5G/6G Academy
TELECOM CERTIFICATIONS · SINCE 2009

Technical Whitepaper
N° 05 · June 2026

TECHNICAL WHITEPAPER

URLLC Reality Check: The Millisecond, Audited

What the 1 ms figure actually measures, why even the standard's own QoS table stops at 5 ms, what it really takes to go faster — and what operators are honestly selling in 2026 instead.

AUTHOR	VERSION	DATE
5G/6G Academy Research	1.0	June 2026

IMT-2020 · 5QI 82–90 · GRANT-FREE UL · TSN · SURVIVAL TIME · XR · L4S

5g6gacademy.com/whitepapers

5G/6G ACADEMY · TECHNICAL WHITEPAPER · N° 05 June 2026 · 5g6gacademy.com

URLLC Reality Check: The Millisecond, Audited

5G/6G Academy Research

a TELCOMA Global publication · Telecom Certifications · Since 2009

***Abstract*—The most famous number in 5G is also its most misquoted. IMT-2020's 1 ms is the radio interface's contribution alone — one way, unloaded, a single user, a 32-byte packet — and 3GPP's own QoS table never promises anything tighter than 5 ms between the device and the core's edge, with up to 5 ms of that reserved for the network behind the radio. This paper audits the millisecond: where the latency budget actually goes (the scheduling handshake, not the radio, is the biggest block); what the Rel-15/16 toolbox — configured grants, PDCP duplication, pre-emption — buys and what it costs; why survival time and TSN integration turn impossible per-packet five-nines into engineerable availability; and what is genuinely for sale in 2026: bounded-jitter products at tens of milliseconds, slices, and L4S. We found no documented commercial public-network service at 1 ms / 99.999% as of June 2026 — and we explain why the XR-class 5QIs at 5–20 ms are where the low-latency market is actually landing.**

1. The Millisecond That Never Was

Read the source document and the famous figure shrinks before your eyes. ITU-R's IMT-2020 requirements define user-plane latency as "the contribution of the radio network" — measured one way, from layer-2/3 packet ingress to egress across the radio interface, "in unloaded conditions (i.e. a single user) for small IP packets," with 1 ms required for URLLC and 4 ms for ordinary broadband [1]. The reliability requirement is similarly precise and similarly narrow: a $1–10^{-5}$ probability of delivering a 32-byte layer-2 packet within 1 ms, evaluated at the coverage edge of one specific test environment [1]. That is the whole promise: one radio hop, one user, no load, one small packet.

Everything else in the path — the transport network, the core, the application server, the return trip — was never inside the number. 3GPP's NR design targets (TR 38.913) are commonly cited as 0.5 ms one-way in each direction for URLLC [6], which is consistent: the radio was designed to spend only half the ITU budget. The gap between "1 ms radio contribution" and "1 ms application experience" is not a failure of engineering; it is a failure of quotation. This paper is about what lives in that gap.

2. Even the Standard Does Not Promise 1 ms

Here is the quiet irony at the center of URLLC: 3GPP's own QoS catalogue contains no 1 ms class. The delay-critical 5QIs — the rows of TS 23.501's Table 5.7.4-1 that exist precisely for automation, smart grid and vehicles — bottom out at a 5 ms packet delay budget, and that budget is defined between the device and the UPF's N6 interface, not across the radio alone [2] (Table 1).

