Journal of Institute of Science and Technology, 31(1), 137-148(2026)
ISSN: 2467-9062 (print), e-ISSN: 2467-9240
https://doi.org/10.3126/jist.v31i1.92519

Research Article

DYNAMIC RESOURCE MANAGEMENT FOR 5G NETWORK SLICING USING O-RAN NEAR-RT RIC

Bini Chand[1], Aashraya Neupane[1], Binu Suwal[1], Nanda Bikram Adhikari[1*]
[1]*Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, Tribhuvan University, Lalitpur, Nepal*
**Correspondence: adhikari@ioe.edu.np*

(***Received:*** *April 3, 2026;* ***Revised:*** *May 25, 2026;* ***Accepted:*** *June 4, 2026)*

ABSTRACT

Network slicing in 5G New Radio (NR) requires the simultaneous satisfaction of heterogeneous Service Level Agreements (SLAs) for ultra-reliable low-latency communications (uRLLC), enhanced mobile broadband (eMBB), and massive machine-type communications (mMTC). The objective of this study is to design and evaluate a dynamic, three-layer radio-resource-management framework, built on the O-RAN Near-Real-Time Radio Intelligent Controller (Near-RT RIC) that meets these three conflicting SLAs concurrently on a shared gNB. At the data plane, three algorithms are proposed: a Weighted Proportional-Fair (WPF) scheduler that maps RIC-issued weights to proportional-fair time windows; a slice-aware pre-processor that estimates per-slice Physical Resource Block (PRB) demand; and a starvation-aware Bandwidth Part (BWP) multiplexer. At the RIC, a heuristic BWP Manager observes per-slice key performance indicators every 100 ms and updates the slice weights through an exponential-moving-average-smoothed proportional update law. The framework is implemented in ns-3 v3.40 with the 5G-LENA NR module, co-simulated with a Python RIC through ns3-gym, and compared against a static-weight TDMA-PF baseline. Preliminary results from a 10-second co-simulation, in which a single gNB serves 5 uRLLC, 10 eMBB and 30 mMTC user equipments across three independent BWPs, show that the dynamic framework attains 100% uRLLC deadline compliance (mean delay 0.49 ms), 100% eMBB throughput success (34.9 Mbps per UE) and zero mMTC packet loss, whereas the static baseline fails all three SLAs. The idealized modelling assumptions and their implications for real deployments are discussed as limitations.

Keywords: BWP management, Weighted proportional-fair scheduling, ns3-gym, uRLLC, eMBB, mMTC, SLA compliance

INTRODUCTION

The fifth-generation New Radio (5G NR) standard virtualizes shared physical infrastructure into logically independent network slices, each optimized for a distinct service category (NGMN Alliance, 2016; Foukas et al., 2017). Three canonical slice types are recognized: uRLLC, targeting end-to-end latencies below 1 ms and near-zero packet loss; eMBB, requiring sustained throughput in the tens-of-Mbps range; and mMTC, supporting massive populations of Internet-of-Things devices with sporadic, low-rate traffic. Delivering all three concurrently is fundamentally challenging, because uRLLC demands short scheduling windows and negligible queuing delay, eMBB demands large bandwidth allocations, and mMTC demands wide coverage with minimal control overhead.

5G NR introduces Bandwidth Parts (BWPs) as a native spectral-partitioning mechanism (3GPP, 2022a). Each BWP carries an independent numerology, defining a sub-carrier spacing of $15 \times 2^{\wedge}\mu$ kHz and a slot duration of $2^{\wedge}-\mu$ ms. Numerology 3 (120 kHz sub-carrier spacing, 0.125 ms slots) is well suited to uRLLC because a scheduling grant can be issued and acknowledged within a single slot. The O-RAN Alliance architecture (O-RAN Alliance, 2023) positions a Near-RT RIC above the gNB for policy-based control on 10–1000 ms timescales, which is appropriate for slice-weight management.

The objective of this work is to design and validate a dynamic three-layer slicing framework that satisfies the uRLLC, eMBB and mMTC SLAs simultaneously, while addressing several gaps in existing approaches. The most closely related

framework, NRflex (Boutiba et al., 2022), addresses BWP-based slicing for eMBB and uRLLC but omits mMTC, uses absolute PRB counts as its control output, does not modify the aggressiveness of the underlying proportional-fair (PF) scheduler, and was evaluated in MATLAB without a full physical/MAC/RLC stack. This work addresses each of these limitations within a complete ns-3 / 5G-LENA simulation. The key contributions are:

WPF Scheduler: a weight to PF time window mapping that integrates with the 5G-LENA OFDMA-PF infrastructure without modifying the scheduler core.

Slice Pre-processor: per TTI Earliest Deadline First PRB sizing for uRLLC, CQI/throughput-aware sizing for eMBB, and robust-MCS bulk sizing for mMTC, embedded at the gNB MAC layer.

Starvation-aware BWP Multiplexer: per-UE dual counters with configurable force-activation thresholds that prevent indefinite starvation.

Heuristic BWP Manager (Near-RT RIC): a violation-proportional weight update with EMA smoothing and a PRB-saturation guard distinguishing radio-caused from core-caused uRLLC delay violations.

mMTC as a first-class slice: an independent BWP at 700 MHz, 10 MHz, numerology 0, with a dedicated SLA metric and pre-processor.

Full ns-3 co-simulation: end-to-end validation with realistic propagation, UE mobility and mixed UDP traffic, compared against a static baseline.

Related work

Radio Access Network (RAN) slicing has been studied extensively since the early work on wireless-resource virtualization. NVS introduced a substrate for virtualizing cellular resources across providers (Kokku et al., 2012 ; Ksentini & Nikaein, 2017) examined the flexibility and resource abstraction required to enforce slicing on the RAN, while FlexRAN provided a flexible and programmable software-defined RAN platform enabling fine-grained control of base-station functions (Foukas et al., 2016). To bound latency for delay-sensitive slices, (Ksentini et al., 2018) proposed a two-level MAC-scheduling framework that applies different per-slice policies over a shared RAN. The broader set of enablers and challenges for serving vertical industries through 5G RAN slicing was surveyed by (Elayoubi et al., 2019).

