gNB-based Local Breakout for URLLC in industrial 5G

Rajendra Paudyal*, Rajendra Upadhyay†, Al Nahian Bin Emran‡, Duminda Wijesekera†

Department of Computer Science*, Department of Cyber Security†, Department of Information Technology‡

George Mason University

Fairfax, VA, USA

*rpaudyal@gmu.edu, †rupadhya@gmu.edu, ‡abinemra@gmu.edu, †dwijesek@gmu.edu

Abstract—**Industrial URLLC workloads-coordinated robotics, automated guided vehicles, machine-vision collaboration require sub-5 ms latency and five-nines reliability. In standardized 5G Multicast/Broadcast Services, intra-cell group traffic remains anchored in the core using MB-SMF/MB-UPF, and the Application Function. This incurs a core network path and packet delay that is avoidable when data transmitters and receivers share a cell. We propose a gNB-local multicast breakout that pivots eligible uplink flows to a downlink point-to-multipoint bearer within the gNB, while maintaining authorization, membership, and policy in the 5G core. The design specifies an eligibility policy, configured-grant uplink. 3GPP security and compliance are preserved via unchanged control-plane anchors. A latency budget and simulation indicate that removing the backhaul/UPF/AF segment reduces end-to-end latency from ≈6.5-11.5 ms (anchored to the core) to ≈1.5-4.0 ms (local breakout), producing sub-2 ms averages and a stable gap ≈10 ms between group sizes. The approach offers a practical, standards-aligned path to deterministic intra-cell group dissemination in private 5G. We outline multi-cell and prototype validation as future work.**

Index Terms—**5G, URLLC, NG-RAN, 6G, multicast, MBS, Industry 4.0, Industry 5.0, private 5G, MB-UPF, local breakout.**

I. Introduction

Industry 4.0 and emerging Industry 5.0 settings blend cyber-physical systems (CPS), AI, and digital twins to achieve adaptive, resilient, and efficient manufacturing. Examples include: (i) coordinated multi-robot assembly with sub-frame motion updates; (ii) UAV swarms requiring fast group state dissemination; (iii) machine vision alarms that must reach multiple controllers within a control cycle and (iv) human robot collaboration where safety relevant signals require deterministic delivery [1], [2]. Private 5G deployments are attractive due to licensed-grade interference resilience, mobility, and QoS controls. However, for *intra-cell* group communication, the standard 5G Multicast/Broadcast Service (MBS) model routes multicast through MB-UPF in the core [3]. Even on a campus with short fiber backhaul, traversing gNB → UPF / MB-UPF → AF / application → gNB adds queueing and processing that can undermine URLLC targets [4], [5].

We address this gap with *RAN-local multicast breakout*. The gNB forwards specific uplink flows from group members to a downlink multicast bearer without sending payloads to the core. Only the user-plane packets are rerouted towards the downlink multicast channel, where all other control-packets are handled by the core. This retains core-driven authorization and session control. Only control-plane packets are routed towards the core by the gNB; data plane qualified packets qualify for local breakout are rerouted by the gNB to the multicast group. This reduces latency to single-digit milliseconds for intra-cell groups. The concept of local breakout is taken from the paper [6].

II. Background

A. MBS in 5G

In the Third Generation Partnership Project (3GPP) Release 17, MBS are introduced as a key enhancement to the 5G system architecture. The release established efficient PTM communication for applications such as public safety, vehicle-to-everything (V2X), Internet Protocol Television (IPTV), and group communications. This architecture builds on the existing 5G framework defined in Technical Specification (TS) 23.501 by incorporating new network functions to support both multicast and broadcast modes. It emphasizes resource efficiency and compatibility with non-roaming scenarios [7], [3], [8]. Specifically, the Multicast/Broadcast Session Management Function (MB-SMF) is responsible for session control, including the creation, activation, deactivation, modification and deletion of MBS sessions, as well as the allocation of Temporary Mobile Group Identities (TMGIs), the derivation of Quality of Service (QoS) parameters, and coordination with the Access and Mobility Management Function (AMF) for the allocation of resources from the Radio Access Network (RAN). Complementing this, the Multicast/Broadcast user-plane Function (UPF) serves as the MBS session anchor. It handles user-plane replication and distribution by receiving a single copy of MBS data from the Application Function (AF) or Multicast/Broadcast Service Transport Function (MBSTF)
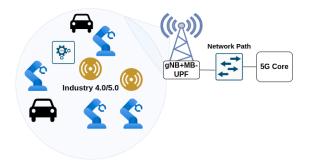
Fig. 1. Architectural overview of local multicast

and forwarding it using General Packet Radio Service Tunneling Protocol User Plane (GTP-U) tunnels to Next Generation RAN (NG-RAN) nodes or other UPFs for downstream delivery.

