial bandwidth but face tougher propagation and indoor penetration than FDD anchors, and DL-heavy TDD patterns leave fewer UL opportunities. Together with typical device constraints (two transmit chains, SAR, battery, and thermal limits), these factors increase pressure on UL headroom at both user and cell levels.

Baseline technique i.e., R15 ULCA improves coverage and capacity but impose trade-offs on wide TDD UL utilization: there is one transmit chain per component carrier, the UE cannot sustain two UL layers on the mid-band TDD carrier, limiting capacity even as FDD improves coverage.

Given these constraints, optimizing the use of UL transmission opportunities on TDD bands is essential to improve uplink capacity and spectral efficiency. This chapter will examine two complementary techniques defined by 3GPP and evaluated under FR1 TDD+FDD combinations on SA:

Uplink Transmit (UL Tx) Switching (Release 16, with R17/R18 extensions): Tx Switching dynamically assigns the UE’s two transmit chains in the time domain. When UL symbols are available on the TDD carrier, the UE prioritizes 2‑layer UL MIMO on the mid‑band TDD carrier; when the TDD carrier is DL‑occupied or conditions favor FDD, one chain switches to the FDD carrier. This “Switched Uplink” operation yields either 2Tx on TDD or 1Tx on FDD at a given instant. Release 17 introduces dynamic variants (e.g., 2CC 2Tx‑2Tx, 3CC 1Tx‑2Tx, 3CC 2Tx‑2Tx) and UL‑MIMO coherence indications; Release 18 extends flexibility across more bands.

Uplink with Three‑Antenna Transmission (3Tx): This enables concurrent 2x2 UL MIMO on the TDD carrier plus 1Tx UL on the FDD carrier, removing time‑domain compromises. It builds on existing UL CA combined with two UL layers on one NR CC. Today, 3Tx is best suited to CPE/FWA platforms with sufficient power and thermal headroom; higher power classes (e.g., HPUE) further support sustained two‑layer UL on TDD.

5.2.2 Overview

UL Tx Switching, introduced in 3GPP Release 16, is designed for dynamic management of uplink data transmission. This feature aims to enhance uplink performance in various scenarios, including Supplementary Uplink (SUL) and Uplink Carrier Aggregation (UL CA).

The fundamental concept behind UL Tx Switching is to optimize the use of transmission opportunities with 2 transmit antennas in the uplink to maximize efficiency whenever possible.

Release 17 extends the Tx Switching framework to cover dynamic multi-CC operation within two bands and introduces signaling to preserve uplink MIMO characteristics through switching events:

- Dynamic 2CC 2Tx–2Tx switching: This enhancement allows both aggregated carriers to be eligible for two transmit paths, with time-domain reassignment of the full two-path resource between them. The network can instruct the UE to shift the two-path configuration between carriers based on configured policies, without requiring simultaneous dual-carrier transmission of two paths.
- Dynamic 3CC 1Tx–2Tx switching: This mode supports three component carriers across two bands. In a dual‑band 3CC setup (two CCs in one band, one CC in the other), the UE

time–shares its transmit chains—applying 2Tx to one of the two same–band carriers during their UL slots, and maintaining 1Tx on the cross–band carrier when that band is in DL—with no concurrent multi–carrier 2Tx.

- Dynamic 3CC 2Tx–2Tx switching: Same time–domain switching logic as the 3CC 1Tx–2Tx mode, but all three carriers across the two bands are eligible for the two–path boost.

To enable the above modes, Release 17 extends UE capability reporting so that supported dynamic switching variants and valid band/component-carrier combinations can be explicitly advertised. This allows the network to configure switching behavior consistent with the UE's declared support and the defined CA set.

Release 18 scales the Tx Switching scope beyond two bands to enable switching across three or four bands. The multi-band extension generalizes the time-domain reassignment of the two-path resource to a wider set of carriers distributed over additional bands, with corresponding capability advertisement to indicate supported band combinations and switching domains.

Although UL Tx switching significantly improves UE uplink performance, its inherent constraints can lead to suboptimal use of available resources.3Tx adds a third, permanent uplink transmit path so the UE can use UL MIMO on a TDD carrier and UL CA with an FDD carrier at the same time. Practically, this enables simultaneous transmission of TDD 2Tx (two UL layers on the TDD CC) plus FDD 1Tx, removing the time-domain overheads inherent to Tx switching and exploiting both wide TDD bandwidth and FDD robustness in parallel.

From standards point of view, UL 3Tx in 5G NR is achieved by combining UL CA (across FDD and TDD bands) and 2x2 MIMO on the TDD band, using standard UE features. No additional UE capabilities are needed beyond standard UL CA and 2-layer MIMO support. The main requirement is that the network (gNB) must have software capable of scheduling simultaneous uplink transmissions on two component carriers, with 2x2 MIMO on one of them.

Due to form-factor, antenna layout constraints, and the Tx power consumption, current implementations are offered exclusively for FWA platforms and are not available for handheld devices.

For clarity, Figure 3 illustrates the evolution of uplink Tx switching and 3Tx. In Release 15, UEs support two simultaneous uplink transmissions (1Tx + 1Tx) on two bands using ULCA or SUL, with semi-static configuration. In Release 16, uplink transmission introduces time-domain Tx switching: one transmit path dynamically switches between two bands based on traffic, configuration, and channel conditions; only one band is active at a time, but the switching is dynamic. Release 17 enhances this mechanism so that up to 2Tx paths can be switched between two bands, with support for up to three component carriers. Release 18 (5G-Advanced) further generalizes dynamic switching of up to 2Tx across three or four bands, supporting up to five component carriers.3Tx builds on the prior 2Tx capability and, for devices with more relaxed power and form-factor constraints, enables up to three concurrent uplink transmit paths across multiple bands and component carriers with dynamic power sharing to boost throughput, extend coverage, and improve robustness.

In all cases, the figures illustrate how the active transmit resource is reassigned in time to the selected carrier, while adhering to the UE's Tx constraint. Each 5G release increases the flexibility and number of bands/carriers for uplink transmission switching, enabling better adaptation to traffic and channel conditions, and supporting more complex carrier aggregation scenarios.

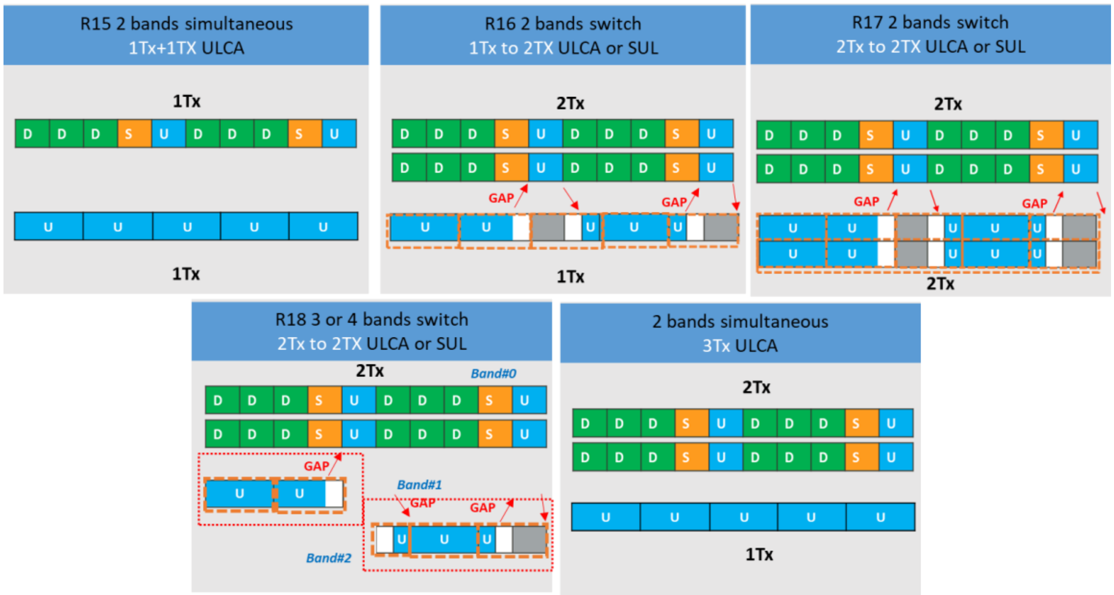

Figure 3. The Evolution of Uplink Tx switching and 3Tx

5.2.3 Performance Analysis

An uplink Tx switching feature evaluation was conducted in a live trial using a commercial UE platform, not a simulation. The uplink CA combination was n78A_n1A, with n78 configured as the PCell and n1 as the SCell. The n78 carrier had 80 MHz bandwidth with 2x2 UL MIMO, while the n1 carrier had 10 MHz bandwidth. The n78 TDD slot configuration used 4D1U as shown in the Table 2. Other key parameters included 14 PUSCH symbols per PRB, 217 resource blocks over the 80 MHz bandwidth, and four UL grant opportunities per 10 ms frame.

Test Configuration	**Pcell**	**Scell**
Band	N1	N78
DL:UL slot configuration	-	4:1
Bandwidth (MHz)	10	80

Table 2.Test Configuration of UL Tx switching

Using previous the configuration for the n1 FDD and N78 peak uplink throughput calculation, the expected throughput is around 60 Mbps for n1 and 200Mbps for N78 UL 2layers.

Below is a demonstration that captures the performance of different techniques evaluated in the trial including n1 UL MIMO, UL Tx switching and 3Tx which was a goal of the exercise.

The FDD-single layer(n1) configuration demonstrates a throughput of 60 Mbps, and TDD Two-layers(n78) configuration demonstrates a throughput of 200 Mbps, indicating their baseline performance. Moving to UL CA (n78_n1), we observe a notable increase in throughput to 160 Mbps. With Release 16, UL Tx switching attains roughly 234 Mbps; relative to the Rel-15 UL CA reference (~160 Mbps), this corresponds to an increase of about 46%. A 3Tx uplink configuration achieves approximately 260 Mbps (+~62% vs the Rel-15 UL CA reference). Release 17 UL CA with Tx switching further increases peak uplink to about 277 Mbps (+~73% vs the Rel-15 UL CA reference). Figure 4Figure 4 illustrates the stepwise increase in peak uplink throughput as carrier aggregation, UL MIMO, and Tx switching are introduced.

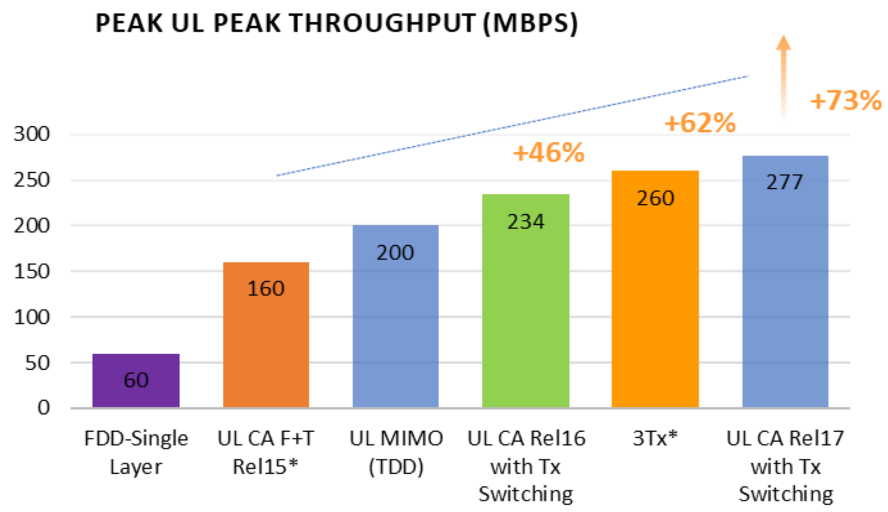

Figure 4. UL Peak Throughput across Technologies

Building on the previous result, the below Figure 5decomposes the previously reported peak uplink throughput into FDD and TDD modes to show how CA, MIMO, and Tx switching affect peak rates in each mode. For Rel-15 UL CA across FDD+TDD, the FDD component is 60 Mbps and the TDD component is 100 Mbps, totaling 160 Mbps. With Rel-16 UL CA incorporating Tx switching, the FDD component is about 34 Mbps (approximately 40% loss than the Rel-15 FDD figure, due to TX path switch), while the TDD component rises to 200 Mbps (roughly double the Rel-15 TDD value), for a total of 234 Mbps. A 3Tx configuration shows 60 Mbps (FDD) and 200 Mbps (TDD), yielding 260 Mbps in total (about 11% higher than the Rel-16 total). Rel-17 UL CA with Tx switching records 77 Mbps (FDD) and 200 Mbps (TDD), summing to 277 Mbps (higher than the 3Tx total). Overall, by introducing Tx switching shifts, the aggregated uplink throughput increases stepwise across configurations, with the principal gains driven by the TDD side.