More recently, machine learning has been applied widely to slice resource management. (Azimi et al., 2022) provide a comprehensive survey of machine-learning techniques for RAN slicing in 5G and beyond. Deep reinforcement learning (DRL) has been used for resource allocation in dynamic multi-tenant networks (Xie et al., 2022), for closed-loop control through O-RAN xApps (Polese et al., 2023), and for slice resource allocation in the O-RAN midhaul (Cheng et al., 2022). DRL-based controllers can achieve strong asymptotic performance, but they require extensive training and can exhibit instability early in learning. In contrast, heuristic controllers converge immediately and provide interpretable logic (Bonati et al., 2020), which makes them well suited to the deterministic, short-duration evaluation considered here; the present framework therefore adopts a heuristic RIC controller while remaining compatible with a future learning-based replacement.

This work extends NRflex (Boutiba et al., 2022) along several axes: it adds mMTC as a third slice; it replaces the absolute-PRB control output with a normalized-weight mechanism; it introduces PF time-window scaling as the control actuator; it adds a second per-UE starvation counter; it implements a PRB-saturation guard and EMA smoothing; and it validates the design in ns-3 / 5G-LENA rather than MATLAB. The treatment of multiple numerologies follows the multi-numerology radio-resource-management direction explored by (Boutiba et al., 2022). The simulation environment relies on the 5G-LENA NR module (Patriciello et al., 2019) and the ns3-gym OpenAI Gym bridge (Gawłowicz & Zubow, 2019), with control issued over the O-RAN E2 interface defined by the (O-RAN Alliance, 2023).

MATERIALS AND METHODS

System architecture

The framework comprises three layers, illustrated in Figure 1. The Service Layer provides static per-slice SLA policies (Table 1) consumed by the Near-RT RIC. The Near-RT RIC Layer runs the Python BWP Manager algorithm, communicating with ns-3 at 100 ms intervals through the ns3-gym OpenGym interface. The gNB Data-Plane Layer hosts the WPF scheduler, the slice pre-processor and the BWP multiplexer, all implemented as extensions to the NrMacSchedulerOfdmaPF class in 5G-LENA.

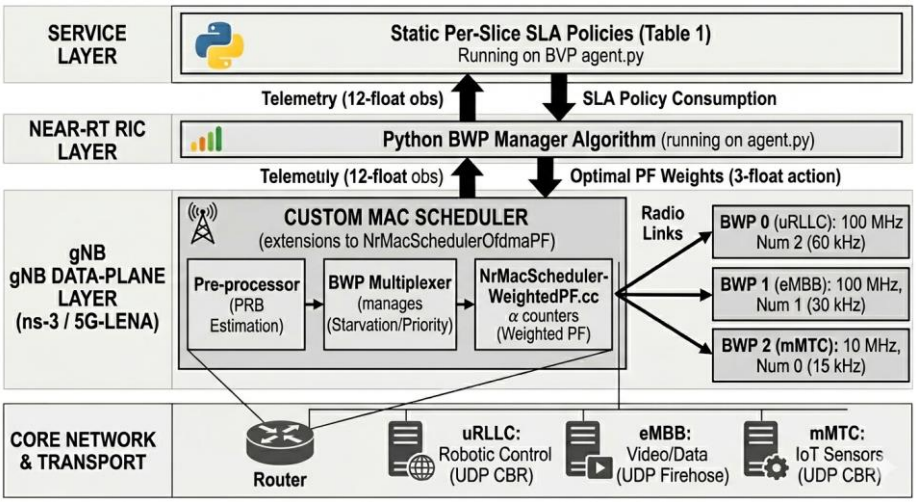

Figure 1. Three-layer architecture, The Service Layer supplies SLA policies; the Near-RT RIC runs the BWP Manager and exchanges a 12-element observation and a 3-element weight action with the data plane every 100 ms; the gNB data plane hosts the pre-processor, BWP multiplexer and weighted-PF scheduler driving three BWPs

As shown in Figure 1, telemetry flows upward from the gNB data plane to the RIC as a 12-element observation vector every 100 ms, and the RIC returns a three-element vector of normalized PF weights that the data-plane scheduler applies to the three BWPs. This closed loop is the principal control mechanism of the framework: the data-plane algorithms translate the weights into per-slice scheduling behavior, while the RIC adjusts the weights in response to observed SLA gaps.

Bandwidth part configuration

Three BWPs are configured on a single gNB. BWP 0 (uRLLC) operates at 6.0 GHz, 100 MHz, numerology 3 (120 kHz sub-carrier spacing, 0.125 ms slots), with Urban Macro (UMa) propagation. BWP 1 (eMBB) operates at 3.5 GHz, 100 MHz, numerology 1 (30 kHz sub-carrier spacing), also with UMa propagation. BWP 2 (mMTC) operates at 700 MHz, 10 MHz, numerology 0 (15 kHz sub-carrier spacing), with Rural Macro (RMa) propagation. Traffic is routed to the BWPs via 3GPP EPS bearer types and EpcTft port-filter rules, mapping uRLLC, eMBB and mMTC flows to BWP 0, BWP 1 and BWP 2 respectively.

Table 1. Per-slice SLA definitions

Parameter	uRLLC	eMBB	mMTC
Minimum throughput (Mbps)	1.0	20.0	0.1
Maximum delay (ms)	1.0	50.0	200.0
Maximum PRB utilization	0.85	0.90	0.80
Target throughput (Mbps)	2.0	20.0	0.5
SLA weight (w_SLA)	5.0	2.0	1.0

Methods and implementation of radio access network model

The simulation models a single gNB serving 45 UEs across the three independent BWPs. The gNB is equipped with a 4 × 8 planar antenna array (32 elements) and each UE with a 2 × 4 array (8 elements), implementing multi-user MIMO with isotropic elements. Ideal direct-path beamforming is applied through the IdealBeamformingHelper, which assumes perfect channel-state information (CSI) at the transmitter. The gNB transmit power is 43 dBm on all three BWPs.