RAN plays a key role in optimizing delivery on the radio interface, dynamically selecting between PTM and point-to-point (PTP) modes per User Equipment (UE) based on factors such as UE density, radio conditions, and mobility [9]. In PTM mode, the NG-RAN transmits a single copy of MBS data over a Multicast Radio Bearer (MRB) to multiple UEs, leveraging multicast transport for shared delivery. This enhances spectral efficiency in scenarios with high UE concentrations [10], [3]. In contrast, the PTP mode involves separate transmissions to individual UEs via unicast Protocol Data Unit (PDU) sessions, ensuring reliability in sparse or edge coverage areas, with seamless switching between modes to maintain service continuity during handovers [9]. This architecture demonstrates robustness and scalability across multiple cells, supporting features like location-dependent services, multiple QoS flows (Guaranteed Bit Rate and non-Guaranteed Bit Rate), and mobility handling through transitions between 5G Core (5GC) shared and individual delivery methods.

However, a notable limitation lies in the centralization of user-plane replication within the core network at the MB-UPF, which acts as a centralized anchor for data ingress and distribution, even when UEs are co-located within a single cell [3]. While this design reuses existing 5G entities to minimize deployment costs and supports efficient multicast transport where available, it can introduce inefficiencies in localized scenarios, as replication for individual delivery or unicast tunnels occurs in the core rather than being distributed closer to the RAN, potentially increasing latency and backhaul load [9].

B. URLLC for Industrial Control

In the context of 5G and beyond networks, 3GPP has established stringent service requirements for cyber-physical control applications. Particularly in vertical domains such as industrial automation, where URLLC

is paramount to ensure deterministic performance with end-to-end latencies as low as 5 ms and reliabilities exceeding 99.9999% for packet transmissions [1]. These requirements, outlined in TS 22.104, address scenarios involving real-time feedback loops in machine-to-machine interactions. It emphasizes the need for robust communication services to support closed-loop control systems without compromising safety or efficiency. The foundational building blocks of URLLC in 5G are shortened transmission time intervals (TTIs) by reduced slot durations, configured-grant enabling grant-free uplink access to minimize scheduling delays, packet duplication across multiple paths for enhanced reliability, robust modulation and coding schemes (MCS) to combat channel impairments, and mini slots for flexible sub slot transmissions [11]. However, extending these mechanisms to group dissemination scenarios, such as multicast URLLC (mURLLC) for synchronized control of device clusters, introduces additional complexities in feedback handling and scheduling, as dynamic resource allocation must balance individual user conditions with group wide efficiency, potentially increasing computational overhead and risking latency violations in dense environments.

Figure 1 illustrates an Industry 4.0/5.0 application connected to the gNB with local user-plane function "gNB+MB-UPF". Industrial endpoints inside the cell (robots, AGVs, controllers, vehicles) generate group updates that, when policy and membership allow, are pivoted at the gNB directly into a downlink point-to-multipoint (PTM) bearer and transmitted once to all in-cell receivers bypassing the network path to the 5G Core. While control-plane functions remain anchored in the core. This avoids the gNB → UPF/AF → gNB path, shortening delivery to UE → gNB → group UEs and driving the latency gains.

C. Local Breakout

Local breakout in cellular networks is a pivotal strategy to optimize traffic routing by anchoring data flows at the network edge. It minimizes backhaul traversal and reduces end-to-end latency. In private 5G deployments, Local breakout enables localized processing by integrating media and control-plane functions directly at the gNB as defined in 3GPP TS 23.501. This approach is particularly advantageous for latency-sensitive applications such as industrial IoT. The proximity to the data source improves performance while alleviating core network congestion. This pivoting traffic flow path reduces backhaul load and supports improving scalability. Furthermore, 3GPP TS 33.501 ensures that security and policy enforcement, including authentication, authorization, and QoS management, are preserved at the core through standardized mechanisms like the Multicast

Broadcast Session Management Function (MB-SMF) and Policy Control Function (PCF).