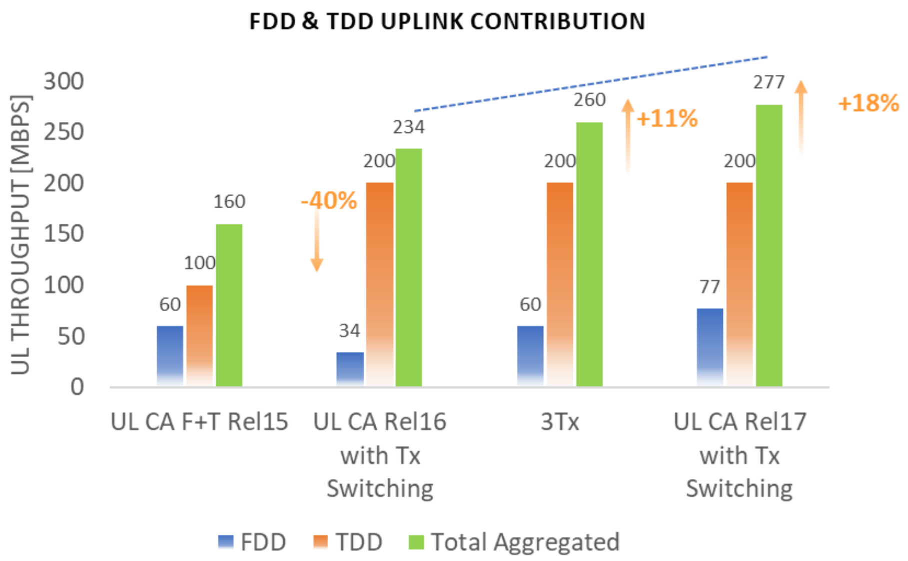

Figure 5. FDD&TDD Uplink Contribution Decomposition

Enabling uplink Tx switching increases overall uplink throughput, however, the FDD component shows a throughput decline due to the UL Tx switching function requires sophisticated scheduling mechanisms at gNB side to ensure proper timing for switching between the carriers, which will introduce some loss because of :

1. Switching delay: During the period of switching, some of the FDD resources will not be available for UL transmission. The actual loss will depend on configuration. For example, in the case of 15 kHz subcarrier spacing (SCS) on FDD carrier, and since UE is using 140us switching delays, this translates to 140/71 = 2 FDD symbols. This means that 2 FDD symbols will not be available for transmission when the UL switching happens from TDD to FDD.

2. TDD SRS in the UL symbols inside the "S" slot. TDD Slot Structure in this trial occupies 2 TDD symbols + 2 Guard Period symbols, with 30KHz SCS, which translates to 2 FDD symbols, this means 2 FDD symbols will not be available for transmission.

3. UL TDD when sent, it creates a gap in FDD (which means FDD can't transmit), and in UL TDD Slot we have 14 symbols which translate to 7 FDD symbols, this means 7 FDD symbols will not be available for transmission.

5.2.4 Summary and deployment suggestions

The introduction of uplink Tx switching and uplink 3Tx transmission marks a meaningful step forward in optimizing 5G NR uplink performance and spectral efficiency. Field evaluations indicate strong gains that should encourage operators to adopt these techniques to advance their networks. Tx switching enables more efficient time-domain allocation of transmit resources, improving system capacity and overall spectral efficiency, while 3Tx delivers higher throughput and added robustness by using UL CA and MIMO concurrently, eliminating switching overheads. Together, these capabilities translate into higher uplink rates, better spectral utilization, and an enhanced user experience, and they are commercially ready on both the UE and network sides.

Looking ahead, uplink Tx switching continues to evolve. While current implementations are limited to switching between two component carriers across two bands, Release 17 expands the feature with dynamic switching options—such as 2CC 2Tx–2Tx, 3CC 1Tx–2Tx, and 3CC 2Tx–2Tx within two bands—alongside signaling for UL-MIMO coherence and corresponding UE capability indications. Release 18 further extends switching across three or even four bands. As the industry prepares for Release 19, ongoing work targets additional uplink throughput gains. Dynamic switching between FDD and TDD across asymmetric 1Tx/2Tx and 2Tx/1Tx configurations, when combined with 3Tx, is projected to deliver more improvement in peak uplink throughput under favorable conditions, materially improving spectral efficiency and overall network performance, subject to appropriate UE and gNB support.

5.3 Dynamic Waveform Switching

5.3.1 Overview

Coverage remains one of the most critical factors for operators when deploying and commercializing cellular networks because it directly influences service quality, user experience, and overall network economics, including CAPEX and OPEX.

In real-world deployments, uplink (UL) performance often becomes the bottleneck, particularly as emerging vertical applications introduce UL-heavy traffic patterns, such as real-time video uploading, industrial IoT, and remote monitoring. Recognizing these challenges, 3GPP introduced significant coverage enhancements in Release 17 under the work item "NR Coverage Enhancements," extending coverage for key uplink channels such as PUSCH, PUCCH, and Msg3. These efforts continued in Release 18, further refining NR coverage capabilities to meet the demands of advanced use cases and stringent performance requirements.

One of the most notable innovations in Release 18 is Dynamic Waveform Switching (DWS) between DFT-s-OFDM and CP-OFDM for uplink transmissions.

Traditionally, the uplink waveform was statically configured via RRC signaling during initial setup, meaning that once a UE was assigned CP-OFDM or DFT-s-OFDM, it could not adapt to changing radio conditions without undergoing a full RRC reconfiguration—a process that introduces latency and signaling overhead. CP-OFDM, the default NR waveform, offers high spectral efficiency and supports advanced MIMO schemes, making it ideal for cell-center users with favorable channel conditions. However, its high Peak-to-Average Power Ratio (PAPR) forces UEs to back off their power amplifiers, reducing coverage and energy efficiency. Conversely, DFT-s-OFDM, also known as SC-FDMA in LTE, provides lower PAPR, improving power efficiency and coverage for cell-edge users, though it lacks flexibility for multi-layer MIMO.

DWS addresses these limitations by enabling the gNB to dynamically select the optimal waveform on a per-grant basis without requiring RRC reconfiguration. This is achieved through a lightweight one-bit indicator in DCI formats 0_1 or 0_2, signaling whether the UE should use CP-OFDM or DFT-s-OFDM for the upcoming PUSCH transmission. For semi-persistent scenarios, MAC Control Elements (MAC-CE) can maintain waveform settings across multiple transmissions, while RRC remains responsible for enabling or disabling DWS capability per Bandwidth Part (BWP).

To ensure decoding consistency when switching waveforms, 3GPP specifies size-alignment rules for DCI fields such as TPMI and SRI, where the maximum bit width is used, and the UE interprets only the relevant bits for the indicated waveform.

5.3.2 Benefits

The benefits of DWS are substantial. From a coverage perspective, dynamically enabling DFT-s-OFDM can improve uplink link budget by approximately 2 dB for cell-edge UEs, enhancing reliability and reducing retransmissions.

From a power efficiency standpoint, lower PAPR reduces power amplifier backoff, extending battery life for UEs operating under challenging conditions.

Throughput optimization is another key advantage: CP-OFDM can be prioritized for high-rank MIMO and favorable channel conditions, while DFT-s-OFDM can be applied when coverage or power constraints dominate. This flexibility allows operators to balance performance and efficiency dynamically without sacrificing spectral efficiency or introducing heavy signaling overhead.

For deployment, the feature DWS is very beneficial for UEs in coverage-limited areas. Overall, dynamic waveform switching represents a significant step forward in making NR uplink more adaptive, efficient, and capable of meeting diverse service requirements, combining the strengths of CP-OFDM and DFT-s-OFDM in a flexible and intelligent manner.

5.4 1024QAM

5.4.1 Overview

1024QAM (Quadrature Amplitude Modulation) represents one of the most significant advancements in high-order modulation for mobile networks. It was introduced in 3GPP Release 15 for LTE downlink and later adopted in Release 17 for 5G NR FR1 downlink (PDSCH) as part of 5G-Advanced enhancements.

The standardization process involved defining the constellation mapping, updating MCS tables, and aligning CQI feedback mechanisms to support this higher modulation order. These specifications are captured in 3GPP TS 38.211 (modulation functions), TS 38.214 (MCS tables), and TS 38.521 (test requirements).

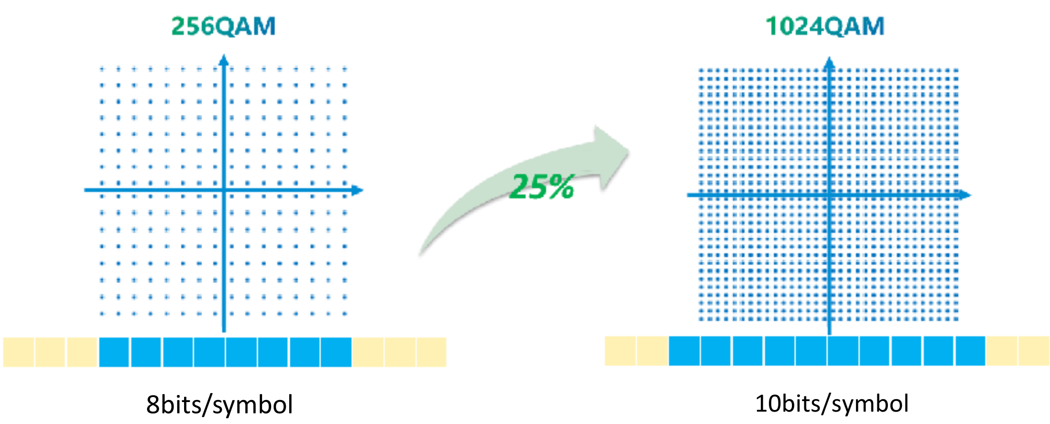

Figure 6. Constellation Mapping :256QAM vs. 1024QAM

5.4.2 Benefits

The primary motivation for introducing 1024QAM is to maximize spectral efficiency and boost data throughput within existing bandwidth resources. This improvement is particularly valuable in scenarios with high SINR, good channel conditions (e.g., LOS or low mobility), and advanced RF capabilities.

- Higher spectral efficiency – Raising modulation order from 256QAM (8 bits/symbol) to 1024QAM (10 bits/symbol) yields a theoretical +25% peak throughput gain at the same bandwidth.
- Capacity & user experience – In high-SINR cells, networks can serve more users or higher bit-rates per user without additional spectrum, enabling smoother 4K/8K streaming, XR/VR, and cloud gaming experiences.
- Ecosystem momentum – Trials across operators and vendors show consistent gains and pave the way for 5G-Advanced commercialization.

In the IMT2020PG 2025 trials, 1024QAM showed extreme speed performance, combined with downlink 5CC CA. For the test, five carriers are aggregated - 160MHz on 2.6GHz (n41), 160MHz on 4.9GHz (n79) and 2x30MHz on 700MHz (n28) and utilized 4-layer downlink transmission and 1024QAM high-order modulation for all component carriers, the test ultimately made new breakthroughs with downlink peak rate of 8Gbps. The validation of 1024QAM completed by chipset vendors in collaboration with industry partners marks a new milestone in wireless communications capabilities and lays a solid foundation for accelerated implementation of cutting-edge use cases. With faster downlink speeds, 5G-A will provide robust support for innovative applications such as immersive communication, real-time cloud rendering, and UHD livestreaming.