Propagation and channel model

Two propagation environments are modelled following 3GPP TR 38.901 (3GPP, 2020). Urban Macro (UMa) is applied to BWP 0 (6.0 GHz) and BWP 1 (3.5 GHz), modelling dense urban deployment with a gNB height of 10 m and UE height of 1.5 m. Rural Macro (RMa) is applied to BWP 2 (700 MHz), modelling wide-area IoT coverage with longer propagation distances, which motivates the choice of numerology 0 (longest cyclic prefix). Shadowing is disabled in both models to isolate the effect of the scheduling algorithms from random shadowing variation across runs.

UE mobility and traffic models

Three mobility models are used, matched to each slice's typical use case. The 5 uRLLC UEs follow Random Walk 2D mobility within a 400 × 400 m area at 1–5 m/s, representing industrial robots or automated guided vehicles. The 10 eMBB UEs follow Random Waypoint mobility confined to a 100 × 100 m region centered on the gNB at 1–3 m/s, representing indoor video streaming with good CQI. The 30 mMTC UEs follow Random Walk 2D mobility over the full area, representing distributed low-mobility sensors. Random Walk and Random Waypoint are standard ns-3 mobility models widely used in RAN-slicing evaluations (Azimi et al., 2022); they are selected here because they reflect the mobility characteristics of the corresponding slice use cases rather than for analytical convenience. Likewise, the UDP constant-bit-rate (uRLLC, mMTC) and UDP on-off (eMBB) traffic generators in Table 2 follow common practice for modelling deterministic control traffic, burst IoT reporting and elastic broadband demand, respectively.

Table 2. Traffic parameters per slice for two scenarios

Slice	Type / size	Rate / interval
Dynamic SDN scenario		
uRLLC	UDP CBR, 512 B	1 ms (~4.1 Mbps offered)
eMBB	UDP OnOff, 1500 B	50 Mbps, always-on
mMTC	UDP CBR, 1280 B	10 ms (~1.0 Mbps offered)
Static baseline scenario		
uRLLC	UDP CBR, 100 B	2 ms (0.4 Mbps offered)
eMBB	UDP CBR, 1400 B	1 ms (11.2 Mbps offered)
mMTC	UDP CBR, 50 B	20 ms (0.02 Mbps offered)

Traffic flows downlink from a remote MEC host to each UE. Applications start at t = 1.0 s to allow radio-link setup and RRC connection establishment to complete before transmission begins.

Idealized modelling assumptions

To isolate the behavior of the proposed scheduler and RIC controller from confounding transport- and channel-related effects, several components are deliberately idealized in this study. The S1-U interface delay and the MEC link delay are set to zero (co-located edge); perfect CSI and ideal direct-path beamforming are assumed; the PHY processing delays and transport-block decoding latency are set to zero at the uRLLC BWP; and the SRS overhead is reduced through a long periodicity. These choices, and their implications for a fielded deployment, are examined in the limitations subsection.

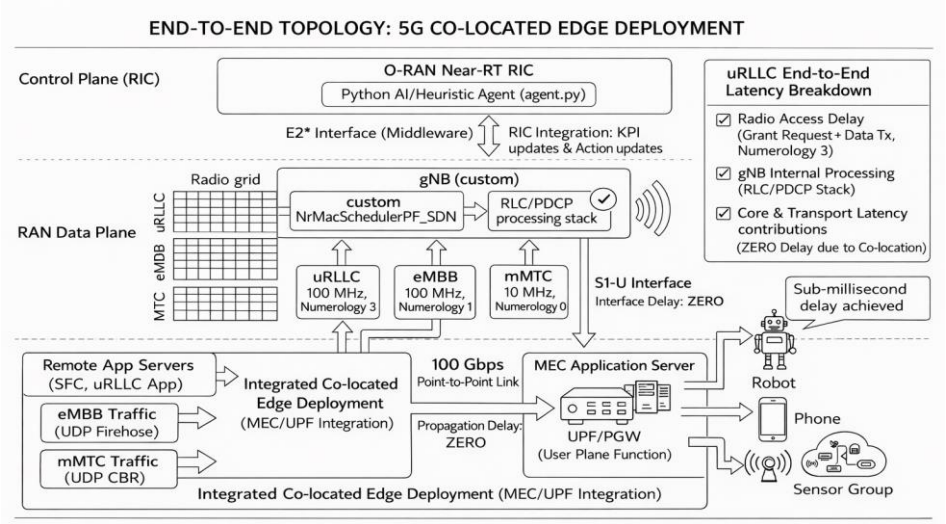

Figure 2. End-to-end network topology, The MEC server is co-located with the PGW/UPF over a 100 Gbps, zero-delay link, eliminating the core-network propagation floor so that the only latency contributions for uRLLC are the radio-access delay at numerology 3 and the RLC/PDCP processing stack

Figure 2 shows the end-to-end topology. The EPC/5GC consists of a PGW/UPF node connected to the MEC application server through a 100 Gbps point-to-point link with zero propagation delay, modelling a co-located edge deployment; the S1-U interface delay is also set to zero. This co-location is the key enabler for the sub-millisecond uRLLC measurement: with the transport floor removed, the only latency contributions are the radio-access delay (scheduling grant plus data transmission at numerology 3) and the RLC/PDCP processing stack.