III. RELATED WORK

Several efforts have addressed reducing core traversal for multicast or group-based delivery in 5G. Völk et al. [12] proposed a Core–Edge Split EPC architecture for public safety networks, where essential core functions are deployed at the network edge in mobile base stations. This allows intra-cell communications to continue locally even when backhaul is unavailable, thereby avoiding core traversal and lowering latency for group voice/video traffic. The feasibility of such base-station–level routing aligns with our gNB-local breakout concept. In one of the earliest 5G Mobile Edge Computing (MEC) trials, Zhang et al. [13] quantified the latency benefits of offloading to a local MEC server compared to a remote core. Although MEC reduced one-way latency to about 17 ms, strict URLLC targets less than 5 ms remained out of reach, motivating further path shortening such as gNB-level forwarding. From the standards perspective, 3GPP TR 23.757 [14] examines Release 18 enhancements for MBS, including the concept of local MBS service where multicast delivery is restricted to a cell or tracking area. The study discusses placing multicast anchors closer to the RAN and even enabling RAN-based switching between unicast and multicast modes, reinforcing the idea that core traversal can be bypassed for local group traffic. Coll-Perales et al. [11] analytically model end-to-end latency in V2X deployments with varying application server placement. Their results show centralized cloud deployments incur tens of milliseconds more delay than edge-hosted ones, confirming that moving processing and replication closer to the RAN yields substantial latency gains. This mirrors the ≈ 10 ms improvement we observe when replacing core-anchored multicast with gNB-local breakout. Complementing these approaches, Säily et al. [15] developed a 5G NR RAN-controlled multicast architecture (5G-Xcast) with dynamic unicast/multicast switching and minimal 5GC impact. Their design demonstrates that multicast replication and group control can reside largely within the RAN.

Our work differs from these prior efforts in three key ways. First, unlike MEC approaches that still require an extra hop through a local core anchor, our method collapses the intra-cell data path entirely into the gNB while preserving 3GPP control-plane compliance. Second, while standards studies and RAN-centric frameworks consider multicast at or near the RAN, we explicitly target deterministic URLLC group control in industrial private 5G settings, with policy-driven eligibility conditions for local breakout. Third, we provide quantitative simulation evidence showing stable ≈ 10 ms latency reduction and sub-2 ms group

delivery, directly linking these gains to the removal of the Backhaul/UPF/AF segment. This work makes the following contributions:

- *RAN–local multicast data path.* We design a gNB-local multicast breakout architecture that pivots eligible uplink flows (s, f) to a downlink point-to-multipoint (PTM) bearer B_g within the gNB that eliminates the core traversal for intra-cell group delivery while retaining core-anchored control.
- *Standards alignment, security, and compliance.* We show how the work fits the 3GPP MBS/5GS architecture [3], [4] without redefining the core roles of authentication, ciphering, and integrity [16].
- *Latency and scalability analysis.* We derive a latency decomposition and demonstrate that local intra-cellular breakout reduces latency ≈ 10 ms. We validate the work using simulation that the average group latency remains 2 ms within the group with a stable ≈ 10 ms gap between local breakout and core-anchored delivery.

IV. SYSTEM MODEL AND PROBLEM STATEMENT

A. Network and Group Model

We consider a private 5G deployment that covers an industrial cell served by a single gNB. Let $\mathcal{U} = \{1, \ldots, N\}$ denote the set of user equipments (UEs) that serve the application such as robots, sensors, controllers, AGVs. attached to the cell. Devices subscribe to one or more *industrial multicast groups* $\mathcal{G} \subseteq 2^{\mathcal{U}}$, where each $g \in \mathcal{G}$ is a subset of $\mathcal{U}$ authorized by policy. For any $g \in \mathcal{G}$, let $\mathcal{R}(g) \subseteq g$ be the receiver set and let $s \in g$ be a designated source that generates time-critical events or state updates to be distributed to $\mathcal{R}(g)$.