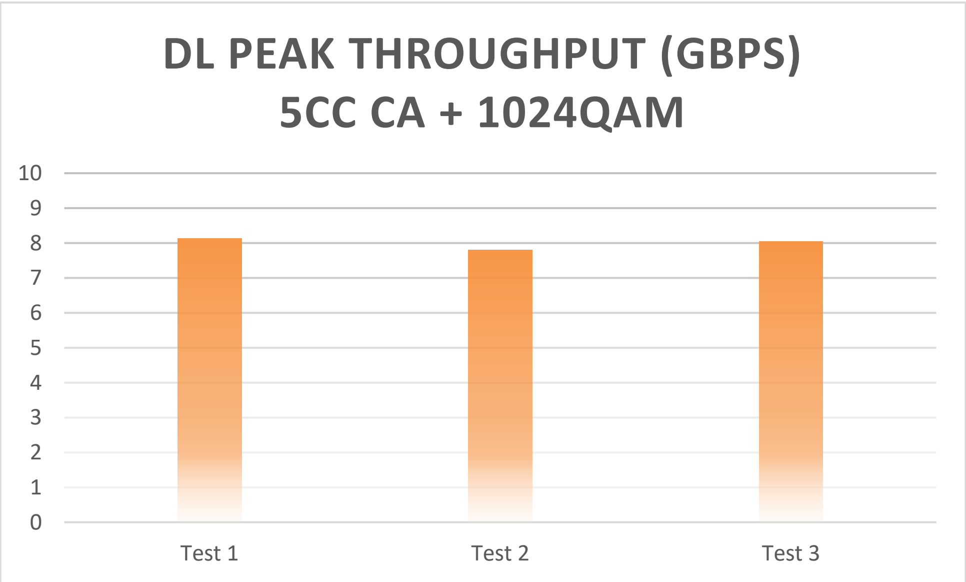

Figure 7.1024QAM Performance in IMT2020PG trials

5.5 L1/L2 Triggered Mobility

5.5.1 Introduction

As 5G advances toward higher frequencies and wider bandwidths, networks are becoming denser with smaller cell footprints. Under dense network deployments, and UE in high mobility, the frequent border crossings are the norm: UEs traverse multiple cells within short windows while beams are continually reselected. In this environment, the legacy, L3-driven handover model—based on measurement reports and RRC signaling—reveals structural limits: long control loops and difficulty compressing service interruption time. Concretely, classic handovers must wait for L3 measurements to trigger and report, the network to decide and send an RRC handover command, and the UE to complete random access and synchronize at the target cell. At high speeds, weak coverage edges, or in high bands, this inflates end-to-end handover latency and interruption time, leading to throughput loss, jitter spikes, higher (RLF) rates, and more ping-pong events—degrading time-sensitive applications such as high-speed moving, XR, URLLC, V2X and industrial control.

To address these bottlenecks, 3GPP introduced LTM (L1/L2 Triggered Mobility) in Rel-18 as part of the “Further NR mobility enhancements” effort. LTM lowers the “break-before-make” trigger path into faster PHY/MAC-level mechanisms.

5.5.2 Overview

In legacy NR mobility, handover is an L3/RRC-driven procedure designed to preserve the radio link while masking mobility-induced disruptions to the ongoing service. The UE conducts measurements of the serving and neighbor cells as configured by the network, typically using downlink reference signals such as SSB and CSI-RS. These observables are filtered over time at L3, and event logic like A3 or A5 compares neighbor and serving signal quality with configurable offsets and time-to-trigger. When a reporting condition is satisfied, the UE sends a measurement report. The source gNB evaluates policy and radio conditions, prepares the target context, and issues an RRC handover command that carries the target cell parameters and access information. The UE then detaches from the source in the sense of traffic discontinuity, synchronizes to the target, and executes random access—generally a contention-free RA where possible—before establishing or resuming SRB/DRB bearers and continuing PDCP state. In parallel, the source and target coordinate context transfer and data forwarding over the Xn interface in intra-RAN cases or NG signaling for core path switch when applicable, as illustrated in left part of Figure 8. In Rel-16, CHO has been raised to improves HO robustness.as depicted in right part of Figure 8. However, the interruption persists: the UE still needs to access the target and complete RRC reconfiguration on execution, so service break duration is essentially unchanged.

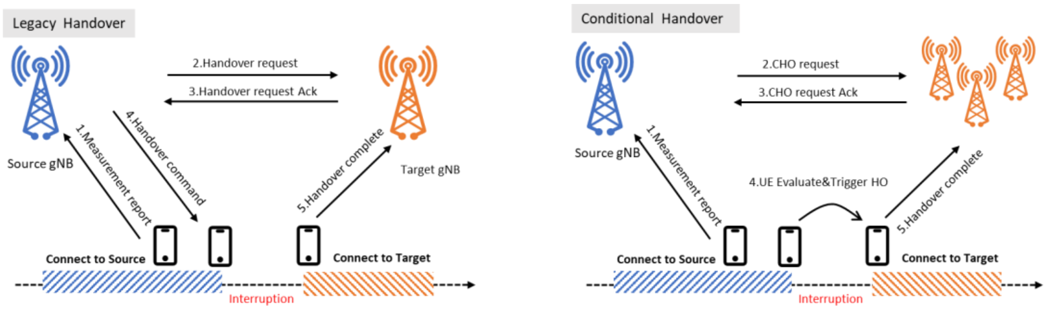

Figure 8.Legacy HO and CHO Procedures

If you trace the legacy handover flow end to end, as illustrated by Figure 9, it becomes clear that the current break-before-make framework inherently creates an interruption window that resists compression. This interruption normally does not impact commonly used services like telephony, video streaming or web surfing. However, for latency-sensitive and interactive services, such as video calling, real-time gaming, XR or remote control and mobility automation, this interruption causes stutter, audio dropouts, increased delay, and can push time-critical tasks past their deadlines. Also, in mid- and high-frequency deployments or at high-speed moving, handovers occur frequently, and these per-event interruptions accumulate into a meaningful share of session time, inflating tail latency and failure risk.

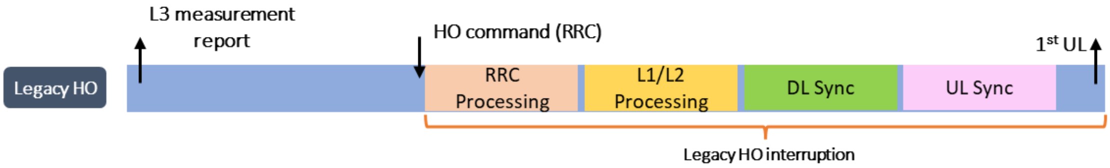

Figure 9.Legacy HO timeline

L1/L2-Triggered Mobility (LTM) introduced in Release 18 re-architects the break-before-make path to address these limits.

The mobility procedure in LTM is performed in four steps.

Step 1 LTM preparation: The serving cell preconfigures a set of candidates cells and delivers parameters either as full per-candidate content or via a compact "reference configuration plus per-candidate delta" model. This keeps air-interface overhead contained while ensuring the UE holds all information needed to act promptly.

Step 2 Early sync: Based on that preparation, the UE performs early configuration decoding and early downlink and uplink synchronization to candidates. Early UL sync can use RACH-based timing-advance acquisition or, when conditions permit, RACH-less timing alignment; If early UL sync uses RACH, the UE may briefly pause activity on the serving cell to send PRACH and retune RF. The standards cap this pause, and its length depends on where the RACH opportunity falls and whether the UE can operate radios concurrently.

Step 3 LTM cell switch execution: When a switch is needed, a MAC Control Element from the serving cell. The command includes the target configuration ID, which identifies the target cell's candidate configuration. The UE switches to the target cell and applies the candidate configuration referenced by that ID.

Step 4 LTM cell switch completion: After the trigger, if the UE does not have a valid timing advance (TA) for the target cell, it performs a random access (RA) procedure to the target cell. Then the UE sends an RRCReconfigurationComplete message to the target cell, completing the LTM cell switch. If the UE performed RA previously, it considers the LTM switch successfully completed once the RA procedure succeeds. Otherwise, the UE considers the switch successfully completed when it determines that the network has received its first uplink (UL) data. Given that SRB1 has a higher logical channel priority, during RACH-less LTM the UE prioritizes assembling the MAC PDU carrying RRCReconfigurationComplete and transmits it to the target cell as the first UL packet after the switch.

The Figure 10 illustrates each step of LTM in a message sequence chart with the signaling between the UE and the gNB, which is also captured in 3GPP TS38.300

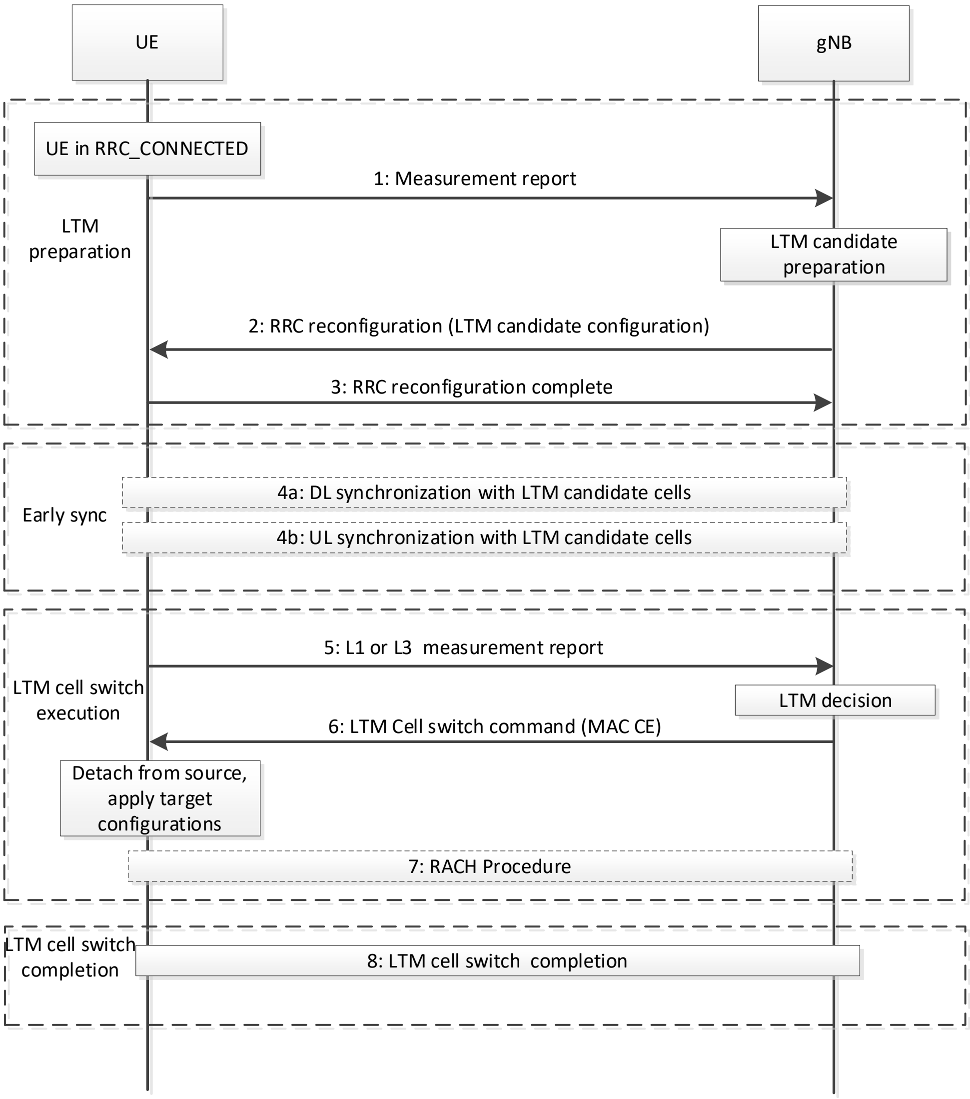

Figure 10. LTM HO procedure

Comparing to legacy HO，LTM reduces the HO interruption by moving expensive parts of search, coarse DL sync, and UL timing acquisition earlier—while the UE is still attached to the source—and by using a low-layer (L1/L2) trigger and lightweight control to keep the execution-time work to a minimum， which is illustrated in Figure 11.

