Predictive Preemptive Backhaul Borrowing for Public-Safety Traffic Prioritization in 5G IAB Networks

Joshua Omorodion, Jinwei Liu
Department of Computer and Information Sciences
Florida A&M University, Tallahassee, FL 32307, USA
Email: {joshua1.omorodion, jinwei.liu}@famu.edu

***Abstract*—Integrated Access and Backhaul (IAB) enables 5G public-safety (PS) coverage by reusing the NR radio interface for access and backhaul, but the shared-medium half-duplex design creates a multi-hop cascade problem: a mission-critical surge at a leaf node overwhelms upstream hops that static slicing and local-only priority cannot prevent. We propose a predictive preemptive backhaul borrowing framework in which a queue-growth predictor fires a path-scoped Borrowing-Advisory before overflow, triggering a coordinated, temporary downgrade of non-critical commercial flows. Evaluation in a discrete-time system-level simulator (1,800 runs, 72 scenario configurations, five policies including a tabular Q-learning baseline) and an ns-3 5G-LENA proof-of-concept shows: over the full sweep, predictive borrowing and Q-learning achieve near-equal aggregate PS delivery (0.939 vs. 0.930); in the hardest stressed regime the heuristic leads by 12.5 points (0.884 vs. 0.759) because its trigger directly encodes signaling-plus-switching delay. Time-division duplexing (TDD) slot borrowing via `SetPattern()` is implementable without modifying core 5G-LENA internals. Commercial delivery remains non-zero under all borrowing policies. The Borrowing-Advisory maps to an O-RAN near-RT RIC xApp and represents a new scheduler extension not present in current 3GPP specifications.**

***Index Terms*—5G, integrated access and backhaul, network slicing, public safety, O-RAN, reinforcement learning**

I. Introduction

Fifth-generation networks increasingly support public-safety traffic through logical prioritization, dedicated QoS treatment, and network slicing [1]. Those mechanisms work well when the access network is fiber-fed and capacity bottlenecks are largely local. In IAB networks, however, a burst at one access node can consume capacity across several wireless backhaul hops [2]. A thermal video stream from a first-responder drone or an incident robot does not only compete with nearby commercial users on the ingress node; it can trigger a capacity cascade along each parent hop between the serving IAB node and the donor gNB [3].

Operational context. IAB is most likely deployed where this cascade matters: post-disaster temporary coverage, rural or underserved areas without fiber backhaul, and large-event surge zones. In such deployments, 3–4 relay hops are common, commercial and mission-critical traffic share the same spectrum, and the half-duplex TDD constraint means that access and backhaul compete for time-frequency resources within each node. When an incident triggers a burst of MCPTT (Mission-Critical Push-to-Talk) voice, MC video, or sensor telemetry, local-only priority at the ingress node is insufficient: upstream hops remain loaded with commercial traffic that was never preempted, creating a bottleneck that the ingress scheduler cannot resolve alone.

Mixed commercial/public-safety IAB therefore introduces a *multi-hop backhaul coordination gap* not addressed by static slice reservation or local-only priority [4]. When a surge is detected too late, commercial access activity, half-duplex frame constraints, and upstream buffer pressure combine to cause queue overflow before the network can react.

This paper proposes a bounded mechanism for preemptive backhaul borrowing: a controller that predicts an imminent public-safety surge, propagates a path-scoped preemption advisory upstream, and temporarily downgrades non-critical commercial resources on the affected IAB chain via *controlled downgrade*, not session destruction. The mechanism reuses existing QoS, slice, and scheduler semantics and can be realized as an O-RAN near-RT RIC xApp.

Our contributions are: (1) a problem framing isolating multi-hop cascade and half-duplex failure modes in mixed commercial/public-safety IAB; (2) a predictive borrowing mechanism with a low-complexity heuristic ($O(H \cdot F \log F)$) and explicit delay modeling; (3) a 1,800-run evaluation across 72 configurations and five policies including a Q-learning baseline; and (4) an ns-3 5G-LENA proof-of-concept via `SetPattern()`.

The scope of this work is concept validation, identifying a standards-relevant control primitive that warrants deeper PHY-level, multi-hop, and multi-vendor study.