Queue management and SRS

Active Queue Management is applied on all network devices using FQ-CoDel with a target delay of 2 ms and an interval of 20 ms, preventing UDP buffer-bloat on the MEC-to-PGW link. The RLC layer uses Unacknowledged Mode for all bearers with a 20 KB transmit buffer per UE, small enough to produce meaningful drop counts during congestion for the mMTC buffer-drop KPI. The SRS periodicity is set to 320 slots, reducing SRS overhead for the 45-UE deployment while maintaining sufficient CQI feedback for the adaptive modulation-and-coding algorithm.

Proposed algorithms

Weighted PF scheduler

The standard PF metric is the ratio of instantaneous achievable rate to the exponentially filtered average rate over a time window T_w. The WPF algorithm maps the RIC weight w ∈ (0, 1] to T_w:

$$T\_w(w) = max(1, T\_base / w), \quad (1)$$

with T_base = 99 ms. A shorter T_w increases scheduling reactivity towards high-CQI UEs. RIC weights are scaled by κ = 3 so that the equal-split weight w = 1/3 maps to a scheduling weight of 1 and preserves T_w = 99 ms.

Slice pre-processor

For uRLLC, all queued bytes are treated as urgent (within two TTIs), consistent with the 1 ms SLA at numerology 3; the required PRBs follow from the queued bytes and the per-PRB capacity (CQI-dependent spectral efficiency η_CQI times 168 resource elements per PRB):

$$P\_uRLLC = [\, 8 \cdot Q\_bytes \,/\, (\eta\_CQI \cdot 168) \,] \quad (2)$$

For eMBB, the demand is the minimum of the PRBs for the target rate (R), the PRBs to drain the queue (N), and the available PRBs:

$$P\_eMBB = min(R, N, P\_avail) \quad (3)$$

For mMTC, robust bulk sizing uses a fixed conservative spectral efficiency of 0.3770 bits per resource element (CQI 3, QPSK):

$$P\_mMTC = min([\, 8 \cdot Q\_agg \,/\, (0.3770 \cdot 168) \,], P\_avail) \quad (4)$$

Starvation-aware BWP ultiplexer

Per-UE counters track the consecutive slots in which eMBB and mMTC were pre-empted. Force-activation thresholds of 8 (eMBB) and 16 (mMTC) guarantee bounded starvation: when a counter reaches its threshold, the corresponding slice is activated for that UE regardless of higher-priority demand, after which the counter resets. uRLLC retains strict priority in all other cases (Nandan & Adhikari, 2021). This dual-counter design extends the single-counter mechanism of NRflex and prevents the indefinite starvation that strict priority alone would cause for the lower-priority slices.

BWP manager (Near-RT RIC)

The RIC observes a 12-element vector composed of per-slice filtered throughput, mean delay, PRB utilization and drop-rate derivative, and derives three SLA-gap KPIs: a uRLLC deadline-failure indicator combining normalized delay excess with the drop-rate derivative (Equation 5), an eMBB throughput-success ratio (Equation 6), and an mMTC buffer-drop indicator (Equation 7).

$$f\_uRLLC = max(0, (d0 - d0\_max)/ d0\_max) + p0 \quad (5)$$

$$f\_eMBB = min(1, T1 / T1*) \quad (6)$$

$$f\_mMTC = p2 \quad (7)$$

Weight updates are violation-proportional. For uRLLC, a radio-caused violation (positive KPI with PRB utilization at or above the saturation threshold) increases the weight in proportion to the violation, while over-performance releases it towards the floor. For eMBB and mMTC, the weights rise with the throughput or PRB-utilization gaps:

$$w0 \leftarrow w0 + \delta(1 + min(2, f\_uRLLC)) \quad (8)$$

$$w1 \leftarrow w1 + \delta\big(3 + 2(1 - f_{eMBB})\big) \quad (9)$$

$$w2 \leftarrow w2 + \delta(1 + f\_mMTC) \quad (10)$$

After clipping to [0.10, 0.70] and renormalizing, an EMA with smoothing factor τ = 0.3 stabilizes the output:

$$w(t+1) = (1 - \tau) \cdot w(t) + \tau \cdot \hat{w} \quad (11)$$

The environment returns a reward each step that rewards SLA compliance across the three slices, weighted by the SLA weights; perfect compliance yields a maximum of about 5.55:

$$r = \Sigma\, w\_SLA, s \cdot ln(1 + max(0, 1 - V\_s)) \quad (12)$$

where V_s is a composite SLA-violation score combining throughput deficit, delay excess, PRB excess and drop rate.

Static baseline scheduler

The static baseline extends NrMacSchedulerTdmaPF. All UEs share a single BWP at 28 GHz, 100 MHz. A slice-weighted PF metric assigns higher priority to uRLLC using fixed weights of 10:3:1 (uRLLC:eMBB:mMTC). No control loop runs; the weights are constant throughout. The baseline therefore employs the same family of data-plane primitives as the dynamic framework but lacks the closed-loop weight adaptation, which isolates the contribution of the RIC controller.

Simulation setup

All simulations use ns-3 v3.40 with the 5G-LENA NR module; the Python RIC connects through ns3-gym on TCP port 5555. Each run lasts 10 s, divided into 99 control steps of 100 ms. Applications start at t = 1.0 s, so the first nine steps constitute a warm-up during which no application traffic flows; these are excluded from the active-phase statistics. KPIs are extracted with the ns-3 FlowMonitor, and steady-state statistics are computed over steps 11–93, after the eMBB and mMTC BWPs reach their utilization plateaux. Table 3 summarizes the parameters for both scenarios.