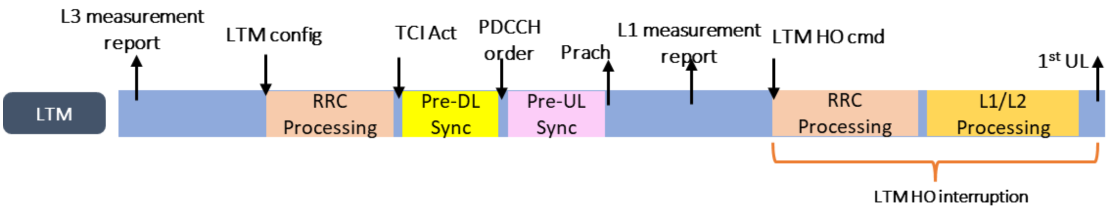

Figure 11. LTM HO timeline

A key scalability property is that once the candidate set and pre-configuration are in place, subsequent LTM operations can proceed consecutively within that set without fresh RRC provisioning each time. This “subsequent LTM” behavior is critical in dense environments where short stays are common, and handovers are frequent. LTM also defines robust fallback and continuity handling. For intra-CU LTM, security context updates are not required solely due to the LTM switch; however, fast-recovery paths must preserve SRB PDCP count continuity to avoid keystream reuse, and the network side may need to handle PDCP sequence gaps with appropriate reordering timers to prevent added latency. The procedure clarifies the interplay with CHO to prevent cascading “fast recovery” attempts; for example, if an LTM fast recovery path fails under certain conditions, the UE proceeds to RRC reestablishment rather than attempting CHO quick recovery immediately, and stored LTM configuration is released if reestablishment proceeds outside the prepared candidate set.

5.5.3 Performance Analysis

To complement the theoretical budget, LTM was evaluated on a joint testbed using a Rel−18 LTM−capable gNB and an LTM−enabled UE prototype. The test campaign covered handovers on both TDD and FDD bands under TCP and UDP traffic, using legacy handover as the comparison baseline.

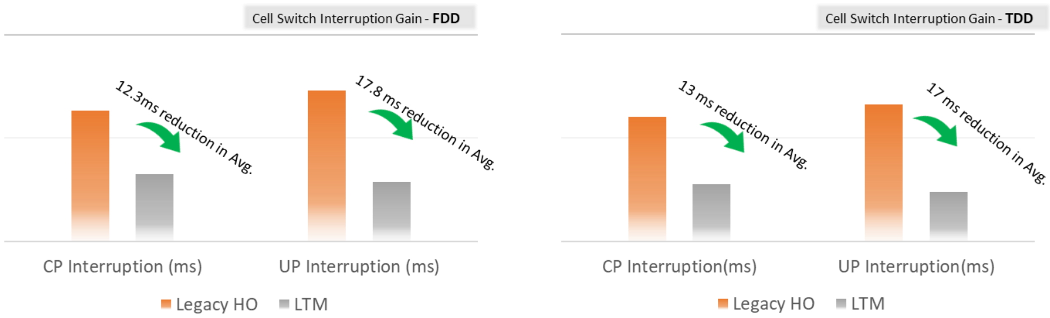

Figure 12.LTM HO interruption vs. Legacy HO interruption

Figure 12 shows that LTM significantly reduces cell switch interruption times compared to Legacy HO in both FDD and TDD modes. The control plane interruption is from LTM cell switch triggered to RRCReconfigurationComplete receiving time, which is calculated by gNB side; The user plane interruption time is from the last PDSCH to the first PDSCH, which is calculated by UE side.

During the comparative tests, we kept the network scheduling strategy consistent and conducted multiple rounds of testing to improve the reliability of the data. For both control plane and user plane interruptions, the average reduction is over 12 ms, with the greatest improvement seen in user plane switching and these reductions are consistent with our previous theoretical analysis. This demonstrates LTM's clear advantage in minimizing service disruption during cell handovers.

5.5.4 Summary

LTM shortens the break‑before‑make gap by shifting target discovery, coarse downlink sync, and uplink timing acquisition into preparation and using L1/L2 triggers instead of an execution‑time L3 round trip. This converts full search and full RA into brief confirmations, reducing median and tail interruption and improving reliability (HOF/RLF and ping‑pong).

Experimental validation on a joint testbed corroborates these trends. Compared with Legacy HO under the same topology and load, LTM demonstrates a clear advantage in minimizing interruption.

From a deployment perspective, the most consistent benefits appear in intra‑CU intra‑/inter‑DU topologies and in same/adjacent‑band scenarios where candidate pre-configuration is accurate and PRACH opportunities are aligned for early TA acquisition, with CFRA enabled.

Looking ahead, the Rel‑19 introduction of inter‑CU LTM is expected to broaden applicability beyond the current intra‑CU scope, simplifying adoption across wider mobility boundaries and making the feature easier to operationalize at scale. Inter‑CU support will extend LTM to more gNB domains and enable consistent latency and reliability gains across heterogeneous deployments. As UE capabilities and gNB software mature, LTM is poised to become a mainstream mobility optimization, with straightforward integration into existing frameworks.

5.6 L4S

5.6.1 Introduction

In today's hyper-connected environment, low and consistent latency is a decisive factor for user experience and service quality. From video conferencing and cloud gaming to vehicle connectivity, enterprise collaboration, and industrial control, applications increasingly depend on real-time interactions that must remain responsive under varying load and radio conditions. When milliseconds matter, operators need mechanisms that minimize queuing delay and jitter without sacrificing throughput, so they can sustain a competitive edge and deliver reliable, efficient digital services at scale.

To understand where latency originates and how to mitigate it, it is useful to decompose total latency—the time for a packet to travel from source to destination—into four primary components:

- Propagation delay: The time imposed by distance and the speed of light. It can be reduced by placing compute and content closer to users (for example, edge sites within roughly 100 km can keep round-trip propagation below about 1 ms).
- Application processing delay: The latency introduced by the software applications themselves when data is processed before sending or after receiving. It depends on code efficiency, hardware processing power, runtime environment, and the complexity of the tasks (e.g., encoding/decoding, encryption, rendering, analytics).
- Interface delay: The finite time required by the access technology to handle packets, including aggregation, protocol/stack processing, and media acquisition/scheduling on a shared medium. Modern systems such as 5G NR drives these into the millisecond range.
- Queuing delay: Arising when packets wait in buffers at routers, switches, or the RAN scheduler.

Among on the latency breakdown, queuing delay is typically the largest and most variable contributor to end-to-end latency—and the chief target for improvement. Even when propagation and interface/processing delays are driven into the (sub–) millisecond range, interactive experiences often degrade when queues build up at true bottlenecks, especially in wireless access.

Why queuing delay dominates? The main factors include:

- Burstiness and demand peaks: Multiple elastic flows (e.g., TCP/QUIC transfers, variable–bit–rate video) can momentarily exceed a link's capacity, forcing packets to wait in buffers and inflating delay and jitter.
- Time–varying wireless capacity: In cellular and Wi–Fi, instantaneous throughput fluctuates with radio conditions and scheduling. When capacity dips, backlogs grow quickly and latency spikes.
- Bufferbloat trade–off: large buffers improve utilization but at the cost of high and variable delay; small buffers cut delay but risk under–utilization.
- Heterogeneous traffic and RTTs: Mixed round–trip times and flow counts make queue dynamics harder to predict and control, further amplifying jitter for real–time applications.

Designed to minimize queuing delay, the IETF's Low Latency, Low Loss, and Scalable Throughput (L4S) architecture replaces drop–based feedback with explicit, fine–grained ECN marking while ensuring clean coexistence with non–L4S traffic.

5.6.2 Overview

L4S emerged from long–standing efforts to mitigate queuing delay without sacrificing utilization.

Traditional AQM improves on drop–tail by acting before buffers fill, but it still communicates congestion mostly implicitly via packet drops or coarse ECN, which adds jitter and retransmission delay—particularly harmful to real–time applications like XR/AR/VR and cloud gaming. Moreover, classic loss–based congestion controls (e.g., Reno, CUBIC) tend to depend on a non–trivial standing queue for stable feedback, making near–zero queuing delay difficult to achieve alongside high link utilization, especially in heterogeneous flow and RTT mixes.

L4S directly targets these drawbacks by retaining AQM's early congestion detection but converting it into frequent, explicit ECN marks at very shallow queue levels. Bottlenecks mark L4S–capable packets (ECT (1)) with CE as soon as a small backlog appears, and Prague–compliant endpoints adjust sending rates quickly and smoothly without relying on loss or standing queues. To ensure clean coexistence with legacy traffic, L4S introduces Dual–Queue, coupled AQM: L4S and Classic packets are separated into distinct queues, the L4S queue uses ECN marking to maintain near–zero delay, and the Classic queue preserves traditional behavior. A coupling mechanism links their dynamics so that growth in the Classic queue increases the L4S marking probability, delivering long–term fairness without starving either class. The result is consistently low queuing delay with low jitter and scalable throughput across diverse flow counts and RTTs, achieved incrementally wherever endpoints and true bottlenecks support L4S.

The L4S suite was standardized in 2023 as three core RFCs: RFC 9330 defines the end–to–end architecture, RFC 9331 specifies the "Prague" requirements for responsive, scalable endpoint congestion control, and RFC 9332 describes the Dual–Queue, coupled AQM at bottlenecks (commonly implemented as DualPI2). Operationally, L4S relies on ECN signaling with ECT (1)/CE in the IP header, accurate ECN feedback in transports such as TCP and QUIC (and RTP/RTCP ECN for media), and correct ECN preservation across tunnels so marks reach endpoints even in encapsulated paths.

5.6.3 Performance Analysis

An end-to-end implementation is preferred for L4S, with all nodes in the way being able to detect congestion and apply the packets marking with ECN. However, the most important nodes are those placed in possible bottlenecks. For a 3GPP wireless network, one such bottleneck is the air interface. Two network elements placed at the edges of this interface are the UE (User Equipment) and the base station: for 5G this is a gNB.

The UE controls uplink congestion, while gNB can do the same for the downlink. Due to reasons such as the subscribers' traffic profile and more transmit power at the base station, downlink performance has always prevailed in wireless networks, and 5G is no exception. This leaves UE to gNB interface as the most probable bottleneck in a 5G network, where buffer congestion may appear, impacting delays on latency-sensitive applications. This is why L4S on the device side is pivotal: it shortens the feedback loop for uplink congestion control, lets the sender react immediately to CE marks, and helps keep the radio-side queue shallow.

Extensive testing on live networks has shown that L4S–enabled CPEs reduce latency significantly compared to AQM and unoptimized baselines, a simple diagram is shown in Figure 13. These gains were observed in these live video–calling trials.

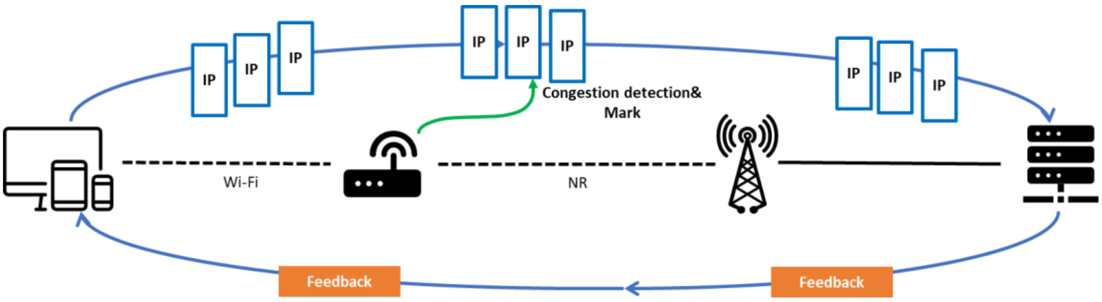

Figure 13. UE side L4S diagram

The tests covered three network conditions: no congestion, a 10 Mbps margin to congestion, and a fully congested network. The result was L4S delivered consistent and stable performance regardless of bandwidth, with significantly lower latency variance than AQM and unoptimized traffic. Service data rates and video call quality remained largely unaffected. Lab results showed even greater performance gains, indicating further optimization potential.