Traffic from s to g is carried out in a flow from uplink identified by (s, f) with a per packet deadline D and a reliability target R (e.g. $D \leq 5$, ms and $R \geq 99.999\%$). We measure end-to-end latency L from MAC ingress at s to MAC delivery at each $r \in \mathcal{R}(g)$; reliability is defined as L;

$$\Pr\left[\, L \leq D \,\right] \; \geq \; R, \qquad (1)$$

B. Local Multicast Forwarding at the gNB

The gNB hosts a *local user-plane function* (local UPF) that maintains a forwarding table

$$\text{FT}: \; (s, f) \; \mapsto \; (g, B_g),$$

where B_g denotes a configured point-to-multipoint (PTM) downlink bearer serving $\mathcal{R}(g)$. Upon receiving an eligible uplink PDU for (s, f), the gNB *pivots* the payload to B_g and schedules a PTM transmission to $\mathcal{R}(g)$ within the next available PDSCH slot. The control-plane functions (admission, authorization, group membership, and policy) remain anchored in the 5GS core. Only the *data* path for intra-cell dissemination

is locally broken out in the gNB. Eligibility for local breakout is governed by policy. When eligibility fails, the gNB is returned to the standard core-anchored path without interruption of service.

C. Latency Decomposition

For intra-cell group delivery, the end-to-end latency under core-anchored multicast can be decomposed as

$$L_{\mathrm{CA}} = T_{\mathrm{rqt}} + T_{\mathrm{UL}} + T_{\mathrm{gNB\_proc}}$$
$$+ T_{\mathrm{BH/UPF/AF}} + T_{\mathrm{DL\_schd}} + T_{\mathrm{DL}} \quad (2)$$

where $T_{\mathrm{BH/UPF/AF}}$ aggregates backhaul, UPF, application processing, and associated queueing. Under local forwarding,

$$L_{\mathrm{LB}} = T_{\mathrm{rqt}} + T_{\mathrm{UL}} + T_{\mathrm{gNB\_proc}} + T_{\mathrm{DL\_schd}} + T_{\mathrm{DL}} \quad (3)$$

Payloads do not traverse the core user-plane. The design thus aims to eliminate $T_{\mathrm{BH/UPF/AF}}$ minimizing the latency.

Assumptions. (i) Single-cell analysis: inter-cell groups revert to core-anchored MBS or use inter-gNB coordination. (ii) Control-plane policies are distributed to the gNB; the local MB-UPF does not weaken authentication, ciphering, or integrity protection. (iii) Mobility events trigger a seamless transition between the local and core paths.

Problem Statement. Given $\mathcal{U}$, $\mathcal{G}$, deadlines D, reliability goals R, and gNB's scheduling resources, design a gNB-local multicast forwarding policy π that (a) pivots eligible uplink flows (s, f) to the downlink PTM bearer B_g and (b) schedules transmissions to minimize latency while meeting reliability, and security. The hypothesis is that $L_{\mathrm{local}} \ll L_{\mathrm{core}}$ for intra-cell groups that yields order-of-magnitude improvements in the presence of nontrivial $T_{\mathrm{BH/UPF/AF}}$.

V. PROPOSED GNB LOCAL MULTICAST FORWARDING

A. High-Level Architecture

The control-plane flow is considered as defined by 3GPP. The MB-SMF, 5G core function, is used for session admission and group or policy dissemination and AMF/PCF cooperation. Only relocating the *intra-cell* user-plane replication from the core MB-UPF to the gNB [3], [4], [5]. Concretely, the gNB hosts a MB-UPF that maintains a local forwarding map for multicast traffic.

$$\mathrm{FT}: \ (s, f) \ \mapsto \ (g, B_g)$$

where (s, f) identifies an authorized uplink traffic flow from source s, g to the multicast group, and B_g is a configured point-to-multipoint (PTM) downlink bearer serving $\mathcal{R}(g)$ on NR as per 3GPP TS 38.300 [17]. For an eligible uplink PDU, the local MB-UPF on gNB pivots

the payload to B_g at PDCP/RLC and schedules PTM PDSCH within the next feasible slot, thus avoiding core traversal of the *data* path while preserving core anchored control and policy.

Group membership and *local-breakout* permissions are distributed to the gNB through MB-SMF/AMF under PCF policy [3], [4]. The local MB-UPF considers a flow pivoting (s, f) eligible if: (i) all intended receivers are currently attached to the serving cell (ii) PRB reservations admit PTM transmission. If any condition fails, the gNB transparently reverts to core-anchored MBS without service interruption.