II. Related Work

IAB resource management. Polese *et al.* [5] surveyed IAB at mmWave; Cudak *et al.* [6] positioned IAB as a 5G enabler. Pagin *et al.* [7] addressed semi-centralized IAB resource management; Alghafari and Haghighi [8] proposed decentralized allocation. Yu *et al.* [3] combined centralized donor coordination with distributed IAB-node reaction for bursty traffic, but do not differentiate traffic classes or apply to public-safety surge scenarios. None of these works model mission-critical prioritization under multi-hop surge conditions.

TABLE I
RELATED WORK POSITIONING

Capability	IAB Mgmt	PS QoS Slicing	This Work
Multi-hop resource alloc.	✓	–	✓
PS prioritization	–	✓	✓
Cross-hop surge coord.	–	–	✓
Predictive trigger	–	–	✓
Delay model	–	–	✓

RAN slicing and public-safety QoS. RAN slicing provides differentiated QoS [9], [10], and public-safety work shows logically dedicated treatment on shared infrastructure [1], [11]. However, these assume single-hop or fiber-backhaul access; slice isolation does not extend across multi-hop wireless backhaul.

Learning-based IAB scheduling. DRL has been applied to IAB spectrum allocation [12] with performance gains over static policies, but trained agents lack per-decision interpretability and require retraining when conditions change.

Gap. Table I positions this work. The gap is the intersection of the three lines above: a surge-response mechanism coordinating backhaul borrowing across IAB hops under half-duplex constraints and bounded control delay. 3GPP TR 38.874 [13] defines the IAB architecture and TS 38.340 [14] specifies BAP, but neither defines a cross-hop preemptive borrowing mechanism. QoS requirements for public-safety bearers derive from TS 22.280 [15] and TS 23.501 [16].

III. System Model

We model the IAB network as a directed graph $G = (V, E)$ rooted at donor node v_0. Each non-root node can have both access UEs and a parent backhaul link. Let $\mathcal{P}(i)$ denote the parent path from node i to the donor.

At node i, the available resource budget in scheduling window t is $W_i(t)$, partitioned among:

- commercial access traffic $\beta_i(t)$,
- public-safety access traffic $\alpha_i(t)$,
- backhaul forwarding budget $\eta_i(t)$.

Because access and backhaul are coupled, these quantities are not independent. For in-band operation,

$$\alpha_i(t) + \beta_i(t) + \eta_i(t) \leq W_i(t). \tag{1}$$

Let $Q_{ps}^{(i)}(t)$ denote the public-safety queue at node i and $\dot{Q}_{ps}^{(i)}(t)$ its short-term growth estimate. Let τ_{sig} denote the end-to-end signaling delay needed to notify upstream schedulers and let τ_{sw} denote the frame-switching or scheduling-reconfiguration delay.

IV. Proposed Mechanism: Preemptive Backhaul Borrowing

A. Predictive Trigger

Preemption should not begin only after the queue is already full. The trigger uses queue growth and delay bounds:

$$Q_{ps}^{(i)}(t) + \dot{Q}_{ps}^{(i)}(t) \cdot (\tau_{sig} + \tau_{sw}) > Q_{crit}^{(i)}. \tag{2}$$

When this condition holds, the node predicts that reactive scheduling will cause overflow and issues a `Borrowing-Advisory` for path $\mathcal{P}(i)$. The predictor uses a constant lookahead $\tau_{pred} = 0.12\,\text{s}$ to ensure timely firing regardless of path depth.

B. Congestion-Aware Downgrade

The mechanism first reclaims capacity from commercialCM2 (elastic best-effort commercial, 5QI 8/9), touching CM1 (GBR commercial) only if the path deficit remains. Reclaimed budget is reserved for backhaul forwarding and PS1/PS2 ingress; commercial shares are restored progressively after the burst ends to avoid synchronized congestion collapse. The `Borrowing-Advisory` maps onto an O-RAN near-RT RIC xApp (Section VII) or existing scheduler-control hooks.

C. Priority Model and Borrowing Workflow

Traffic is modeled using four priority classes mapped to standardized 5QI values from TS 23.501 [16]. **PS1**: 5QI 65 (MCPTT voice, GBR, PDB 75 ms) and 5QI 67 (MC video, GBR, PDB 100 ms). **PS2**: 5QI 69 (MC signaling, Non-GBR, PDB 60 ms) and 5QI 70 (MC data, Non-GBR, PDB 200 ms). Both PS classes are protected. **CM1** (5QI 1–4, GBR) is preserved where possible; **CM2** (5QI 8/9, Non-GBR, PDB 300 ms) is the first downgrade candidate.