Table 3. Principal simulation parameters

Parameter	Value
Simulation duration	10 s (99 steps at 100 ms)
gNB transmit power	43 dBm (all BWPs)
Antennas (gNB / UE)	4×8 / 2×4, isotropic
Beamforming	Ideal direct-path
uRLLC / eMBB / mMTC UEs	5 / 10 / 30

RIC step size δ, EMA τ	0.05, 0.3
Weight bounds (w_min, w_max)	0.10, 0.70
Initial weights	(1/3, 1/3, 1/3)
MEC / S1-U delay	0 ms (co-located edge)
Static carrier / BW	28 GHz / 100 MHz, single BWP
Static scheduler	TDMA-PF, fixed weights 10:3:1

RESULTS

SLA compliance reward

Figure 3 shows the per-step SLA compliance reward over the 99-step episode. Step 10 yields a peak reward of 5.545 (99.8% of the theoretical maximum of 5.55), because at that moment the eMBB and mMTC PRB utilizations are still below their SLA ceilings, so no PRB over-utilization penalty is incurred. From step 11 onward, the eMBB BWP saturates and the mMTC PRB utilization stabilizes above its 0.80 ceiling, both triggering PRB-violation penalties; the steady-state reward consequently settles at 3.75 ± 0.05. Despite the PRB excess, the throughput and delay SLAs are fully satisfied for all three slices throughout the episode.

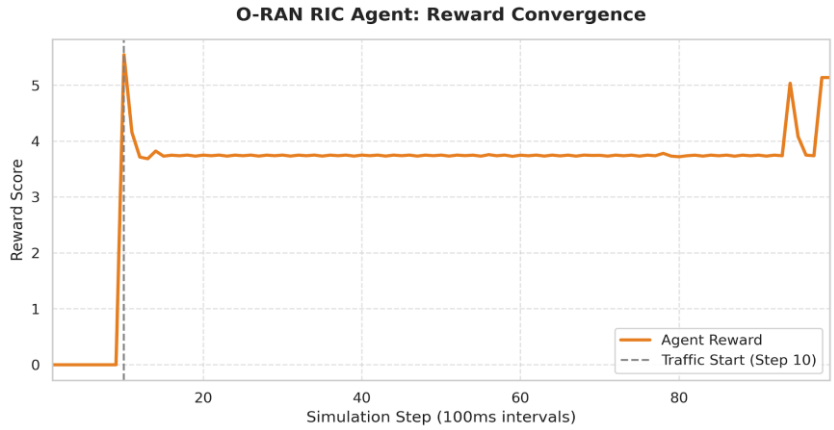

Figure 3. SLA compliance reward over 99 steps. The peak at step 10 coincides with the brief interval before the eMBB and mMTC BWPs reach PRB saturation; the steady-state mean reward is 3.75 ± 0.05

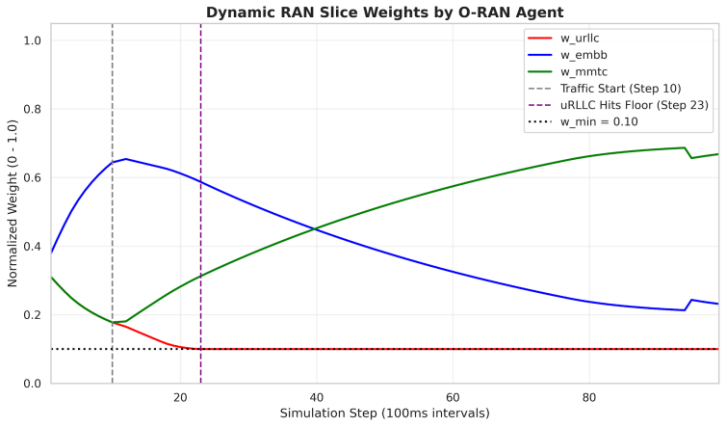

Figure 4. Normalized BWP weight evolution output by the Near-RT RIC. The eMBB weight peaks at step 11, then decreases as the mMTC-boost rule fires continuously; the uRLLC weight reaches the floor of 0.10 by step 23

Weight evolution and RIC adaptation

Figure 4 shows the normalized weight trajectory output by the Near-RT RIC, in which three phases emerge. During the warm-up (steps 1–9), no application traffic flows and the eMBB-boost rule fires repeatedly, driving the eMBB weight from 0.379 to 0.627. At the adaptation onset (steps 10–11), traffic arrives, the eMBB throughput SLA is met, and the eMBB weight peaks at 0.649. In steady state (steps 12–99), because the mMTC PRB utilization remains above 0.80, the mMTC-boost rule fires every step, raising the mMTC weight from 0.179 to 0.686 by step 93; renormalization reduces the eMBB weight symmetrically to 0.214. The uRLLC weight reaches its floor of 0.10 by step 23 because the co-located MEC and zeroed PHY delays keep the uRLLC delay consistently below 0.8 ms, triggering the over-performing release rule.

Per-slice throughput

Figure 5 and Table 4 compare per-slice per-UE throughput. In the dynamic scenario, uRLLC delivers a constant 4.32 Mbps per UE, exceeding the 1.0 Mbps SLA minimum by 4.3 times, whereas the static baseline delivers only 0.51 Mbps per UE because PRB contention on the shared single BWP causes queuing and loss. eMBB delivers 34.93 ± 0.07 Mbps per UE against a 50 Mbps offered load (1.75 times the SLA target); the eMBB BWP is fully saturated, so the 100 MHz BWP, rather than the scheduler, is the bottleneck. The static baseline delivers only 9.87 ± 0.34 Mbps per UE, a 50.6% deficit. mMTC delivers 1.04 ± 0.04 Mbps per UE (10.4 times the SLA minimum) against 0.031 Mbps per UE for the static baseline.

