OPCM: Opportunistic Performance-driven Connectivity Management for 5G/xG Networks

AHMAD HASSAN, University of Southern California, USA
WEI YE, University of Minnesota, Twin Cities, USA
ANLAN ZHANG, University of Southern California, USA
ROSTAND A. K. FEZEU, University of Minnesota, Twin Cities, USA
JASON CARPENTER, University of Minnesota, Twin Cities, USA
RUIYANG ZHU, University of Michigan, USA
SHUOWEI JIN, University of Michigan, USA
MYUNGJIN LEE, Cisco Research, USA
AKSHAY JAJOO, Cisco Research, USA
MORLEY MAO, University of Michigan, USA
ZHI-LI ZHANG, University of Minnesota, Twin Cities, USA
FENG QIAN, University of Southern California, USA

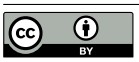
5G and future 6G networks deploy cells with diverse combinations of access technologies, architectures, and radio frequency bands/channels. Cellular operators also employ carrier aggregation for higher data access speeds. We investigate the fundamental question of how to intelligently and dynamically configure and reconfigure a user equipment's serving cells to deliver the best network performance. Through comprehensive measurements across 12 cities in 5 countries, we experimentally show the wide availability, heterogeneity, and untapped performance gains of today's cell deployments. We then present a principled, performance-driven connectivity management framework, dubbed OPCM. It is a centralized solution deployed at the base station, allowing it to coordinate multiple UEs, enforce operator policies, and facilitate user fairness. Extensive evaluations show that OPCM improves the application QoE by up to 65.2%.

CCS Concepts: • **Networks** → **Mobile networks**; *Network measurement*; *Network performance analysis*.

Additional Key Words and Phrases: 5G; xG; Connectivity Management; Band Switching; Band Selection.

ACM Reference Format:
Ahmad Hassan, Wei Ye, Anlan Zhang, Rostand A. K. Fezeu, Jason Carpenter, Ruiyang Zhu, Shuowei Jin, Myungjin Lee, Akshay Jajoo, Morley Mao, Zhi-Li Zhang, and Feng Qian. 2025. **OPCM**: Opportunistic Performance-driven Connectivity Management for 5G/xG Networks. *Proc. ACM Netw.* 3, CoNEXT4, Article 23 (December 2025), 23 pages. https://doi.org/10.1145/3768970

Authors' Contact Information: Ahmad Hassan, University of Southern California, USA, ahmadhas@usc.edu; Wei Ye, University of Minnesota, Twin Cities, USA, ye000094@umn.edu; Anlan Zhang, University of Southern California, USA, anlanzha@usc.edu; Rostand A. K. Fezeu, University of Minnesota, Twin Cities, USA, fezeu001@umn.edu; Jason Carpenter, University of Minnesota, Twin Cities, USA, carpe415@umn.edu; Ruiyang Zhu, University of Michigan, USA, ryanzhu@umich.edu; Shuowei Jin, University of Michigan, USA, jinsw@umich.edu; Myungjin Lee, Cisco Research, USA, myungjle@cisco.com; Akshay Jajoo, Cisco Research, USA, ajajoo@cisco.com; Morley Mao, University of Michigan, USA, zmao@umich.edu; Zhi-Li Zhang, University of Minnesota, Twin Cities, USA, zhzhang@cs.umn.edu; Feng Qian, University of Southern California, USA, fengqian@usc.edu.



ACM 2834-5509/2025/12-ART23
https://doi.org/10.1145/3768970

1 Introduction

5G New Radio (NR) supports a wide range of frequency bands, including Low-Band (<1GHz), Mid-Band (1–6GHz), and High-Band (mmWave, 24–40GHz). Since its rapid global rollout in 2019, 5G has largely coexisted with 4G/LTE. To leverage both technologies, most operators deploy Non-Standalone (NSA) mode with Dual Connectivity (DC), while some have also introduced Standalone (SA) mode. These architectures coexist, adding enormous complexity to modern networks. Like 4G, 5G employs Carrier Aggregation (CA) to boost the overall bandwidth and throughput [22, 77]. As a result, operators often configure multiple cells per base station, and use one or a combination of them to serve a user equipment (UE).