B. Algorithm description and fallback

The gNB maintains a forwarding table keyed by UE and traffic identifiers to enable fast local multicast. When a packet arrives, it checks if a forwarding entry exists and if local breakout is permitted. If either check fails, the packet is passed to the core and served via the standard MB-UPF path. Otherwise, the associated multicast group and bearer are retrieved, and the payload is queued in the appropriate PDCP. Select PTM on the next available PDSCH transmission slot. This approach delivers ultra-low-latency multicast within a cell while reverting to the core when UEs move across cells or policies change.

Alogrithm 1 provides a high-level view of this local multicast forwarding process at the gNB. It shows how the forwarding table (FT) and policy checks determine whether an uplink PDU is sent directly to the core or pivoted into the appropriate downlink multicast bearer for PTM transmission.

Algorithm 1 gNB Local Multicast Breakout

Require: UL PDU (UE_id, $flow\_id$, $payload$)
  State: FT (forwarding table), Policies Π, Group G, Multicast Bearer B_G
1: **if** (UE_id, $flow\_id$) $\notin FT$ **then**
2:     **SendToCore**(); **return**
3: **if** $\neg \Pi$.LOCALBREAKOUTALLOWED(UE_id, $flow\_id$) **then**
4:     **SendToCore**(); **return**
5: $G \leftarrow FT[UE\_id, flow\_id].group$
6: $B_G \leftarrow FT[UE\_id, flow\_id].bearer$
7: **EnqueueToPDCP**($payload$, B_G, QoS_marking)
8: **SchedulePTM_PDSCH**(NextEligibleSlotAlgnToWnd)
9: **CollectLimitedNAKs**()
10: **if AnyNAKs**() **then**
11:     **SelectiveUnicastRepair**()

VI. METHOD AND DISCUSSION

We model a single private 5G cell with one gNB and up to $N=150$ UEs uniformly distributed within a radius 100 m. The gNB is at $(0, 0, 30)$ with 4 antennas and each UE has 2 antennas. NR uses 30 kHz SCS (*slot* = 0.5, ms), 100, RB over 100 MHz in 3.5 GHz, up to 2 MIMO

layers and 64QAM in UL/DL. Each UE generates on/off traffic using `networkTrafficOnOff` with OnTime= 10 ms, OffTime= 90 ms, DataRate= 1 Mbps, and packet size ≈ 1002, bits. Inter-arrival times and payloads are sampled per slot; packets are bit-padded to satisfy NR layer mapping constraints. A CDL-D (`nrCDLChannel`) model is used with carrier-consistent sampling, maximum Doppler 10 Hz, and per-link seeds for repeatability. Decoders are invoked with zero noise variance to isolate path latency from PHY (physical layer) errors. The uplink uses unicast PUSCH (Physical Uplink Shared Channel) from each UE to the gNB whereas downlink uses PTM PDSCH multicast to all receivers. PRB sets span the entire bandwidth for clarity of comparison. The MATLAB simulation implementation is available on: https://github.com/rajendra1124/LocalBreakout.

Scenarios: *Core-anchored MBS:* gNB→MB-UPF/AF→gNB adds a fixed user-plane delay of 5-12 ms per multicast packet [18]. The variation in delay is based on the network design and configuration. *Local breakout:* gNB pivots eligible UL payloads to the DL PTM bearer locally, and radio parameters are identical across both paths.

A. Latency Analysis

We decompose end-to-end latency L for intra-cell group delivery:

$$L = T_{\text{UL\_tx}} + T_{\text{gNB\_proc}} + T_{\text{path}} + T_{\text{DL\_schd}} + T_{\text{DL\_tx}} \quad (4)$$

From Table I and according to (4), the end-to-end latency decomposes into four parts for intra-cell group delivery under 30 kHz SCS. Figure 2 contrasts the two delivery paths and highlights that removing the Backhaul / UPF / AF segment with gNB-local forwarding yields a stable ≈ 10 ms reduction in end-to-end latency across group sizes, which matches the decomposition in (4). UL grant + tx (0.25–1.0 ms) covers access and PHY transmission with request-based UL and queuing. A gNB processing (1.0–2.0 ms) accounts for PDCP/RLC handling, group mapping, and local breakout and PTM scheduling. The Backhaul + UPF + AF segment 5.0–10.0 ms of transport and core processing exists only for core-anchored MBS and disappears with local breakout because payloads never leave the gNB user-plane. DL scheduling + tx (0.25–1.0 ms) reflects the wait to the next PTM opportunity and PDSCH time which get shortened by mini-slot alignment and reserved PRBs. Summing these yields 6.5 to 11.5 ms for core-anchored MBS versus 1.5 to 4.0 ms for local breakout. The ranges are primarily driven by numerology, scheduler protection, and any selective repair or duplication mechanism.