When the trigger fires, the node issues a `Borrowing-Advisory` scoped to the upstream path. Parent schedulers apply the minimum downgrade to CM2 (CM1 only if the deficit persists), redirect reclaimed budget to PS1/PS2 forwarding, and restore commercial shares gradually after the burst. Three guardrails bound each event: a per-event downgrade cap, a maximum borrowing duration, and a progressive restoration slope (see Fig. 4).

V. Low-Complexity Heuristic

Solving the cross-hop borrowing problem exactly is infeasible under scheduling deadlines. We define a greedy heuristic: (1) rank upstream hops by predicted backhaul stress; (2) rank commercial flows by downgrade cost-per-recovered-resource; (3) reclaim only the minimum to satisfy the predicted deficit plus a guard margin. If $\mathcal{F}_i$ is the eligible set at node i, the borrowed capacity is $B_i(t) = \sum_{f \in \mathcal{F}_i^{sel}} r_f(t)$, where $\mathcal{F}_i^{sel}$ is the minimal prefix. Runtime is $O(H \cdot F \log F)$; for $H \leq 4$, $F = 30$, the p95 prototype time ranges from 3.28 μs in the representative stressed scenario to 3.5 μs across all 72 configurations.

VI. ns-3 5G-LENA Proof-of-Concept

To confirm scheduler-level implementability, we built a scenario in ns-3.47 [17] with the 5G-LENA NR module [18]: one donor gNB and two relay nodes across a three-node chain, with six UEs distributed unevenly: one PS UE attached to the deepest relay and five commercial UEs distributed across all three nodes. TDD slot borrowing is realized via runtime calls to `NrGnbPhy::SetPattern()`, switching from an access-heavy to a backhaul-heavy pattern when the borrowing condition fires. A per-relay `BapForwarder` handles IPv4 routing across the access/backhaul boundary. In a 6 s run with a PS burst at t=2–4 s, borrowing delivers the PS flow at 2.20 Mbps with 16.3 ms mean delay and zero packet loss. Without borrowing, the same PS flow delivers only 443/500 packets (88.6 %) at 487 ms mean delay. CM delivery remains at 97.5 % under borrowing, confirming that no commercial session is destroyed. This scenario is not a full PHY-accurate multi-hop study; its purpose is to confirm the mechanism works in 5G-LENA without modifying core internals. All policy comparisons in Section VIII use the system-level simulator.

VII. Adoption Path

A plausible near-term realization is an O-RAN near-RT RIC xApp on the E2 interface (10 ms–1,s control loop), placing the borrowing controller outside the gNB real-time path. The intended flow is: the leaf node queue predictor generates a Borrowing-Advisory; the near-RT RIC propagates scheduler-weight updates to parent IAB nodes via E2 Control; the non-RT RIC holds operator borrowing policy via A1. A concrete mapping candidate is O-RAN E2SM-RC (RAN Control) [19]: its REPORT service style can carry per-relay queue telemetry to the near-RT RIC, and its CONTROL service style can push scheduler weight updates back to individual IAB nodes. Whether current E2SM-RC exposure types suffice or require new E2 Setup Request information elements is an open implementation question. The Borrowing-Advisory is a paper-level abstraction, not an existing 3GPP or O-RAN primitive. Mapping it to specific F1AP procedures or BAP extensions requires prototyping work beyond this study.

VIII. Evaluation and Results

A. System-Level Simulator and Scenario Grid

We implemented a discrete-time system-level simulator modeling multi-hop IAB forwarding, per-hop commercial and public-safety queues, signaling delay, blockage-induced capacity drops, and policy-dependent scheduler decisions. Time steps are 10 ms; scheduling decisions apply at 40 ms frame boundaries, approximating slotized TDD updates rather than continuous reallocation.

Traffic model. Commercial traffic is modeled using a Poisson process at each hop (offered load: $0.4\times/0.7\times/0.9\times$ hop capacity for low/high/extreme) [20]. PS traffic is a constant bit rate (CBR) burst at $0.3\times$ hop capacity, representing aggregated MCPTT, MC video, and sensor telemetry, starting near t=15 s ($\pm$3 s jitter per seed) and lasting 15 s. Queue growth is estimated by an EWMA over 120 ms (equal to τ_{pred}), so the estimate spans exactly the borrowing lookahead horizon.