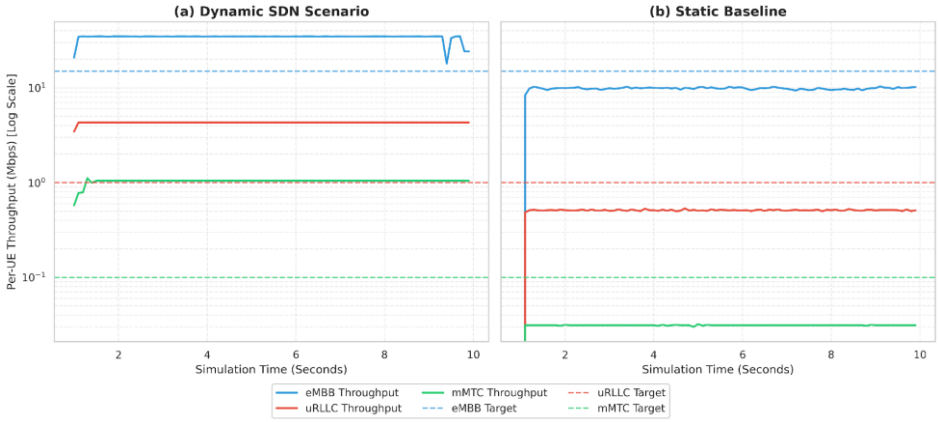

Figure 5. Per-slice per-UE throughput on a logarithmic scale: (a) dynamic SDN scenario and (b) static baseline, Dashed lines indicate SLA minimum targets; the dynamic framework meets all three targets while the static baseline fails all three

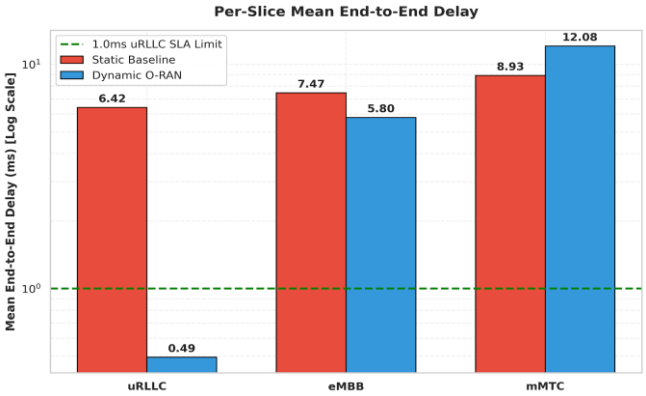

Figure 6. Per-slice mean end-to-end delay on a logarithmic scale, the dynamic uRLLC delay (0.49 ms) is about 13 times below the SLA limit, while the static baseline (6.4 ms) exceeds it in every interval

Latency compliance

Figure 6 shows the per-slice mean end-to-end delay. In the dynamic scenario, uRLLC mean delay is 0.493 ± 0.010 ms, well below the 1 ms limit, with a deadline-failure rate of zero in 100% of active steps. The co-located MEC and zeroed PHY processing delays reduce the irreducible floor to the scheduling wait time at numerology 3, and the residual delay is attributable to radio-layer queuing at roughly 15% PRB utilization. In the static baseline, uRLLC mean delay is 6.418 ms, exceeding the 1 ms SLA in every interval (0% compliance). eMBB mean delay is 5.738 ± 0.012 ms (within the 50 ms SLA) and mMTC mean delay is 12.14 ± 0.53 ms (within the 200 ms SLA).

Buffer management and PRB utilization

Figure 7 shows the three derived KPIs over time. From step 10 onward, the uRLLC deadline-failure indicator is zero, the eMBB throughput-success indicator is one, and the mMTC buffer-drop indicator is zero, confirming simultaneous SLA compliance for throughput and delay across all three slices. The conservative CQI-3 spectral efficiency used in the mMTC pre-processor correctly sizes PRB demand even at the cell edge of the 700 MHz RMa channel, so no mMTC packets are dropped. Figure 8 shows PRB utilization per BWP. BWP 0 (uRLLC) stabilizes at 15.7 ± 1.6%, reflecting the low duty cycle of the 512 B CBR traffic. BWP 1 (eMBB) reaches 100% from step 11, confirming that the 100 MHz BWP capacity is the bottleneck. BWP 2 (mMTC) stabilizes at 93.3 ± 1.6%, exceeding the 0.80 ceiling and driving the continuous mMTC weight growth seen in Figure 4. In the static baseline, PRB utilization is much lower across all slices because the shared 28 GHz BWP at default numerology cannot schedule all UEs efficiently, producing large queuing delays despite moderate PRB usage.

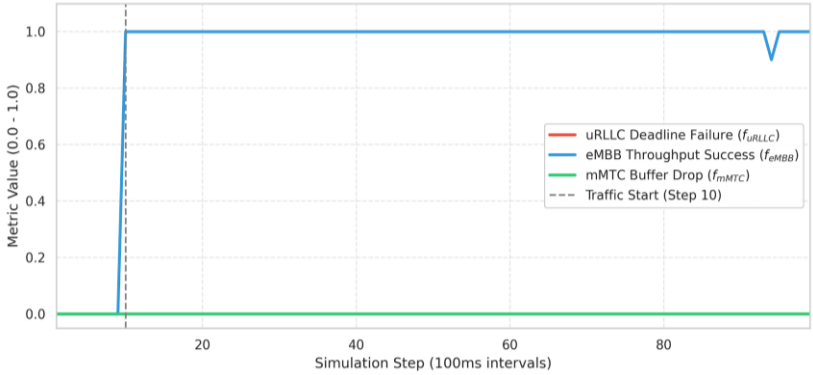

Figure 7. Derived KPI metrics over 99 steps. From step 10 onward the uRLLC deadline-failure and mMTC buffer-drop indicators are zero and the eMBB throughput-success indicator is one, confirming simultaneous SLA compliance for all three slices

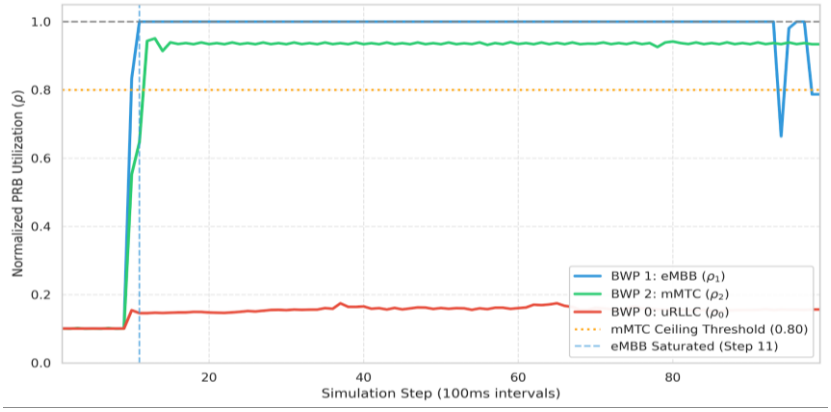

Figure 8. PRB utilization per BWP in the dynamic scenario, BWP 1 (eMBB) is fully saturated from step 11, and BWP 2 (mMTC) exceeds the 0.80 ceiling, driving continuous mMTC weight boosts