Figure 3 shows that the local breakout architecture consistently achieves a group latency of sub 2 ms

TABLE I
INDICATIVE LATENCY (INTRA-CELL GROUP DELIVERY).

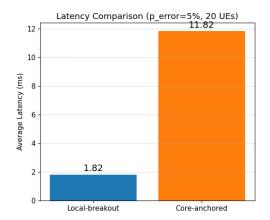| Component	Core-anchored MBS	Local breakout
UL grant + tx	0.25–1.0 ms	0.25–1.0 ms
gNB processing	1.0–2.0 ms	1.0–2.0 ms
Backhaul + UPF + AF	5.0–10.0 ms	0
DL scheduling + tx	0.25–1.0 ms	0.25–1.0 ms
Total	**6.5–11.5 ms**	**1.5–4 ms**

Fig. 2. Average latency of Core-anchored Vs gNB local breakout forwarding for intra-cell groups.

between 10 to 150 receivers, while the core-anchored architecture remains around 12 ms on average well above a URLLC deadline of 5 ms. The nearly flat trend with group size reflects the efficiency of PTM (one DL transmission for all receivers). The constant gap between the curves of 10 ms is the removed core user-plane segment that the local breakout method eliminates. Overall, the figure substantiates two claims: (i) intra-cell local breakout meets tight industrial deadlines with comfortable margin, and (ii) multicast scales with group size without degrading average latency.

For private 5G in Industry 4.0/5.0, the local-breakout method provides a practical path to deterministic group dissemination without architectural upheaval. The pivoting of data from the user-plane moves to the gNB, while the control-plane remains as per the 3GPP MBS architecture [3].

Our design preserves 3GPP security and compliance guarantees while adding the local UPF function collocated at the gNB to connect industries 4.0/5.0 scenarios into next-generation networks. The gNB's local breakout changes the location of replication from the core to the gNB, while leaving the security anchors unchanged. The authorization, group membership, and policy stay anchored in the core. So authentication, ciphering, and integrity protection remain exactly as specified in 3GPP TS 33.501/23.501 [16], [4].
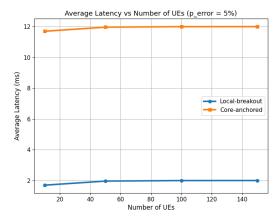
Fig. 3. Latency comparison of Core-anchored Vs local breakout when number of UEs changes.

We acknowledge limits and threats to validity, as the results are from a single-cell setting with controlled traffic and simplified radio/error models. The multi-cell operation, inter-cell groups, handover transients, and correlated interference could change the latency margin. Hardware contention or misconfiguration on the gNB could affect isolation. Future work should therefore evaluate multi-cell deployments with realistic mobility and channel dynamics.

VII. CONCLUSIONS AND FUTURE WORKS

We presented a gNB-local breakout architecture for industrial URLLC in private 5G that pivots eligible up-link traffic at the gNB to a downlink PTM bearer, while leaving authorization, membership, and policy anchored in the core. This local forwarding, commonly referred to al 'local breakout', collapses the intra-cell data path by removing the Backhaul/UPF/AF segment. Hence the end-to-end group latency reduced to roughly sub 2 ms averages in our settings. There is also a stable ≈ 10 ms gap to the core-anchored delivery across loads and group sizes. Crucially, the method preserves 3GPP security and compliance: authentication, ciphering, and integrity. This method is practical for coordinated robotics, AGVs, machine vision, and human–robot collaboration in Industry 4.0 / 5.0 that require low single-digit latency. Although our results are from a single-cell model with controlled traffic, results indicates RAN-local data forwarding is a simple, effective for URLLC grade group dissemination.