Scenario grid. We sweep 3 hop counts (2/3/4) $\times$ 3 commercial loads $\times$ 2 signaling delays (low 20 ms / high 300 ms) $\times$ 2 blockage conditions (none / moderate) $\times$ 2 burst severities (medium / severe) = 72 base configurations, each run with 5 random seeds, across 5 policies, yielding $72 \times 5 \times 5 = 1{,}800$ total runs. Fixed simulation parameters are summarized in Table II.

TABLE II
SYSTEM-LEVEL SIMULATOR PARAMETERS

Parameter	Value
Time step	10 ms
Simulation horizon	45 s
Burst window (mean)	t=15–30 s, $\pm$3 s onset jitter
Nominal hop capacity	70 Mbps
Default backhaul share	0.55 (all policies at rest)
Reactive shift target	0.70 (after threshold crossing)
Predictive shift target	0.85 (while advisory is active)
Queue threshold Q_{crit}	500 KB
Queue limit per class/hop	2 MB
Signal delay sweep	20–300 ms
Blockage model	middle-hop factor 0.68 during burst
Seeds per scenario	5
Scenario configs	72 (3 hops $\times$ 3 loads $\times$ 2 delays $\times$ 2 block. $\times$ 2 bursts)
Total runs	1,800 (72 configs $\times$ 5 seeds $\times$ 5 policies)

Backhaul share semantics. The three share values (0.55/0.70/0.85) are *not* per-policy tuning knobs; all policies share the 0.55 default and temporarily shift upward only during active borrowing. The comparison therefore measures each policy's ability to *time* a resource shift, not a permanently elevated allocation.

B. Metrics

Primary metrics are PS delivery ratio (fraction of generated traffic reaching the donor), PS drop ratio (queue overflow losses), PS p99 latency, Commercial (CM) delivery ratio, and scheduler decision time (p95 wall-clock of the prototype heuristic, included to validate low computational overhead rather than deployment-time firmware performance).

C. Aggregate Results and Policy Tradeoff

Fig. 2 compares all five policies in the representative stressed scenario. Predictive borrowing achieves 0.884 PS delivery with 3.28 μs scheduler time; the Q-learning baseline reaches 0.759 despite training on this scenario. Q-learning preserves more commercial delivery (0.718 vs. 0.634), confirming the heuristic is more aggressive in the regime it targets. Reactive global (0.573) and local-only (0.569) are close, confirming that global coordination adds little without the predictive trigger.

D. Controlled Fairness Ablation

To isolate the prediction benefit from the share-level benefit, Table III and Fig. 1 hold the maximum backhaul share constant across trigger types. When both reactive and predictive policies are given the same 0.85 ceiling, predictive borrowing achieves 0.844 PS delivery versus 0.562 for reactive, a 28-point gap that

TABLE III

CONTROLLED FAIRNESS ABLATION (REPRESENTATIVE STRESSED SCENARIO, 5 SEEDS). MAX SHARE HELD CONSTANT ACROSS ROWS TO ISOLATE TRIGGER TIMING.

Trigger	Max Share	PS Del.	CM Del.	Adv.
Reactive	0.70	0.560	0.618	1.0
Reactive	0.85	0.562	0.619	1.0
Predictive	0.70	0.659	0.614	6.0
Predictive	0.85	0.844	0.631	14.8

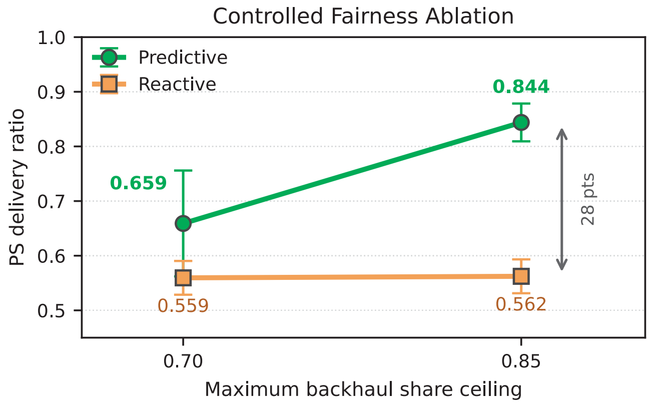

Fig. 1. Controlled fairness ablation: PS delivery at equal maximum share. Reactive delivery barely changes as the share ceiling rises (0.560→0.562); predictive delivery rises by 28 points at the same ceiling. The trigger timing, not the allocated share, determines outcome.

traces to the predictive trigger firing before overflow rather than after. Reactive barely improves going from 0.70 to 0.85 (0.560→0.562) because its threshold-crossing trigger fires too late and queue overflow has already occurred. Even when limited to the same 0.70 ceiling as the original reactive policy, predictive borrowing still reaches 0.659, 10 points above reactive at the same share. Fig. 1 shows this asymmetry: reactive delivery is flat as the share ceiling rises, while predictive delivery scales steeply. The trigger timing, not the allocated share, accounts for the difference.