TABLE 1 · The delay-critical 5QIs — what the standard actually promises (TS 23.501, Table 5.7.4-1) [2]

5QI	Delay budget (UE ⇔ UPF)	Packet error rate	Max data-burst volume	Example services (from the spec)
82	10 ms	10^{-4}	255 B	discrete automation (small packets)
83	10 ms	10^{-4}	1354 B	discrete automation
84	30 ms	10^{-5}	1354 B	intelligent transport systems
85	5 ms	10^{-5}	255 B	electricity distribution, remote driving
86	5 ms	10^{-4}	1354 B	V2X collision avoidance
87–90	5 / 10 / 15 / 20 ms	10^{-3}	up to 63 kB	XR: motion tracking, split rendering (Rel-17/18)

It gets tighter. From each budget, a static core-network share — 1, 2 or 5 ms depending on the class — is set aside for the path behind the radio, leaving the air interface with what remains [2]. A 5QI 85 flow with its 5 ms budget is really a radio budget of about 4 ms plus a core allowance. If your application genuinely needs 1 ms end-to-end, the standardized QoS table has nothing to sell you: you are, by construction, outside it — in the land of private deployments, bespoke configurations, and the toolbox of Section 4.

Notice also what the newest rows say about where 3GPP thinks the demand is. The XR classes added in Rel-17/18 — 5QIs 87 through 90 — target 5 to 20 ms with large burst volumes for split rendering and motion tracking [2]. The standard's own growth area for "low latency" sits an order of magnitude above the headline number.

3. Where the Milliseconds Go

Follow one uplink packet through a sub-6 GHz TDD cell and the budget spends itself fast (Fig. 1). The device processes the packet, waits for an uplink opportunity in the TDD pattern, then — in the default, grant-based procedure — sends a scheduling request and waits for the network's grant before it can transmit at all. Peer-reviewed measurement work from EPFL and RWTH traces exactly this path and finds the handshake costs a full TDD period; on their testbed (n78, 0.5 ms slots, a DDDU pattern), one-way latencies measured roughly 1–3 ms downlink and 2–6 ms uplink [5].

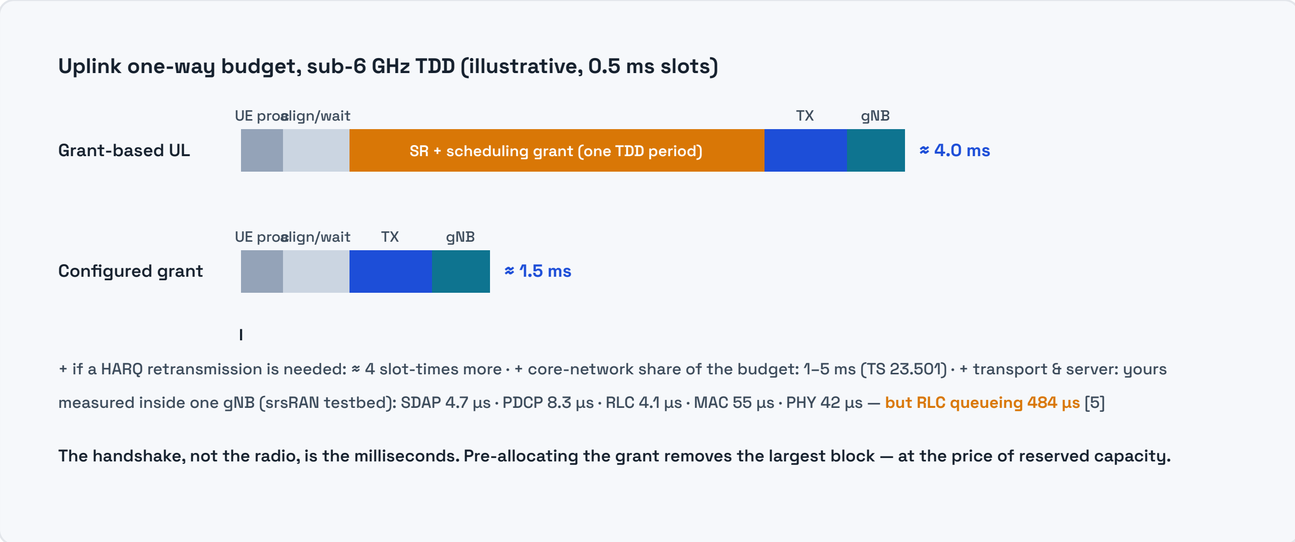

Fig. 1. An illustrative uplink budget on sub-6 TDD, after Maghsoudnia et al. [5] and TS 23.501 [2]. The scheduling-request-and-grant handshake dominates; a configured grant removes it entirely. Inside the gNB, protocol processing is tens of microseconds — queueing is what costs.