Across these scenarios, enabling L4S produced substantial reductions in queuing delay while maintaining high link utilization. Under real–time, interactive workloads, we observed decreases of up to roughly 65 % compared with classic AQM, and up to about 5-fold versus unoptimized queues. In addition, tail latencies and jitter were noticeably lower, and rate adaptation stabilized more quickly after capacity changes.

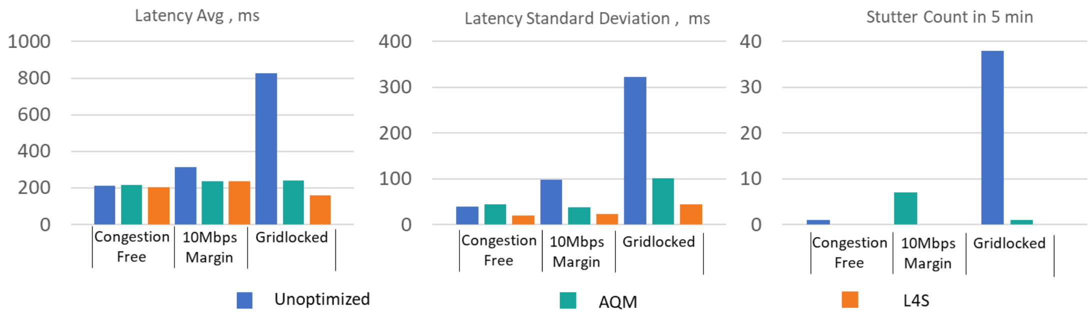

Figure 14. Performance comparison: L4S vs AQM vs Unoptimized

The practical impact is most evident in real–time and interactive workloads. For video conferencing, shorter and more stable queuing delays translate to clearer audio, smoother motion, and minimal conversational lag. Streaming applications benefit from uninterrupted playback and consistent picture quality even during busy periods, and in live trials we saw stalling effectively eliminated across the tested conditions when L4S was enabled end to end. Interactive entertainment shows similar benefits: online games respond more quickly and exhibit reduced jitter, yielding a more fluid experience and, in competitive settings, an observable advantage. Emerging use cases such as XR and cloud gaming also profit from the faster feedback loop and shallow queues, which keep interactions responsive even as radio capacity fluctuates.

5.6.4 Summary

L4S replaces loss-driven signaling with explicit ECN marks, Prague-compliant endpoint control, and a Dual-Queue coupled AQM at bottlenecks, aiming to keep queues near zero while preserving high utilization and ensuring long-term fairness with classic traffic.

Device-side uplink L4S is particularly important. The UE has the most immediate view of its transmit buffers and queue depth. When the UE reacts quickly and smoothly to CE marks with Prague behavior, the control loop shortens, the radio-side queue stays very shallow, and uplink-dominant interactive paths (for example, voice upstream and screen sharing) see meaningful reductions in delay and jitter.

In evaluations spanning uncongested, near-saturation, and fully saturated conditions, end-to-end L4S delivered consistently tighter latency distributions, substantially lower tail delay and jitter, and stable throughput and service bitrates.

5.7 UE Aggregation

5.7.1 Introduction

As uplink-intensive applications proliferate, the limitations of individual UE have become increasingly apparent in real-world deployments. For example, the rise of live broadcast applications, from outdoor events to mobile content creation, places stringent requirements on uplink rate, coverage, and reliability, as users move into parks, concerts, and remote locations.

3GPP has introduced the multipath relay feature—more intuitively referred to as UE aggregation, which will be the term used throughout this whitepaper. First standardized in Release 18, this feature can be leveraged to help overcome uplink bottlenecks in demanding scenarios. Rather than replacing single-UE operation, the UE aggregation enables multiple terminals to cooperate, forming an aggregation group in which a Remote UE can utilize one or more nearby relay UE(s) to enhance its uplink path. This approach is particularly valuable in situations where a single device's transmit power, channel count, or bandwidth becomes a limiting factor. By pooling the capabilities of multiple commercial-grade devices, UE relay-based aggregation offers a cost-effective and flexible alternative to high-power, custom hardware, while also supporting differentiated service levels and terminal types.

5.7.2 Overview

UE aggregation builds a cooperative uplink using two legs from a single Remote UE to the same gNB: a direct Uu path maintained by the Remote UE itself, and an indirect path contributed by a nearby Relay UE that forwards the Remote's user plane over its own Uu. In Rel‑18, the control plane remains anchored on the direct path, and both direct and indirect path operate under the same gNB. This ensures stable control procedures while allowing the user plane to be parallelized or duplicated when conditions and policy justify it. The two UEs are connected by a inter‑UE link, which can be realized with a non‑3GPP (N3C) link or with 3GPP sidelink (PC5). Under N3C, only the Layer‑2 multi‑path relay architecture is supported; under PC5, both Layer‑2 and Layer‑3 relay architectures are supported, with different implications for the RAN.

At Layer‑2, the split and merge occur at PDCP and are coordinated end‑to‑end between the Remote UE and the gNB. The Remote UE hosts a multi‑path split bearer and decides per PDCP PDU whether to send on the direct leg, the indirect leg, or both. Split mode apportions distinct PDCP PDUs across the two legs to lift effective uplink throughput and bandwidth headroom; duplication sends the same PDCP PDU concurrently on both legs to harden reliability and compress tail latency, with the gNB discarding the later arrival. The gNB terminates both legs for the same DRB at PDCP, applies sequence–number-based reordering and de–duplication, and delivers in–sequence SDUs upward. Because the RAN is path–aware at PDCP, it can tune reordering windows and timers to the observed delay skew between legs and enforce QoS consistently across the coordinated bearers. User–plane ciphering and integrity remain end–to–end between the Remote UE and gNB; the Relay forwards opaque PDCP PDUs and cannot access payload contents. Figure 15demonstrated UE aggregation at layer-2 via N3C interface.

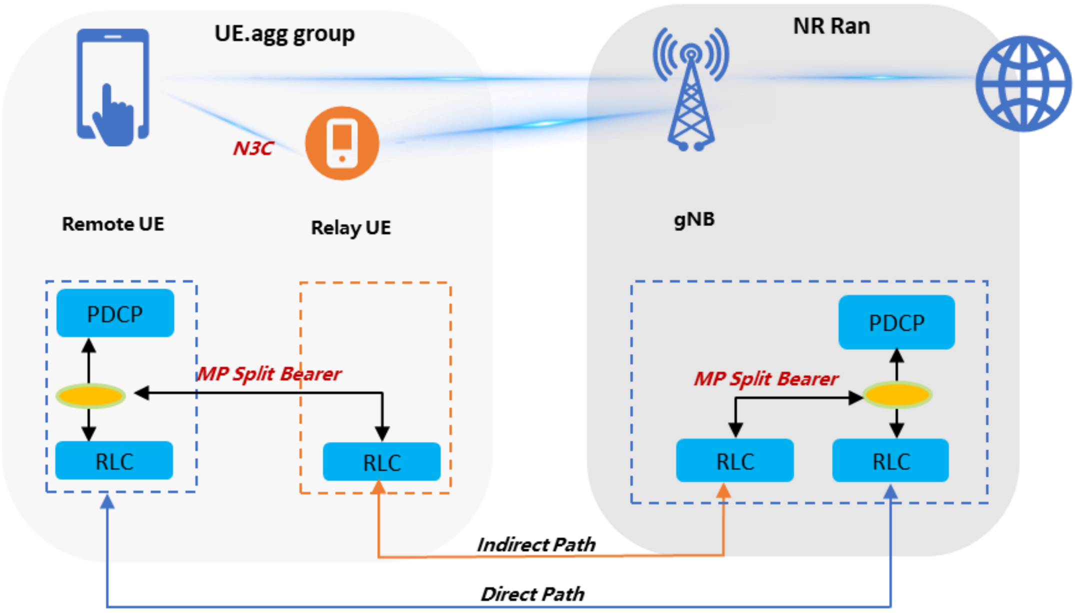

Figure 15.UE Aggregation at Layer-2 via N3C interface

N3C is the preferred realization of the inter–UE connection for early deployments. The Remote UE uses a widely available short–range non–3GPP link—typically Wi–Fi Direct or Bluetooth—to carry its uplink PDCP PDUs to the Relay UE. Path composition is additive: the indirect leg is brought up on top of an existing direct leg under the same gNB and can be suspended or torn down rapidly if N3C quality or Relay headroom degrades. Note that, operational constraints in Rel–18 bind one Relay to one Remote at a time on N3C, with the Remote allowed to keep a candidate set of Relays. Minimal framing is introduced only as needed to preserve bearer mapping and segmentation context. The Relay injects the Remote's PDCP PDUs into its Uu stack via a one–to–one mapping from the Remote's DRBs to dedicated RLC entities and separates forwarding traffic from local services with distinct logical channels for gNB scheduling and accounting.

The Figure 16illustrates the UE aggregation establishment flow in detail.

In step1 and step2, both the Relay UE and the Remote UE access the cell and inform the gNB of their respective roles.

Next, if the Remote UE is already associated with a Relay UE via the N3C link and is ready to initiate UE aggregation, it should report at least the list of C-RNTIs and cell IDs of candidate Relay UEs—information exchanged over the N3C link—using the UEAssistanceInformation message to the gNB, as shown in step 3.

Then,the gNB then sends RRC reconfiguration messages to both the Relay UE and

Remote UE, specifying the configuration for split data radio bearers and the parameters for data splitting between the direct and indirect paths, as described in step 4 and 5.

After the reconfiguration is complete, data transmission begins, with the Remote UE's data being forwarded simultaneously through both the direct path to the gNB and the indirect path via the Relay UE, enabling efficient multipath communication.

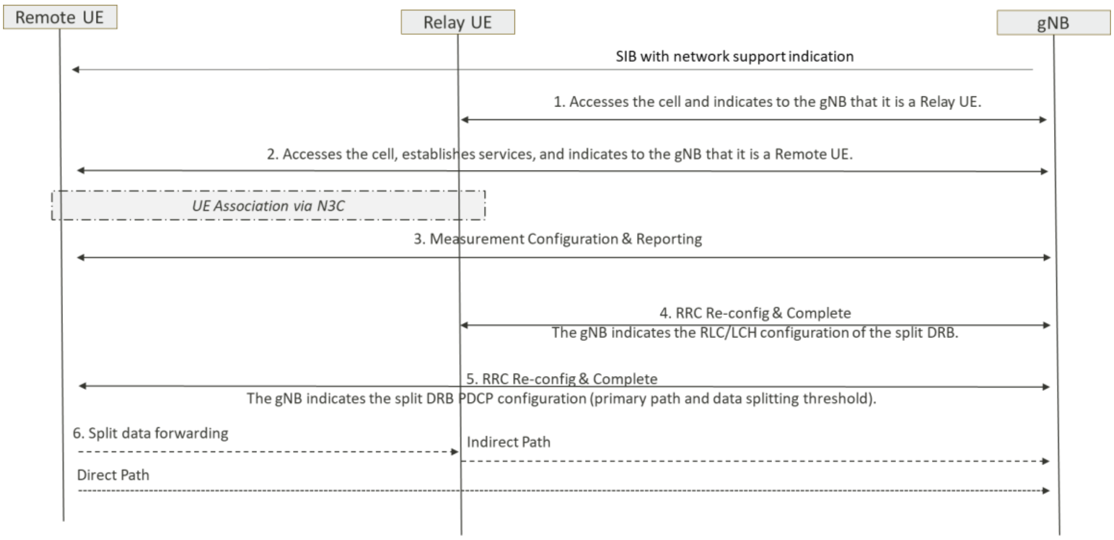

Figure 16.UE Aggregation Establishment Flow

Mobility under Rel–18 respects the same–gNB constraint; handovers that change the serving gNB or CU typically require tearing down the indirect leg and re–establishing it after the handover completes. Failures or degradations on N3C or the Relay's Uu are reported over the surviving leg where possible, and fallback follows baseline single–UE procedures to preserve stability.