E. Tabular Q-Learning Baseline Comparison

The tabular Q-learning agent uses a 15-state encoding (5 fill × 3 growth) and three actions (normal/reactive/predictive share), trained in-loop with ε=0.1. It serves as a lightweight learning comparator rather than a full RL benchmark.

Over the full sweep, Q-learning achieves 0.930 mean PS delivery versus 0.939 for predictive borrowing, effectively equivalent in aggregate. The difference emerges in the hardest regime (0.884 vs. 0.759, a 12.5-point gap), because the heuristic's trigger directly models $\tau_{sig} + \tau_{sw}$ while the Q-agent's state space omits signaling delay. To confirm that state representation is the bottleneck rather than the RL architecture, we extended the Q-agent to a 90-state variant encoding signaling delay level and hop count in addition to queue fill and growth. With comparable warm-up training, the 90-state agent achieves 0.725 PS delivery, within noise of the 15-state baseline (0.759) and still below the heuristic (0.884). Neither variant encodes the delay lookahead that makes the heuristic effective; a deep RL agent with continuous delay-aware state would be needed to narrow the gap (Section IX). Both Q-agent variants adjust the per-frame backhaul share directly; neither issues path-scoped Borrowing-Advisories. The advisory count is accordingly zero because the comparison tests the effect of state representation on share selection, not advisory issuance.

Beyond performance, the heuristic provides auditability: each advisory carries a timestamp, duration, and list of downgraded flows, which is important for post-incident review in public-safety operations.

F. Public-Safety Throughput Over Time

Fig. 3 shows per-second PS delivery in the representative stressed scenario. Static and local-only policies fail as the backhaul saturates; reactive global improves modestly but lags the burst onset. Predictive borrowing issues its first advisory before the queue fills, reducing the undelivered window. The Q-learning agent fires less aggressively early, then improves as queue state moves into regions where the trained policy selects more aggressive actions, consistent with its better CM preservation but weaker PS peak delivery.

G. Signaling Delay Sensitivity

Signaling delay was swept from 20 to 300 ms to evaluate sensitivity to control-path latency. Predictive borrowing degrades gracefully across this range, declining from 0.884 to 0.836 PS delivery at the highest delay. The Q-learning baseline remains comparatively stable at 0.765–0.771 because its learned policy responds to queue state rather than an explicit fixed delay estimate. Both policies remain above the reactive and local-priority baselines throughout the tested range, while the predictive mechanism retains its advantage by explicitly incorporating signaling-plus-switching delay into the trigger condition.

H. Representative Stressed Scenario Detail

In the representative stressed scenario, predictive borrowing raises PS delivery to 0.884 (±0.030), compared with 0.759 (±0.031) for the Q-learning baseline, 0.573 (±0.036) for reactive global, 0.569 (±0.036) for local-only priority, and 0.447 (±0.028) for static slicing. The borrowing controller issues approximately 15 advisories over the burst window, each holding the backhaul share at 0.85 for a bounded interval.

Across borrowing policies, aggregate CM delivery ranges from 0.634 (predictive) to 0.718 (Q-learning). The static baseline achieves 0.636 CM delivery, comparable to predictive borrowing, but drops more than half of all PS packets (0.447 PS delivery). The Q-learning agent preserves more CM traffic because it borrows less aggressively in the stressed regime. Fig. 5 plots PS delivery against CM delivery across all five policies over the full scenario sweep. Predictive borrowing sits at the high-PS end with CM delivery matching the static baseline, while static slicing achieves the highest CM delivery only by abandoning PS protection entirely. No policy dominates on both axes; the choice depends on the operator's PS priority weight.