Table 4. Steady-state performance comparison (dynamic steps 11–93; static t = 1.1–9.9 s)

Metric	uRLLC	eMBB	mMTC
Dynamic – Throughput (Mbps/UE)	4.32	34.93	1.04
Dynamic – Delay (ms)	0.493	5.738	12.14
Dynamic – PRB utilisation	0.157	1.000	0.933
Dynamic – Throughput SLA (%)	100	100	100
Dynamic – Delay SLA (%)	100	100	100
Static – Throughput (Mbps/UE)	0.511	9.873	0.031
Static – Delay (ms)	6.418	7.466	8.929
Static – Throughput SLA (%)	0	0	0
Static – Delay SLA (%)	0	100	100

Table 5. Feature comparison: this work versus NRflex

Feature	NRflex	This work
Slice types	eMBB + uRLLC	+ mMTC
RIC output	Absolute PRBs	Normalized weights
Scheduler modification	None	PF window scaling
PRB saturation guard	No	Yes
Starvation counters	1 per UE	2 per UE
EMA smoothing	No	Yes ($\tau = 0.3$)
Validation platform	MATLAB	ns-3 / 5G-LENA
Maximum numerology	2	3
MEC delay-floor removal	No	Yes

DISCUSSION

The dynamic framework outperforms the static baseline in every throughput metric: uRLLC by 8.5 times (4.32 versus 0.51 Mbps per UE), eMBB by 3.5 times (34.9 versus 9.87 Mbps per UE), and mMTC by 33.5 times (1.04 versus 0.031 Mbps per UE); the uRLLC latency improvement is 13 times (0.49 versus 6.42 ms), converting a 0% compliance rate to 100%. Table 5 summarizes the architectural differences with NRflex (Boutiba et al., 2022), the most directly comparable framework. Relative to DRL-based controllers such as ColO-RAN (Polese et al., 2023) and the DRL allocators of (Xie et al., 2022) and (Cheng et al., 2022), the proposed heuristic controller trades asymptotic optimality for immediate convergence and full interpretability, and it additionally treats mMTC as a first-class slice and removes the MEC delay floor, neither of which is addressed in NRflex.

Among the four proposed components, the Near-RT RIC BWP Manager is the single most decisive contributor to SLA compliance. The static baseline employs the same data-plane primitives (a slice-weighted PF scheduler) but, lacking the closed-loop weight adaptation, fails all three SLAs. The WPF scheduler and the pre-processor are necessary enablers, but it is the RIC's violation-proportional weight update that converts the per-slice KPIs into the dynamic reallocation responsible for the measured gains.

Limitation

The results reported here are obtained under deliberately idealized conditions chosen to isolate the contribution of the proposed scheduling and control logic, and they should be interpreted accordingly. First, the co-located MEC and the zero S1-U/transport delay remove the core-network propagation floor; in a geographically distributed deployment an additional fixed delay, typically of several milliseconds, would

be added to the end-to-end latency of every slice, reducing the uRLLC margin substantially. Second, ideal direct-path beamforming with perfect CSI at the transmitter was assumed; realistic channel-estimation error and feedback delay would lower the achievable spectral efficiency and could intermittently threaten the eMBB throughput target near saturation. Third, the PHY processing delays and transport-block decoding latency were set to zero at the uRLLC BWP; non-zero processing budgets would raise the irreducible uRLLC delay floor (Nandan & Adhikari, 2021). Fourth, the SRS overhead was reduced through a long (320-slot) periodicity. Consequently, the absolute latency and the rapid convergence observed here are best regarded as best-case bounds that characterize the scheduler and RIC behavior rather than as predictions for a fielded network; the fast convergence in particular follows from the deterministic heuristic update law acting on stationary traffic, and would be slower under non-stationary load. The relative improvement over the static baseline, evaluated under the same idealizations, is expected to be more robust to these assumptions. Quantifying the framework under realistic transport delay, imperfect CSI and full SRS overhead is left to future work.

CONCLUSION

This paper presented a dynamic 5G network-slicing framework integrating an O-RAN Near-RT RIC with three gNB-side algorithms: a weighted-PF time-window scaler, a slice-aware pre-processor and a starvation-aware BWP multiplexer. Implemented in ns-3 v3.40 with 5G-LENA and ns3-gym, and validated against a static-weight TDMA-PF baseline over a 10-second co-simulation, the framework achieved 100% uRLLC deadline compliance (mean delay 0.493 ms versus 6.42 ms for the static baseline), 100% eMBB throughput success (34.9 Mbps per UE versus 9.87 Mbps per UE), and zero mMTC packet drop (1.04 Mbps per UE versus 0.031 Mbps per UE), with a steady-state reward of 3.75 against an estimated 1.07 for the static baseline.

A notable finding is that the mMTC BWP PRB utilization (about 93%) persistently exceeds the 0.80 SLA ceiling, driving a continuous weight transfer from eMBB to mMTC and reducing the reward below its theoretical maximum. This reveals an intrinsic tension between throughput-SLA compliance and PRB-efficiency SLAs under heavy mMTC load on the narrow 10 MHz BWP: the system satisfies all performance SLAs but at the cost of spectral efficiency. As discussed in the Limitations subsection, the absolute figures depend on idealized transport, channel and processing assumptions. Future work will quantify the framework under realistic transport delay, imperfect CSI and full SRS overhead, explore dynamic mMTC BWP bandwidth expansion to resolve the efficiency trade-off, and investigate DRL-based weight optimization and multi-gNB inter-cell coordination.