Two measured details deserve emphasis. First, the protocol layers themselves are nearly free — SDAP, PDCP and RLC each processed packets in well under ten microseconds on the testbed — but *queueing* in RLC alone averaged 484 µs: half the headline budget spent standing in line inside one node [5]. Second, the same paper's worst-case analysis of minimal TDD configurations found that meeting 0.5 ms one-way in both directions required the most uplink-friendly pattern with grant-free transmission; relax to FDD, or allow 0.25 ms mini-slots, and the target comes within reach by analysis — but their actual over-the-air testbed, on a realistic pattern, did not hit it [5]. Add the HARQ question — a single retransmission costs roughly four slot-times — and the engineering truth crystallizes: **the radio is fast; the waiting is slow**. Latency work is queue work, schedule work, and handshake work.

4. The Toolbox and Its Price

3GPP built the tools to attack each block of Fig. 1, mostly in Rel-15 and Rel-16 [4], [13]. Against the handshake: **configured grants** — pre-allocated uplink resources a device may use without asking, up to twelve configurations per bandwidth part. Against collisions with ordinary traffic: **downlink pre-emption** (a control message telling broadband devices which of their resources were punctured for an urgent packet) and **uplink cancellation**. Against decoding failures: compact, more-robust control formats, sub-slot feedback so a NACK doesn't wait for the slot boundary, and back-to-back mini-slot repetitions that cross slot boundaries. Against single-path failure: **PDCP duplication** over up to four legs, redundant PDU sessions over independent paths, and dual transport tunnels — the same bits flying multiple routes simultaneously [4].

Every tool pays in the same two currencies. Pre-allocated grants reserve capacity whether or not it is used, so grant-free uplink does not scale to a stadium of devices — it scales to a production cell of them. Duplication multiplies the radio resources per delivered packet by up to four. Pre-emption taxes everyone else's throughput. The honest summary the marketing never printed: **latency, reliability, capacity — pick two**. URLLC engineering is the art of choosing which two, per square meter of factory.

5. Determinism, Not Speed: TSN and Survival Time

The industrial world's actual requirement is usually not "fast" but "on time, every time" — and the standard's response is its most under-reported part. The 5G system can present itself to an IEEE 802.1 TSN network as a **logical TSN bridge**: translators at the device side (DS-TT) and network side (NW-TT) implement hold-and-forward buffering, so packets leave on the TSN schedule regardless of when the radio delivered them — jitter is absorbed, not just minimized. The 5G system participates in 802.1AS time synchronization as a time-aware system, with a clock budget of at most 900 ns across the whole 5G span [2], [3].

The conceptual hero is **survival time**: how long an application can tolerate silence before it must act. TS 22.104's motion-control rows make it concrete — 50-byte messages every 500 µs with a survival time of one interval, for up to 20 devices in a 50×10 m cell; relax to 2 ms cycles and a hundred devices fit [3]. The arithmetic this enables is the whole game: if the application survives one lost burst, then 99.9% per-packet reliability becomes 99.9999% *service availability* — the spec's own conversion [3]. And since Rel-17, the scheduler is told the survival time (via the TSC assistance information) so it can spend robustness only when a burst has just been lost [2]. Five nines stops being a per-packet miracle and becomes a state machine.

6. Reality Check: What 2026 Actually Sells

The peer-reviewed verdict is blunt: four-plus years into commercial 5G, "real-world implementations of URLLC satisfying the specified requirements remain elusive," achievable "under very specific circumstances with stringent hardware and software conditions" [5]. The measurement record, compiled in the same work and labeled here with its vintage (2020–2023), tells the story graphically (Fig. 2): the best published number is a hardware-accelerated, single-device lab link at 0.8 ms downlink; demos cluster at 2–5 ms; campus networks measure 2–8 ms one-way; and commercial millimeter-wave achieved sub-millisecond delivery for only 4.4% of packets in one study — a reminder that the second R, reliability, is the harder half [5], [7].