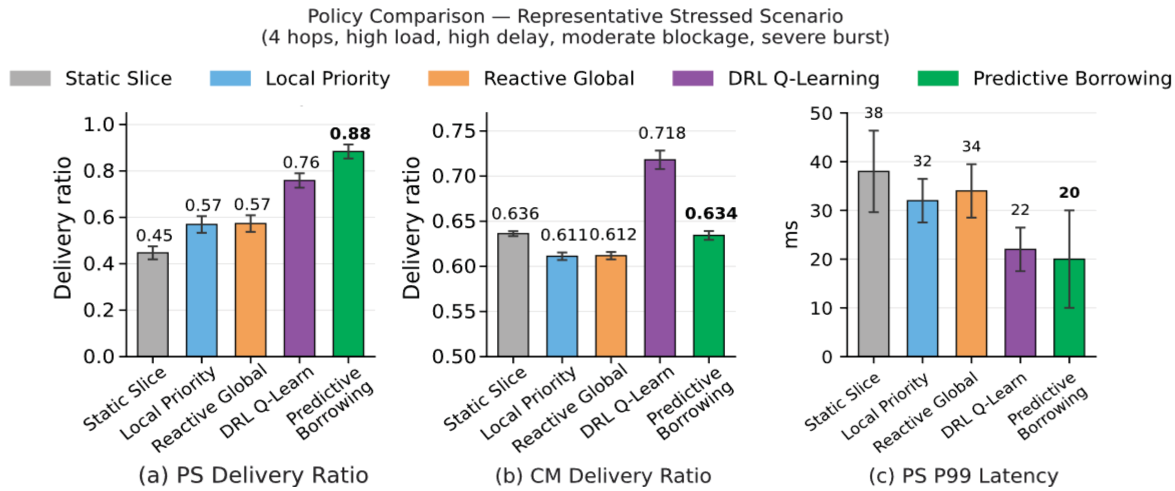

Fig. 2. Five-policy comparison in the representative stressed scenario (4 hops, high load, high delay, moderate blockage, severe burst). Error bars show ±1 std over five seeds. Predictive borrowing leads on PS delivery; Q-learning shows better CM preservation at the cost of lower PS delivery in the hardest regime.

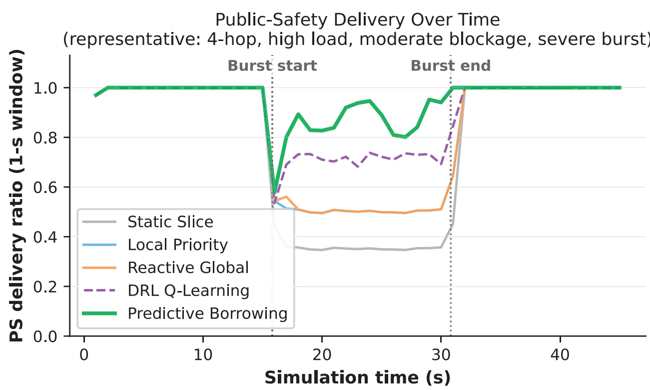

Fig. 3. Per-second public-safety delivery rate (representative stressed scenario). Dashed vertical lines indicate burst onset and end. Predictive borrowing and Q-learning both outperform reactive and static baselines; predictive borrowing maintains a higher delivery rate during the early burst phase.

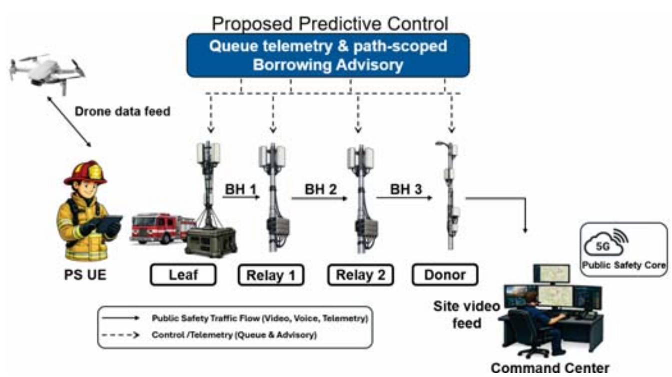

Fig. 4. Predictive backhaul-borrowing workflow across a multi-hop IAB chain. Public-safety traffic enters at the leaf and traverses successive wireless backhaul hops toward the donor and public-safety core. Queue telemetry informs the predictive controller, which issues a path-scoped Borrowing-Advisory to coordinate temporary resource reallocation along the affected IAB path.