Layer–3 multipath schemes, which split above PDCP (for example at IP or application), do not offer gNB–side PDCP merge/de–duplication and leave ordering and reliability to higher layers, reducing RAN determinism for the uplink augmentation problem.

PC5 sidelink is acknowledged as an alternative inter–UE connection, but its commercial penetration remains far behind N3C, making N3C markedly easier to deploy at scale today.

5.7.3 Performance Analysis

A lab test has been conducted to assess the performance of UE aggregation. Following results come from a cabled setup as demonstrated in Figure 17: the UEs and the gNB are connected with cable programmable attenuators to emulate a range of "good/poor coverage" radio conditions. Note that the setup does not model the physical distance, contention, or interference of the N3C link (for example, Wi–Fi channel load, obstacles, or Bluetooth coexistence). Consequently, the measured gains reflect L2 multipath behavior under controlled radio conditions, while real–world outcomes may be higher or lower depending on N3C link quality (channel load, coexistence, obstacles, etc.) or capability.

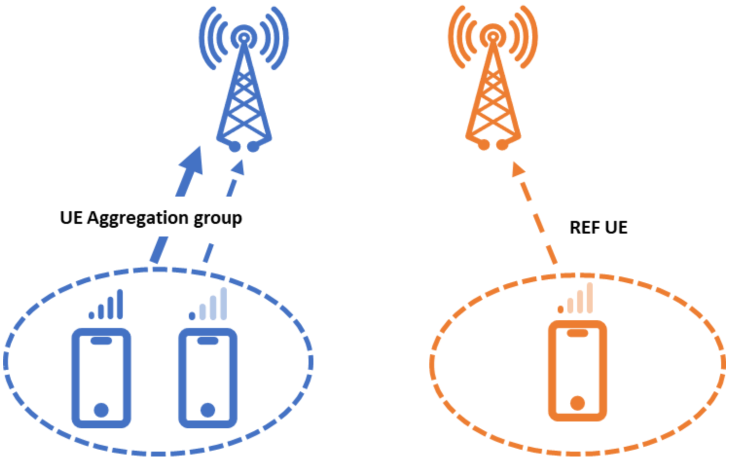

Figure 17. UE Aggregation lab trial setup

Two primary metrics are reported: uplink throughput and uplink latency, reflecting the different objectives of split and duplication mode.

In split mode, PDCP PDUs are distributed across the direct and relay legs, leveraging parallel capacity and path diversity; under sustained load this aggregates bandwidth and reduces queuing on any single path, so throughput improves (roughly 2–10×). The results are depicted in Figure 18.

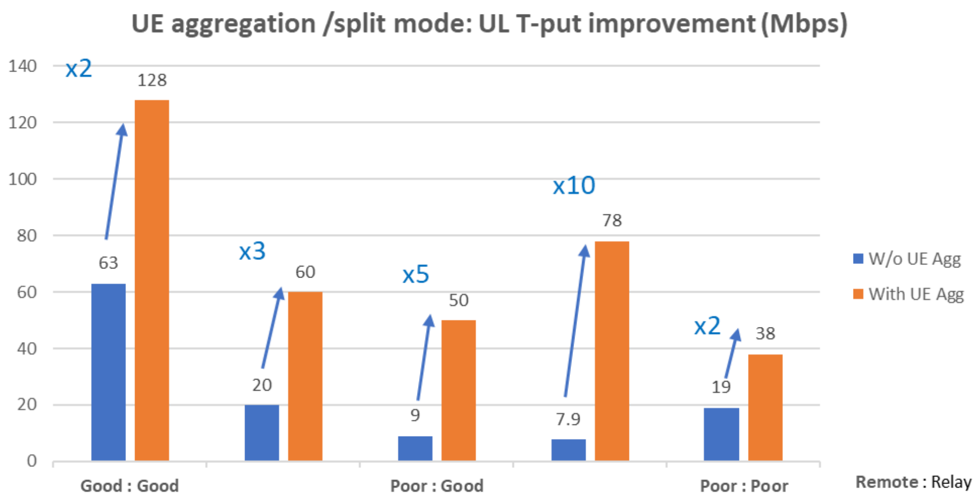

Figure 18.UE Agg/Spilt Mode UL T-put improvement

In duplication mode, the same PDCP PDU is sent on both legs and only the first arrival is accepted; when one leg suffers deep fading, retransmissions, or queue build-ups, the other leg still has a strong chance to "win the race," thereby curbing long-tail delays and reducing timeouts. Duplication does consume resources and device energy on both legs, so it is best applied selectively to reliability- or latency-sensitive flows, while split mode is the workhorse for bulk throughput, as depicted Figure 19.

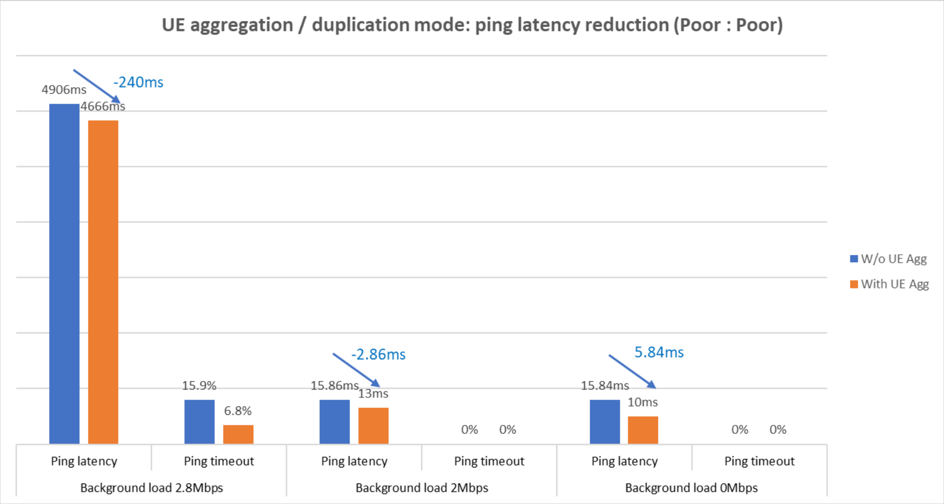

Figure 19.UE Agg/Duplication Mode Latency Reduction

5.7.4 Summary

UE aggregation at L2 (PDCP split/duplication) augments the uplink by pairing the Remote UE's direct path with an indirect path via a Relay UE. Experiments confirm the expected behavior, with higher throughput and lower latency under varied conditions.

In practice, N3C—typically Wi–Fi or Bluetooth—is the preferred inter–UE connection due to its ubiquity and low deployment friction. It requires only moderate adaptation on the inter–UE connection, preserves end–to–end PDCP security, and performs well when path selection, PDCP reordering, and duplication are sensibly tuned.

Current Release 18 constraints—same–gNB coordination and one Relay per Remote—are acceptable for near–term deployments, particularly at the cell edge and for busty uplink. Release 19 work is considering enhancements, including potential relaxation of the same–gNB constraint, which would broaden applicability if adopted.

Looking ahead, the path to network adoption is proving use cases compelling enough to justify operator investment. As high–value scenarios succeed, deployment will accelerate, and the inter–device connectivity ecosystem will deepen.

5.8 RedCap/eRedCap

5.8.1 Introduction

As a continuous evolution and key component of the 5G roadmap, 3GPP introduced the new 5G RedCap technology in the R17 standard—following eMBB and uRLLC terminals—because not all application scenarios require extreme performance under the unified 5G air interface. The capabilities provided by 5G technologies targeting the three main scenarios often exceed actual needs, leading to issues such as high terminal cost, high power consumption, and redundant network capacity, which become barriers to the widespread adoption of 5G. Therefore, in the second 5G evolution release, R17, the 5G RedCap device type was introduced.

RedCap (Reduced Capability), often referred to as "NR-Light," achieves its design objectives by reducing key technical parameters such as system bandwidth, peak data rate, number of antennas, and modulation order. This relaxation lowers the requirements on terminal RF transceivers and baseband processing capabilities, thereby reducing device cost and power consumption, enabling more devices to access 5G networks, promoting the popularization and application of 5G, and expanding new application fields for 5G.

R18 eRedCap is an evolved version of the R17 RedCap.By reducing the antenna capability to a single antenna and lowering the downlink peak rate to 10Mbps, it further reduces the terminal cost and power consumption, meeting the requirements of business scenarios with lower rates and lower power consumption.

5.8.2 Overview

The year 2025 is the first year of RedCap's large-scale development. The industry generally believes that it is necessary to continue to cultivate and support the commercial large-scale development of RedCap. On the one hand, the network still needs to be continuously upgraded and optimized to achieve the same coverage as 5G. On the other hand, larger-scale commercial shipments and product iterations of terminals and chips are required. Such as MIFI, wearable devices, lightweight terminals, in-vehicle communication, industry modules, etc.

5G RedCap is comparable to LTE Cat-4 in terms of technical capabilities and RedCap has obvious advantages. Firstly, the spectral efficiency has been improved. Under the same 20MHz bandwidth, the downlink peak rate has increased from 150Mbps to 220Mbps, and the uplink peak rate has increased from 50Mbps to 120Mbps. Secondly, the frequency range has been expanded from the traditional Sub-3GHz band to the 5G Sub-6GHz band, which is conducive to more extensive deployment by operators.

Meanwhile, RedCap technology inherits the native 5G features such as low latency and high reliability under the 5G unified air interface, network slicing, power saving, 5G LAN, high-precision positioning, high-precision timing, and small data packet transmission. By providing customized solutions to balance performance and cost, it meets the demands of more iot scenarios and significantly reduces the high cost of terminal devices. A research company stated that the price of RedCap modules is 80% lower than 5G eMBB modules, and they also save 90% of power consumption. With the maturity and upgrade of 5G network deployment, as well as the large-scale development of the terminal industry, 5G RedCap is expected to reach the same price level as Cat.4 in the future.

3GPP Release 18 introduces eRedCap (enhanced RedCap), which further simplifies 5G RedCap. eRedCap focuses solely on FR1 spectrum.Irrespective of other features that may be activated, an eRedCap device will only be able to achieve a peak throughput of 5Mbps in the uplink and 10Mbps in the downlink direction. By reducing further the peak data throughput for more constrained use cases, eRedCap reduces the costs and energy usage.

5.8.3 Performance Analysis

The Rel-17 RedCap UE supports 20 MHz, 1 Rx, 1 layer, DL 64QAM, UL 64QAM, FDD or TDD. In addition, cases with 2 Rx, 2 MIMO layers, and HD-FDD features are optionally evaluated.

	Simple eMBB FR1	RedCap FR1
Maximum transmit/receive bandwidth	100 MHz	20 MHz
Supported number of receive antenna branches and downlink MIMO layers	2 or 4	2 or 1
Supported number of transmit antenna branches and uplink MIMO layers	1	1
Maximum downlink modulation	256 QAM	256 QAM/64 QAM
Maximum uplink modulation	64 QAM	256 QAM/64 QAM
DL Peak Rates	500 Mbps	225 Mbps(FDD 256QAM)
UL Peak Rates	230 Mbps	120 Mbps(FDD 256QAM)

Table 3.Comparison of the simplest regular 5G device with RedCap

The peak rates of RedCap are sufficient to fulfill the requirements of most of the intended use cases.

Power-saving enhancement Release 17 implements for RedCap is the use of system frame number (SFN). It allows an extension of disconnected receiver time (eDRX) to a maximum of 10.24 seconds, reducing power consumption in devices with long idle periods during operation.

The Rel-17 RedCap UEs and the Rel-18 UEs being compared have the same number of antenna branches, the same number of layers, the same maximum supported modulation order, and the same duplex mode (among HD-FDD, FD-FDD, and TDD). Irrespective of other features that may be activated, an eRedCap device will only be able to achieve a peak throughput of 10 Mbps in the uplink or downlink direction. Take the eRedCap BW3/PR3 scheme as an example:

- Option BW3:

- 5 MHz BB bandwidth only for PDSCH (for both unicast and broadcast) and PUSCH with 20 MHz RF bandwidth for UL and DL.

- The other physical channels and signals are still allowed to use a BWP up to the 20 MHz maximum UE RF+BB bandwidth.