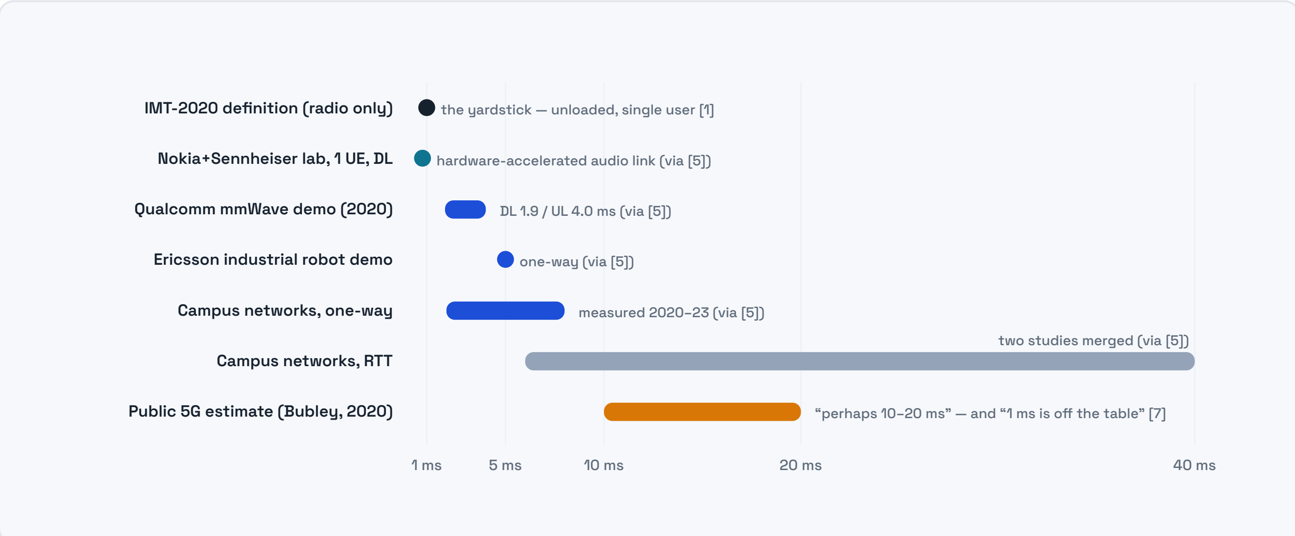

Fig. 2. The yardstick versus the record: published latency results, 2020–2023, as compiled in [5] (originals not independently re-measured), against the IMT-2020 definition [1] and a 2020 public-network estimate [7]. Points are single results; bars are reported ranges.

What do operators sell instead? Honest products at honest numbers. Ericsson's time-critical-communication framing studies a ~20 ms one-way RAN budget at 99% reliability inside a 20–40 ms end-to-end budget — service-requirement framing, not an SLA [8]. T-Mobile US switched on L4S nationwide in July 2025, the first US carrier, marketing *consistency* for cloud gaming, XR and remote driving — with no millisecond figure attached [9]. European operators sell slices with guaranteed bitrates and priority [14]. Even T-Mobile's own URLLC explainer concedes the capability "will likely show up in private 5G networks first" [10]. As of June 2026, we found no documented commercial public-network service offering 1 ms at 99.999% — a negative we state carefully and invite readers to falsify; the likeliest counterexamples are Chinese deterministic-networking and power-grid pilots whose public documentation we could not verify [5], [16].