I. Boundary Cases

In a 2-hop, low-load baseline scenario, all borrowing-capable policies achieve near-full PS delivery, confirming

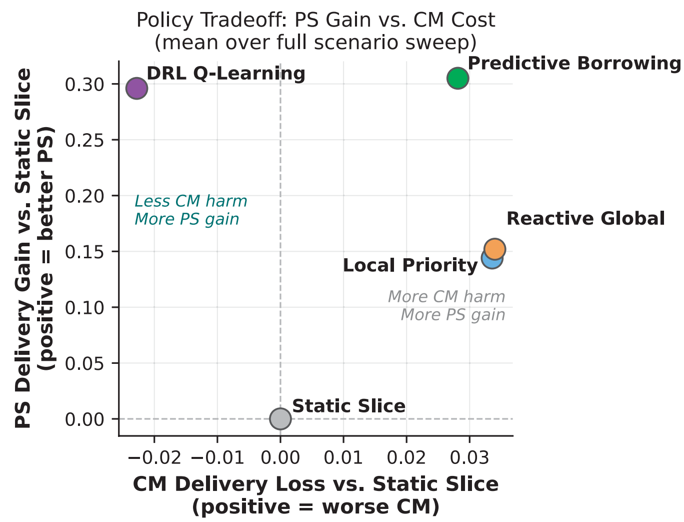

Fig. 5. PS delivery versus CM delivery across all five policies (full 72-configuration sweep). Each point aggregates one policy's mean over all seeds and scenario conditions. Predictive borrowing reaches the highest PS delivery at a CM cost comparable to the static baseline; static slicing maximizes CM delivery while dropping the majority of PS traffic.

the mechanism is most impactful under stressed multi-hop conditions. In the false-positive case (no incident, high load), predictive borrowing emits zero advisories and CM delivery matches the static baseline within noise, confirming the trigger does not fire spuriously. In the harsher 4-hop, extreme-load, severe-burst case, PS delivery remains above 0.88 while CM delivery falls to 0.480 versus 0.475 for static, a bounded commercial penalty relative to the PS gain.

J. Parameter Sensitivity

An ablation over three guard margins (1.02, 1.08, 1.15) and three restoration slopes (0.05, 0.10, 0.20) in the stressed scenario shows PS delivery ranging from 0.867 to 0.899. The default (guard 1.08, slope 0.10) at 0.884 is chosen for advisory economy rather than commercial protection: the most aggressive setting (1.02, 0.20) reaches 0.899 with CM delivery unchanged within noise (0.633 vs. 0.634).

IX. Limitations and Future Work

The simulator uses a fluid TDD approximation; it omits inter-cell interference, beam management, and HARQ. Results are control-logic validation, not PHY-layer predictions. The ns-3 proof-of-concept confirms implementability in a single scenario; a full multi-UE multi-hop study remains future work. All topologies are linear chains; branching introduces competing-leaf coordination challenges not modeled here. The Q-learning baseline is a lightweight 15-state comparator; a deep RL agent with delay-aware state would likely narrow the stressed-regime gap. The controlled ablation (Table III) isolates the prediction benefit at equal share levels; a full share-value grid sweep remains future work.

Next steps include: (1) full PHY-accurate ns-3 multi-hop evaluation; (2) investigation of whether the Borrowing-Advisory maps to existing F1AP, BAP, or OAM messages; (3) branching topologies with concurrent incidents; and (4) comparison against a deep RL agent with delay-aware state encoding.

X. Conclusion

We have presented and evaluated a predictive preemptive backhaul borrowing framework for mixed commercial and public-safety traffic in 5G IAB networks. The core finding is that path-aware coordination, triggered by a constant-horizon queue-growth predictor, improves public-safety delivery in stressed multi-hop regimes where static slice reservation and local-only priority both fail. Over 1,800 runs spanning 72 scenario configurations, predictive borrowing achieves 0.939 mean PS delivery with a 3.28 μs prototype scheduler overhead in the representative stressed scenario (3.5 μs mean across all configurations), near-equal to the tabular Q-learning baseline (0.930) in aggregate but 12.5 percentage points ahead in the hardest stressed regime (0.884 vs. 0.759). The heuristic's advantage in that regime stems from its explicit modeling of signaling-plus-switching delay; its advantage in operational settings stems from per-event auditability. Reactive global (0.786) and local-only priority (0.778) confirm that global coordination without prediction yields only modest gains over local treatment. Commercial delivery remains non-zero under all borrowing policies.