Our ongoing work focuses on explicit HARQ and Negative Acknowledgment feedback and selective uni-cast repair. We are taking it to the next level by extending to multi-cell setups, ensuring smooth handover and gNB coordination with a fallback to default core routing for reliability in busy Industry 5.0 settings. To nail down that five-nines reliability, we're going to use probabilistic models and smart tools to predict and dodge packet failures, even in tough network conditions, ensuring steady performance for small data packets. Also, prototyping the scheduler on a real gNB/O-RAN platform to validate latency, observability, and compli-ance under realistic interference and load.

REFERENCES

[1] *3GPP TS 22.104: Service Requirements for Cyber-Physical Control Applications in Vertical Domains*, 3GPP Std., 2024, release 17+.

[2] Ericsson, "Industry 4.0: Manufacturing in the smart way," 2025, accessed Aug. 11, 2025.

[3] *3GPP TS 23.247: Architecture enhancements for 5G System (5GS) to support multicast and broadcast services (MBS)*, 3rd Generation Partnership Project (3GPP) Std., 2024, release 17 and later.

[4] *3GPP TS 23.501: System Architecture for the 5G System (5GS); Stage 2*, 3GPP Std., 2024, release 17+.

[5] *3GPP TS 23.502: Procedures for the 5G System (5GS); Stage 2*, 3GPP Std., 2024, release 17+.

[6] R. Paudyal and S. Shakya, "An approach towards backbone net-work congestion minimization in software defined network," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 412–416.

[7] N. Chukhno, O. Chukhno, D. Moltchanov, S. Pizzi, A. Gaydamaka, A. Samuylov, A. Molinaro, Y. Koucheryavy, A. Iera, and G. Araniti, "Models, methods, and solutions for multicasting in 5g/6g mmwave and sub-thz systems," *Commun. Surveys Tuts.*, vol. 26, no. 1, p. 119–159, Jan. 2024. [Online]. Available: https://doi.org/10.1109/COMST.2023.3319354

[8] E. Garro, M. Fuentes, J. Carcel, H. Chen, D. Mi, F. Tesema, J. Gimenez, and D. Gomez-Barquero, "5g mixed mode: Nr multicast-broadcast services," *IEEE Transactions on Broadcast-ing*, vol. PP, 03 2020.

[9] V. K. Shrivastava, S. Baek, and V. Gupta, "5g mbs – unleashing the potential of multicast and broadcast communication in 5g," *Samsung Research Blog*, 2025, blog post providing an overview of 5G Multicast-Broadcast Services (MBS) in Release 17.

[10] O. Landrove, R. Cabrera, E. Iradier, E. Jimenez, P. Angueira, and J. Montalban, "Broadcast/multicast delivery integration in b5g/6g environments," *Journal of Network and Computer Applications*, vol. 230, p. 103934, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804524001115

[11] B. Coll-Perales, M. C. Lucas-Estañ, T. Shimizu, J. Gozalvez, T. Higuchi, S. Avedisov, O. Altintas, and M. Sepulcre, "End-to-end v2x latency modeling and analysis in 5g networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 4, pp. 5094–5109, 2023.

[12] F. Völk, R. T. Schwarz, M. Lorenz, and A. Knopp, "Emergency 5g communication on-the-move: Concept and field trial of a mobile satellite backhaul for public protection and disaster relief," *International Journal of Satellite Communications and Networking*, vol. 39, no. 4, pp. 417–430, 2021.

[13] J. Zhang, W. Xie, F. Yang, and Q. Bi, "Mobile edge computing and field trial results for 5g low latency scenario," *China Communications*, vol. 13, no. 2, pp. 174–182, 2016.

[14] 3GPP, "TR 23.757: Study on architectural enhancements for 5G multicast-broadcast services (Release 18)," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2021, v0.4.0.

[15] M. Säily, C. B. Estevan, J. J. Gimenez, F. Tesema, W. Guo, D. Gomez-Barquero, and D. Mi, "5g radio access network ar-chitecture for terrestrial broadcast services," *IEEE Transactions on broadcasting*, vol. 66, no. 2, pp. 404–415, 2020.

[16] *3GPP TS 33.501: Security Architecture and Procedures for 5G System*, 3GPP Std., 2024, release 17+.

[17] *3GPP TS 38.300: NR; NR and NG-RAN Overall Description; Stage 2*, 3GPP Std., 2024, release 17+.

[18] 5.7 qos model. TechSpec. Extracted from 3GPP TS 23.501 (Release 18).