7. The Pragmatic Middle Ground — and What To Do

The market found its level an order of magnitude above the slogan. Rel-18's XR work — PDU Sets that treat a video frame as one QoS object, delay budgets and error rates per set, L4S congestion marking, an XR-aware RAN — is the standard's biggest current investment in latency, and it lives at 5–20 ms [2], [11], [12]. That is not URLLC's failure; it is the demand curve speaking. Interactive media, cloud gaming and tele-operation are real, paying, and satisfied by bounded jitter at tens of milliseconds.

For engineers scoping a genuinely tight application, the audit reduces to five rules. **Quote budgets end-to-end**, and ask every vendor which segment their number covers — the 1 ms confusion survives only because nobody asks. **Buy determinism, not speed**: specify survival time and availability, then let the network spend its tools (configured grants, duplication) only where the state machine demands. **Assume private or dedicated infrastructure** for anything below ~10 ms end-to-end; the public-network record supports nothing tighter. **Design the TDD pattern with the uplink in mind** — the frame structure, not the technology generation, sets the floor. And **treat reliability as the budget item**: sub-millisecond sometimes is easy; sub-millisecond always is the entire discipline. The millisecond was never a lie — it was a laboratory sentence quoted as a service contract. Engineer to the sentence that was actually written, and 5G's low-latency machinery is genuinely excellent.

References

[1] ITU-R, Report M.2410-0, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," Nov. 2017, §4.7, §4.10. itu.int

[2] 3GPP, TS 23.501 (ETSI TS 123 501 V19.7.0, Mar. 2026), §5.7.3–5.7.4 (QoS, 5QI table), §5.27–5.28 (TSC/TSN). etsi.org/deliver

[3] 3GPP, TS 22.104 (ETSI TS 122 104 V19.2.0, Jan. 2026), "Service requirements for cyber-physical control applications," Tables 5.1-1, 5.2-1, §5.6. etsi.org/deliver

[4] 3GPP, TR 21.916 (ETSI TR 121 916 V16.2.0), "Release 16 description," §5.2 (URLLC PHY), §5.3 (NR IIoT). etsi.org/deliver

[5] A. Maghsoudnia, E. Marin, D. Dimitrova et al., "Ultra-reliable low-latency in 5G: a close reality or a distant goal?," ACM HotNets 2024. conferences.sigcomm.org/hotnets/2024

[6] 3GPP, TR 38.913 V18.0.0, "Study on scenarios and requirements for next generation access technologies" (0.5 ms one-way design target). etsi.org/deliver

[7] D. Bubley, "Low-latency and 5G URLLC — a naked emperor?," Disruptive Wireless blog, Sep. 2020. disruptivewireless.blogspot.com

[8] Ericsson Technology Review, "Advanced connectivity solutions for 5G time-critical communication." ericsson.com/en/reports-and-papers

[9] T-Mobile US, "T-Mobile is first to unlock L4S in wireless," newsroom, Jul. 2025. t-mobile.com/news/network

[10] T-Mobile for Business, "5G URLLC brings the potential for near-real-time communications." t-mobile.com/business/resources

[11] 3GPP, "5G architecture support for XR and media services (XRM, Rel-18)," technology briefing. 3gpp.org/technologies/xr-sa2

[12] Anon., "Overview of NR enhancements for extended reality (XR) in 3GPP 5G-Advanced," arXiv:2412.00741. arxiv.org/html/2412.00741v1

[13] Techplayon, "5G NR DCI formats in detail" (DCI 2_1 pre-emption per TS 38.213 §11.2). techplayon.com

[14] GlobalData IT Connection, "5G network slicing services launch with increasing frequency — but what exactly are they offering?," Sep. 2025. itcblogs.currentanalysis.com

[15] R. Abreu et al., "On the resource utilization of multi-connectivity transmission for URLLC services in 5G New Radio," arXiv:1904.07963. arxiv.org

[16] 5G/6G Academy editorial note: negative findings reflect a targeted June 2026 search of operator and analyst sources; corrections welcome at 5g6gacademy.com.