The mechanism is framed as a proposed scheduler extension, rather than a feature of existing 3GPP specifications. Its proximity to O-RAN near-RT RIC control patterns and existing QoS primitives suggests a feasible prototyping path, but the specific signaling requirements and multi-vendor integration challenges remain open and warrant deeper PHY-level and standards-track evaluation.

Acknowledgment

This research was supported in part by the U.S. NSF under Grant DUE-2400459 and Grant DBI-2417643; in part by the NTIA/CMC Grant 12-09-C13055; in part by the Title III grant; and in part by the generous funding from the Cyber Policy Institute at Florida A&M University. The authors thank Top Tier Quality, LLC for providing the computational infrastructure used to conduct the simulation experiments in this work. Dr. Jinwei Liu is the corresponding author.

References

[1] A. Othman and N. A. Nayan, "Public safety mobile broadband system: From shared network to logically dedicated approach leveraging 5G network slicing," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2109–2120, 2021.

[2] P. Fiore, I. Filippini, and D. D. Donno, "Meeting future mobile traffic needs by peak-throughput design of next-gen ran," *IEEE Trans. Mob. Comput.*, vol. 25, no. 5, pp. 6083–6101, 2026.

[3] M. Yu, Y. Pi, A. Tang, and X. Wang, "Coordinated parallel resource allocation for integrated access and backhaul networks," *Computer Networks*, vol. 222, p. Art. 109533, 2023.

[4] A. Zhivtsova, V. Beschastnyi, Y. Koucheryavy, and K. Samouylov, "A survey of delay-oriented dynamic link scheduling policies for 5g/6g integrated access and backhaul systems," *IEEE Access*, vol. 12, pp. 118 565–118 586, 2024.

[5] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated access and backhaul in 5G mmwave networks: Potentials and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, 2020.

[6] M. Cudak, A. Ghosh, A. Ghosh, and J. Andrews, "Integrated access and backhaul: A key enabler for 5G millimeter-wave deployments," *EEE Commun. Mag.*, vol. 59, no. 4, pp. 88–94, 2021.

[7] M. Pagin, T. Zugno, M. Polese, and M. Zorzi, "Resource management for 5G NR integrated access and backhaul: A semi-centralized approach," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 2, pp. 753–767, 2022.

[8] H. Alghafari and M. S. Haghighi, "Decentralized joint resource allocation and path selection in multi-hop integrated access backhaul 5G networks," *Computer Networks*, vol. 207, p. Art. 108837, 2022.

[9] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On radio access network slicing from a radio resource management perspective," *IEEE Wirel. Commun.*, vol. 24, no. 5, pp. 166–174, 2017.

[10] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, 2018.

[11] A. Othman and N. A. Nayan, "Efficient admission control and resource allocation mechanisms for public safety communications over 5G network slice," *Telecommun. Syst.*, vol. 72, no. 4, pp. 595–607, 2019.

[12] W. Lei, Y. Ye, and M. Xiao, "Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 970–979, 2020.

[13] 3GPP, "Study on integrated access and backhaul," 3GPP, Technical Report TR 38.874, 2019, release 16.

[14] 3GPP, "NR; backhaul adaptation protocol (BAP) specification," 3GPP, Technical Specification TS 38.340, 2020, release 16.

[15] 3GPP, "Mission critical services common requirements (MCCoRe); stage 1," 3GPP, Technical Specification TS 22.280, 2022, release 17.

[16] 3GPP, "System architecture for the 5G system (5GS)," 3GPP, Technical Specification TS 23.501, 2022, release 17.

[17] ns-3 Consortium, "ns-3 network simulator," [Online]. Available: https://www.nsnam.org, 2024.

[18] N. Patriciello, S. Lagén, B. Bojović, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simul. Model. Pract. Theory*, vol. 96, p. Art. 101933, 2019.

[19] O-RAN Alliance, "O-RAN WG3: E2 service model (E2SM) RAN control (RC)," O-RAN Alliance, Technical Specification O-RAN.WG3.E2SM-RC-R003-v03.00, 2023, june 2023, Release R003.

[20] N. Polyakov and A. Platonova, "Assessing latency of packet delivery in the 5g 3gpp integrated access and backhaul architecture with half-duplex constraints," *Future Internet*, vol. 14, no. 11, p. 345, 2022.