- Option PR3:

- Restriction of maximum number of PRBs for PDSCH and PUSCH.

- For 15 kHz SCS, the maximum number of RBs is 25.

- For 30 kHz SCS, the maximum number of RBs is 11 or 12.

- The restricted number of PRBs in Option PR3 is a hardcoded limit.

Reduced UE bandwidth	Rel-15 reference	Rel-17 RedCap	eRedCap BW3
RF: Total	**100%**	72.46%	72.43%
BB: Total	**100%**	26.26%	20.31%
RF+BB: Total	**100%**	**44.74%**	**41.15%**

Table 4.Average UE complexity reduction achieved by BW reduction options for FD-FDD 1Rx

Reduced UE bandwidth	Rel-15 reference	Rel-17 RedCap	eRedCap BW3
RF: Total	**100%**	52.04%	52.01%
BB: Total	**100%**	17.66%	13.68%
RF+BB: Total	**100%**	**31.41%**	**29.01%**

Table 5.Average UE complexity reduction achieved by BW reduction options for TDD 1Rx

eRedCap’s cost-complexity optimization focuses on BB domain narrowbanding and algorithm simplification. For high-volume, low-data-rate use cases, eRedCap’ss cost advantage will drive adoption in industrial IoT, smart cities, and wearables.

5.8.4 Conclusion and deployment suggestions

From a deployment perspective, RedCap networks are becoming increasingly mature. Using China market as an example, China Mobile has now completed the upgrade for continuous coverage of its 700MHz RedCap network, achieving county-level continuous coverage. Additionally, China Mobile is fully rolling out 2.6GHz RedCap in 2025 to enhance the quality of contiguous coverage and expand network capacity. China Telecom and China Unicom are deploying RedCap networks based on co-building and co-sharing, actively promoting deployment in over 300 cities in 2025 to achieve contiguous coverage in key areas.

However, overall, out of approximately 4.7 million 5G base stations in China, only about 1.7 million have been upgraded with RedCap capabilities. With the increase in the penetration rate of 5G site SA and the commercialization of RedCap features, RedCap will gradually replace 4G in the market share of medium-speed IoT and consumer-grade medium-speed scenarios. Operators need to upgrade their RedCap networks to achieve coverage parity with mainstream 5G.

Following the inaugural year of 5G RedCap commercialization in 2024, 2025 is poised to be a pivotal year for its scaled commercial expansion. Currently, various segments across the industry chain are actively advancing the large-scale development and further breakthroughs of 5G RedCap. There is a broad consensus within the industry that sustained cultivation and support for RedCap's commercial scaling remain essential. On one hand, continuous network upgrades and optimization are required to achieve coverage parity with mainstream 5G. On the other hand, greater volumes of commercial shipments and iterative improvements in terminals and chips—such as MiFi devices, wearables, lightweight terminals, in-vehicle communication systems, and industrial modules—are imperative for driving widespread adoption.

In terms of market drivers, vigorously expanding consumer-facing (ToC) use cases is critical. The ToC market—spanning wearables, smartphones, MiFi devices, and similar applications—not only represents a substantial scale capable of catalyzing industrial growth and reducing marginal costs across the supply chain, but it will also further spill over into and support ToB industry applications such as power, video surveillance, and vehicle-mounted communications. Ultimately, this synergy will contribute to a thriving ecosystem for the entire industry of 5G RedCap.

For eRedCap, with the 3GPP technical framework now relatively mature and well-defined, it is essential to establish corresponding industry standards to guide and steer relevant product development. Key tasks include conducting verification of critical technologies and device-network compatibility, as well as performing lab and field testing within trial networks. Additionally, the low-power consumption technologies of eRedCap—such as C-DRX (Connected Mode Discontinuous Reception), WUS (Wake-Up Signal), PEI (Paging Early Indication), PDCCH Skipping, and RRM measurement relaxation—require device-network collaboration and thorough technical validation.

Regarding terminals/chips, initial validation of key technologies and device-network compatibility can be conducted by upgrading existing commercial RedCap chips to support R18 eRedCap, thereby allowing networks to advance their upgrade and deployment.

However, supporting eRedCap technology necessitates the development of entirely new terminal chip architectures by enterprises. Chip R&D involves high investment costs and long development cycles, requiring clearer consensus across the industry. In particular, operators—as the primary demand drivers—must promote the continuous refarming of spectrum from 4G to 5G. Only when R18 network upgrades are deployed and the demand for migration from 4G Cat.1 to 5G eRedCap becomes more defined will the roadmap for commercial eRedCap chip development gradually come into focus.

With the further maturation and deployment of 5G networks, alongside the migration demand of LTE Cat.1 terminals from 4G to 5G, R18 eRedCap is expected to enter commercial use in the future. Together with existing technologies, it will help complete 5G's capability portfolio for medium- to low-speed connectivity scenarios.

5.9 NTN

5.9.1 Introduction

Satellite connectivity extends coverage beyond the limits of terrestrial networks, providing communication capabilities in areas where ground-based infrastructure is absent or impractical. It effectively complements terrestrial deployments by addressing their coverage limitations. Countries, companies, and institutes around the world have already set up satellite systems, e.g., GlobalStar, Inmarsat, Iridium, Starlink, Tiantong, and ChinaSat. However, different technologies are considered to be deployed for these satellite systems. It causes the market fragmentation.

Drawing from the great success of the terrestrial communication network, there is a growing consensus on the need for a unified global solution for satellite communications. As a global standard organization, 3GPP firstly introduced NR NTN and IoT NTN in Rel-17, which was built on NR, NB-IoT and eMTC, respectively. This standardization enables 5G devices — including handheld UEs and IoT UEs — to connect to satellite networks with appropriate enhancements.

5.9.2 Overview

In Rel-17 and Rel-18, transparent forwarding architecture is considered for NR NTN and IoT NTN, and one illustration is shown as Figure 20,which is also captured in 3GPP TR38.821.

Satellites only enable radio frequency-related functions. Thus, only low processing capability is required on satellite, which is beneficial to facilitate NR/IoT NTN rapid deployment.

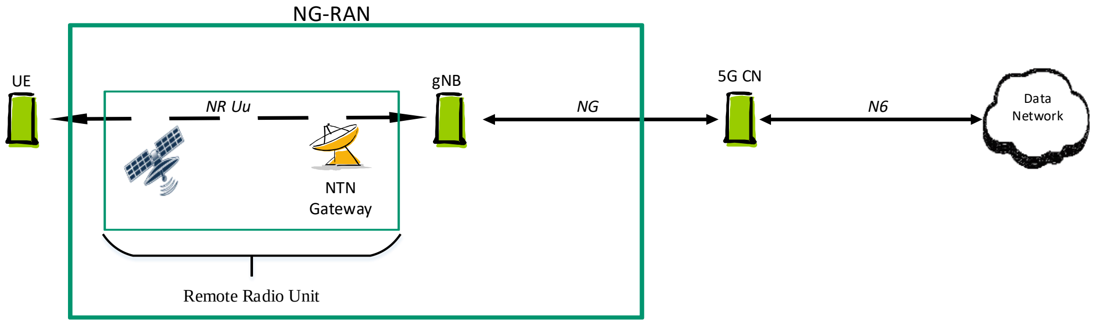

Figure 20.Networking-RAN Architecture with Transparent Satellite

Due to the long distance communication between satellites and terminals and the high mobility of satellites, the transmission delay and Doppler shift are much higher than that of terrestrial systems, which brings challenges to NTN. Thus, 3GPP Rel-17/Rel-18 NR/IoT NTN has done some enhancements in the air interface on top of NR/IoT/eMTC, including but not limited to the following aspects:

- Time/Frequency Synchronization: For the downlink, UE needs to detect downlink synchronization reference signals to perform time and frequency correction, as well as cell ID detection. In each NTN cell, ephemeris information and common Timing Advance (TA) parameters are broadcasted via SIBs. GNSS-capable UEs could calculate the Doppler frequency offset and the uplink timing advance on the service link based on the GNSS position information and the ephemeris information. As a result, UE could autonomously pre-compensate the Timing Advance corresponding to uplink synchronization reference point, as well as the frequency Doppler shift on service link for uplink transmission. The management of Doppler shift experienced over the feeder link is left to the network implementation.
- Timing management: In terrestrial mobile communication systems, the signal propagation delay is usually less than 1 ms. In the satellite communication system, the propagation delay depends on the altitude of the satellite and the type of satellite payload, generally ranging from tens to hundreds of milliseconds. To accommodate the propagation delay in NTNs, several timing relationships involving uplink and downlink timing interaction are enhanced by Koffset and kmac. Koffset is a configured scheduling offset that approximately corresponds to the TA of UE for uplink transmission. kmac is a configured time offset that approximately corresponds to the RTT between the uplink synchronization reference point (RP) and the gNB.
- HARQ enhancement: In legacy, a maximum of 16 HARQ processes per cell are supported. In addition, NW needs to wait for feedback from UE before sending new data. In the case of NACK, NW may need to retransmit data packets, which would result in introducing additional delays in the communication protocol. In terrestrial networks, where round-trip latency is typically within 1 ms, the impact of the stop-and-wait mechanism introduced by existing HARQ transmission is negligible. However, in NTN

systems, the transmission latency is significantly higher, leading to increased waiting time and reduced efficiency. To address this issue, 3GPP Release 17 NR NTN supports a maximum of 32 HARQ processes per NTN cell, allowing for more efficient scheduling of data transmissions during the long propagation delay. It also supports the deactivation of HARQ-ACK feedback for certain HARQ processes, thereby avoiding the need for a stop-and-wait mechanism and reducing transmission latency. Furthermore, in Release 18 IoT NTN, similar functionality is introduced to deactivate HARQ-ACK feedback, aiming to improve throughput and reduce latency in satellite-based IoT scenarios.

- Mobility: To enable mobility in NTN, the network provides serving cell's and neighboring cell's satellite ephemeris. For NR NTN, in Rel-17, basic functions of mobility are supported, including HO, CHO and cell (re)selection between NTN and NTN, NTN and TN; in Rel-18, mobility enhancements are considered to optimize performance, e.g., RACH less HO between NTN and NTN, cell (re)selection based on timing and location. For IoT NTN, due to limited time, mobility is not considered in Rel-17; in Rel-18, it supports UE in RRC connected state to perform neighbor RRM measurement.

Regarding voice communication via NTN, the industry has made significant efforts to develop GEO satellite based IoT NTN voice communication, primarily due to the large number of operational GEO satellites already in orbit. However, the IoT NTN communication system based on GEO satellites suffers from low air interface data rates and large propagation delays, both of which seriously affect the call setup delay and make it difficult to support high-bit rate traditional speech codecs, thereby degrading the user experience.

To address these challenges, a voice call solution based on IMS (IP Multimedia Subsystem) enhanced has been widely considered in the industry. Key enhanced technologies for IMS enhanced based Voice over IOT NTN may include:

- AI-based speech encoding/decoding technologies to maintain high voice quality under low-bandwidth and high-delay conditions.
- Reduce the size and number of SIP signaling in the process of establishing voice calls.
- Protocol stack optimization to support dedicated voice bearers and improve Quality of Service (QoS).
- Semi-static scheduling to reduce scheduling signaling overhead and improve resource utilization efficiency.

5.9.3 Performance Analysis

To assess the performance of the enhanced HARQ mechanisms under live network conditions, evaluations on both of NR NTN and IoT NTN were conducted, respectively.

For NR NTN, an evaluation on HARQ enhancement, i.e., the number of HARQ processes increased from 16 to 32, was conducted in a laboratory environment using a commercial UE platform, rather than relying on simulations. In the laboratory, the fully dynamic NTN channel conditions for LEO with transparent forwarding payload (about 30ms Koffset) is simulated by using a channel simulator. The evaluation was performed using Band n255 with a 20 MHz bandwidth.