ACKNOWLEDGMENTS

The authors acknowledge the open-source contributions of the ns-3, 5G-LENA and ns3-gym communities that provided the simulation infrastructure used in this work.

AUTHORS CONTRIBUTION

Conceptualization: BC, AN, BS; Methodology: BC, AN, BS, NBA; Validation: NBA Investigation: BC, AN, BS; Data Analysis: BC, AN, BS; Writing-original draft: BC, AN, BS; Writing - review & editing: NBA

FUNDING

None

ORCIDs

Bini Chand:
https://orcid.org/0009-0005-5718-2516
Aasharya Neupane:
https://orcid.org/0009-0007-1903-0954
Binu Suwal
https://orcid.org/0009-0001-1708-6117
Nanda Bikram Adhikari:
https://orcid.org/0000-0003-1862-3671

CONFLICT OF INTEREST

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ETHICAL STATEMENT

The authors state that it is their original work and has not been previously published or submitted for publication elsewhere.

DATA AVAILABILITY STATEMENT

The complete simulation source code and raw output CSV files will be made available by the corresponding author upon reasonable request.

SUPPLEMENTARY INFORMATION

None

REFERENCES

3GPP. (2020). Study on channel model for frequencies from 0.5 to 100 GHz (Technical

Report TR 38.901). 3rd Generation Partnership Project.

3GPP. (2022a). NR; Physical channels and modulation (Technical Specification TS 38.211, Release 17). 3rd Generation Partnership Project.

3GPP. (2022b). NR; Physical layer procedures for data (Technical Specification TS 38.214, Release 17). 3rd Generation Partnership Project.

3GPP. (2022c). NR; Overall description; Stage-2 (Technical Specification TS 38.300, Release 17). 3rd Generation Partnership Project.

Azimi, Y., Yousefi, S., Kalbkhani, H., & Kunz, T. (2022). Applications of machine learning in resource management for RAN-slicing in 5G and beyond networks: A survey. *IEEE Access*, *10*, 106581–106612.

Bonati, L., Polese, M., D'Oro, S., Basagni, S., & Melodia, T. (2020). Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead. *Computer Networks*, *182*, 107516.

Boutiba, K., Bagaa, M., & Ksentini, A. (2022). Radio resource management in multi-numerology 5G new radio featuring network slicing. In IEEE International Conference on Communications (ICC) (pp. 359–364). IEEE.

Boutiba, K., Ksentini, A., Brik, B., Challal, Y., & Balla, A. (2022). NRflex: Enforcing network slicing in 5G new radio. *Computer Communications, 181*, 284–292.

Cheng, N. F., Pamuklu, T., & Erol-Kantarci, M. (2022). Reinforcement learning based resource allocation for network slices in O-RAN midhaul. arXiv preprint arXiv:2211.07466.

Elayoubi, S. E., Ben Jemaa, S., Altman, Z., & Galindo-Serrano, A. (2019). 5G RAN slicing for verticals: Enablers and challenges. *IEEE Communications Magazine, 57*(1), 28–34.

Foukas, X., Nikaein, N., Kassem, M. M., Marina, M. K., & Kontovasilis, K. (2016). FlexRAN: A flexible and programmable platform for software-defined radio access networks. In Proceedings of the 12th International Conference on Emerging Networking Experiments and Technologies (CoNEXT) (pp. 427–441). ACM.

Foukas, X., Patounas, G., Elmokashfi, A., & Marina, M. K. (2017). Network slicing in 5G: Survey and challenges. *IEEE Communications Magazine*, *55*(5), 94–100.

Gawłowicz, P., & Zubow, A. (2019). ns3-gym: Extending OpenAI Gym for networking research. arXiv preprint arXiv:1910.01523.

Kokku, R., Mahindra, R., Zhang, H., & Rangarajan, S. (2012). NVS: A substrate for virtualizing wireless resources in cellular networks. *IEEE/ACM Transactions on Networking*, *20*(5), 1333–1346.

Ksentini, A., Frangoudis, P. A., Amogh, P. C., & Nikaein, N. (2018). Providing low latency guarantees for slicing-ready 5G systems via two-level MAC scheduling. *IEEE Network, 32*(6), 116–123.

Ksentini, A., & Nikaein, N. (2017). Toward enforcing network slicing on RAN: Flexibility and resources abstraction. *IEEE Communications Magazine*, *55*(6), 102–108.

Nandan R. K. & Adhikari N. B. (2021). A Multi-connectivity Framework and Simulation Analysis of Ultra-Reliable Low Latency Communication (URLLC) in 5G Network. *Journal of The Institution of Engineers (India): Series B, 102* (5), 895-902.

NGMN Alliance. (2016). Description of network slicing concept (NGMN 5G White Paper). Next Generation Mobile Networks Alliance.

O-RAN Alliance. (2023). O-RAN architecture description 07.00 (O-RAN WG1 Technical Report). O-RAN Alliance.

Patriciello, N., Lagen, S., Bojovic, B., & Giupponi, L. (2019). An E2E simulator for 5G NR networks. Simulation Modelling Practice and Theory, *96*, 101933.

Polese, M., Bonati, L., D'Oro, S., Basagni, S., & Melodia, T. (2023). ColO-RAN: Developing machine learning-based xApps for open RAN closed-loop control on programmable experimental platforms. *IEEE Transactions on Mobile Computing, 22*(10), 5787–5800.

Xie, Y., Kong, Y., Huang, L., Wang, S., Xu, S., Wang, X., & Ren, J. (2022). Resource allocation for network slicing in dynamic multi-tenant networks: A deep reinforcement learning approach. *Computer Communications*, *195*, 476–487.

.