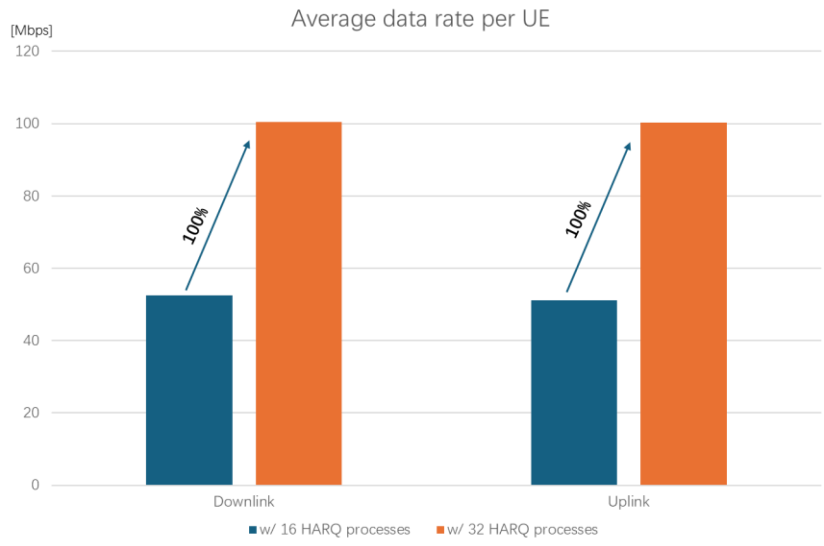

Figure 21.Average data rate per UE w.r.t. #HARQ processes

From this Figure 21, the performance gain is also nearly doubled when the number of HARQ Processes is doubled. It can be anticipated that even greater gains may be achieved with an increased number of HARQ processes, especially in scenarios where the propagation delay between the satellite and the UE is significantly large. This is because having more HARQ processes provides sufficient HARQ resources that can be flexibly utilized, rather than being constrained by waiting periods due to limited resources.

An evaluation on HARQ enhancement on IoT NTN, i.e., the mechanism of HARQ disabled, was conducted in a live trial using a commercial UE platform (not simulation-based). The UE was connected to a Geosynchronous Equatorial Orbit (GEO) constellation, namely the Tiantong system. The measured SNR was approximately 8 dB, and Band n256 was used. Other parameters included a 180 kHz downlink bandwidth and a 15 kHz uplink bandwidth, 2 HARQ processes, and a large TBS (Transport Block Size) of 2536 bits.

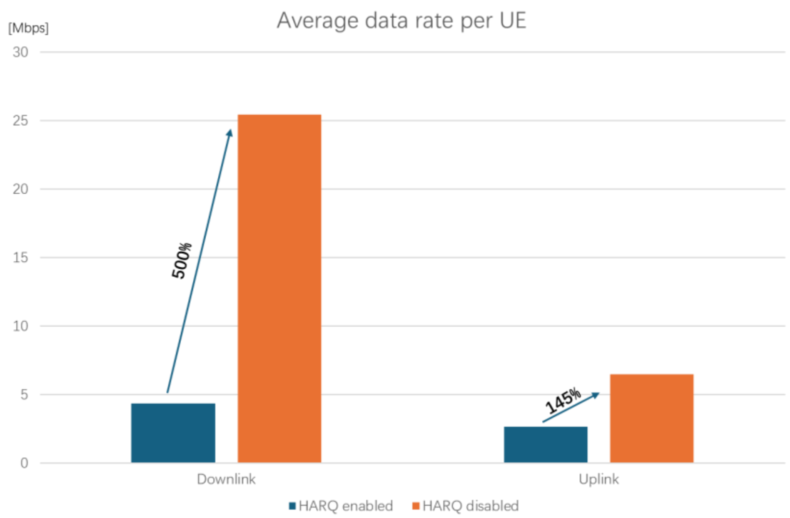

Figure 22. Average data rate per UE w.r.t. HARQ enabled/disabled

As can be seen from Figure 22, the performance gain achieved by deploying the

HARQ-disabled mechanism is significant. For the downlink, the average data rate per UE increases by nearly 500%, whereas for the uplink, the improvement is approximately 145%. By disabling HARQ feedback, the system can release the constraint of waiting for an available HARQ process, thereby improving resource utilization, and reducing transmission latency.

In the laboratory, a voice communication test environment based on IoT NTN was established between two VoNTN (Voice over NTN) terminals using a channel simulator and base station equipment loaded with a high orbit satellite model. This setup was used to evaluate the effectiveness of the voice call solution based on IMS enhanced.

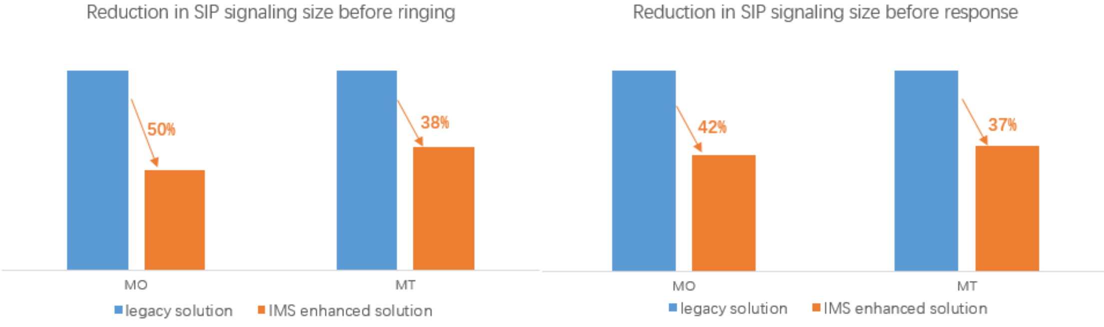

Figure 23. Reduction in SIP signalling size based on IMS enhanced solution

Figure 23 illustrates that, compared to legacy voice call solution, the voice call solution based on IMS enhanced significantly reduces the amount of SIP signaling data exchanged between the mobile originating (MO) terminal and the network by reducing the size of each SIP signaling. Specifically, On the mobile originating (MO) side ,the SIP data exchange before ringing is reduced by 50%, and data exchange before call answer is reduced by 42%. On the mobile terminating (MT) side, the SIP data exchanged before ringing decreases by 38%, and the data exchanged before responding drops by 37%. As a result of this reduction in SIP signaling data, the voice call setup time is also notably shortened. Test data indicates that with the proposed enhanced solution, the average voice call establishment time has been brought down to within 30 seconds.

5.9.4 Summary

The introduction of NR NTN and IoT NTN has significantly enriched the cellular communication ecosystem, enabling a single device to achieve ubiquitous global connectivity from virtually anywhere. Field evaluations have demonstrated that, with the implementation of advanced technologies, performance and user experience can be effectively guaranteed. These innovations encourage operators to deploy NTN networks that allow direct satellite access for UEs. Currently, the UE-side solution has reached a stage of commercial readiness, paving the way for widespread adoption.

Looking ahead, both NR NTN and IoT NTN are expected to continue evolving. In Release 19, enhancements include capacity improvements (e.g., the use of Orthogonal Cover Codes for data channels), coverage enhancements (e.g., repetition transmission for common channels), and the extension of SSB periodicity (i.e., to 160 ms). Meanwhile, the IoT NTN voice communication solution based on GEO satellites is currently being standardized in Release 20. As NTN technologies continue to advance, they are expected to introduce more enhanced features to address a wider range of diverse service scenarios.

6 Conclusion

As 5G coverage expands and Standalone deployment accelerates, 5G‑Advanced is moving from specification to execution. From a chipset perspective, there're two directions of particular importance for devices: improving in performance and power efficiency, and diversification of device and link portfolios.

Ecosystem triggers will shape chipset roadmaps and timelines. Operator commitments and rollouts of Release 17/18 SA features, agreed device capability profiles, and clear spectrum plans provide the demand signals that determine which features are prioritized and when they land. Accordingly, this whitepaper reports validated, measurable device‑side gains for key features on commercial or prototype hardware, and seeks to support industry convergence around practical deployment guidance, harmonized device capability profiles, and acceptance KPIs.

Looking ahead, scaling 5G-Advanced will depend on common playbooks and rigorous, end-to-end interoperability and certification to turn promising features into consistent, network-wide quality. With harmonized implementation and deployment profiles—anchored by acceptance KPIs—and shared configuration guidance, complemented by realistic lab and field validation, the ecosystem can move in step so that today's triggers translate into predictable roadmaps, credible adoption timelines, and reliable user-level gains.

7 References

1. 3GPP TR 21.917 Release 17 Description; Summary of Rel-17 Work Items (version 17.0.0 Release 17)
2. 3GPP TR 21.918 Release 18 Description; Summary of Rel-18 Work Items (version 18.0.0 Release 18)
3. 3GPP TS 38.211 NR; Physical channels and modulation (version 18.7.0 Release 18)
4. 3GPP TS 38.212 NR; Multiplexing and channel coding (version 18.7.0 Release 18)
5. 3GPP TS 38.213 NR; Physical layer procedures for control (version 18.7.0 Release 18)
6. 3GPP TS 38.214 NR; Physical layer procedures for data (version 18.7.0 Release 18)
7. 3GPP TS 38.306 NR; User Equipment (UE) radio access capabilities (version 18.7.0 Release 18)
8. 3GPP TS 38.331 NR; Radio Resource Control (RRC); Protocol specification (version 18.7.0 Release 18)
9. 3GPP TS 38.101-1 NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (3GPP TS 38.101-1 version 17.7.0 Release 17)
10. V. Gupta, B. Sundaramani, S. Maximov, M. A. El-saidny, D. Verma:"5G Uplink Enhancement Techniques: A Comprehensive Evaluation of 5G Uplink Transmit Switching & Uplink with Three-Antenna Transmission Solutions", https://www.mediatek.com/whitepapers
11. 3GPP TS 38.215 NR; Physical layer measurements (version 18.4.0 Release 18)

12. 3GPP TS 38.133 NR; Requirements for support of radio resource management (version 18.11.0 Release 18)
13. "The only way is up for Vodafone and MediaTek as they test 5G uplink speeds of 277 Mbps", https://www.vodafone.com/news/newsroom/technology/the-only-way-is-up-for-vodafone-and-media-tek-as-they-test-5-g-uplink-speeds-of-277-mbps
14. "爱立信携手联发科技完成 IMT-2020（5G）推进组 LTM 技术测试，引领 5G-A 低时延移动性新范式", https://www.ericsson.com/zh-cn/press-releases/2/2025/11/imt-2020-5g-ltm-5ga
15. MediaTek," Improving FWA Performance with Advanced Congestion Avoidance Techniques", https://www.mediatek.com/tek-talk-blogs/improving-fwa-performance-with-advanced-congestion-avoidance-techniques
16. W.Coomans, K.De Schepper, O.Tilmans L4S," low latency, low loss, and scalable throughput", https://www.nokia.com/asset/213410/
17. [RFC9330] B. Briscoe , K. De Schepper, M. Bagnulo and G. White, "Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture", https://www.rfc-editor.org/rfc/rfc9330.html
18. [RFC9331] K. De Schepper and B. Briscoe, "RFC 9331 The Explicit Congestion Notification (ECN) Protocol for Low Latency, Low Loss, and Scalable Throughput (L4S)," https://www.rfc-editor.org/rfc/rfc9331.html
19. [RFC9332] K. De Schepper, B. Briscoe and G. White, "RFC 9332 Dual-Queue Coupled Active Queue Management (AQM) for Low Latency, Low Loss, and Scalable Throughput (L4S)," https://www.rfc-editor.org/rfc/rfc9332.html
20. 3GPP TS 38.323 NR; Packet Data Convergence Protocol (PDCP) specification(version 18.5.0 Release 18)
21. "中国移动联合产业伙伴完成业界首个 R18 UE 聚合技术端到端能力验证", https://www.cww.net.cn/article?id=604652
22. 3GPP TR38.875 Technical specification group radio access network; study on support of reduced capability NR devices (Release 17)
23. 3GPP TR38.865 Technical Specification Group Radio Access Network;Study on further NR RedCap UE complexity reduction(Release 18)
24. "中国移动携手中兴通讯、紫光展锐完成全球首个手机直连高轨卫星 NTN 语音通话实验室验证", https://www.10086.cn/aboutus/news/groupnews/index_detail_50179.html