To support these technological advancements, most research and commercial efforts have concentrated on resource management within a single cell [21, 32, 33, 35, 66]. However, the foundational question of *which cell(s) a user should be served by* in the first place remains relatively under-explored, despite its growing relevance in today's heterogeneous and multi-cell network landscapes. To configure a UE's cells, 3GPP defines a set of procedures that we collectively refer to as Connectivity Management (CM) procedures for brevity. These include cell selection, reselection, handover, and CA/DC. The decision-making criteria for configuring these CM procedures are primarily based on radio link quality; specific configurations to make these decisions are largely left to operators.

Issues with Legacy CM Schemes. Legacy CM has three key limitations. ***(i)*** While radio link quality criterion has historically ensured seamless connectivity, legacy CM is performance-oblivious and lacks support for metrics like throughput and energy efficiency. ***(ii)*** CM procedures operate independently, leading to redundant logic, increased management complexity, misconfigurations [25, 46, 58, 79], and, in extreme cases, network outages [29, 59, 60]. ***(iii)*** Legacy CM is network-centric, applying uniform criteria to all UEs. However, different traffic types demand different solutions—*e.g.,* one cell may be better for latency-sensitive traffic, while another offers higher throughput. Existing schemes [24, 44] do not support such UE-level performance personalization. These limitations contribute to the perception that 5G has been disappointing, with some even claiming it fares worse than 4G [67], despite its potential for higher throughput [61].

Our Proposal. We propose a unified, performance-driven framework for CM that provides a *higher-level abstraction* over existing CM procedures. By decoupling CM actions (adding, removing, or modifying a UE's cells) from performance criteria (throughput, latency, energy, *etc.*), our approach allows operators to flexibly plug in diverse metrics as needed (see Fig. 1). This design enhances user performance, reduces CM complexity, and supports UE-level personalization, all while remaining backward-compatible with existing infrastructure. For example, the network may aggregate high-bandwidth cells for backlogged flows, select low-latency cells (with higher subcarrier spacing) for real-time traffic, or prefer energy-efficient cells for lightweight workloads. In cases prioritizing cell coverage, standard radio link quality metrics remain applicable.

The rationale for our proposal can be distilled into three core insights. ***First***, extensive measurements across 12 cities in 5 North American and European countries reveal the *wide availability and heterogeneity* of cell deployments (§2.3). In the median case, UEs can access 7+ unique cell combinations spanning NSA-5G, SA-5G, and LTE in broad frequency ranges. These deployments remain stable both spatially and temporally. ***Second***, large-scale experiments highlight significant, *untapped performance gains* enabled by the diversity of available cells (§2.4). The missed potential can be up to 70.1% depending on the Quality of Experience (QoE) metric (*e.g.,* video bitrate, per-frame latency and energy consumption). This contrasts with earlier deployments, which prioritized seamless connectivity over application performance. ***Third***, while operators have traditionally used simple heuristics for CM, there has been an interesting *shift in trends* recently. Newer 3GPP specification

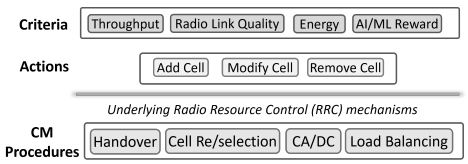


Fig. 1. Decoupling CM criteria and underlying CM procedures via an abstraction layer.

Table 1. Comparing existing CM techniques with OPCM.

Features		Legacy	iCellSpeed	OPCM
Supports Perf. Metric	Link Quality	✓	✓	✓
	Throughput	✗	✓	✓
	Latency	✗	✗	✓
	Energy	✗	✗	✓
Fair to other UEs		✓	✗	✓
Respects RAN Policy		✓	✗	✓
Is 3GPP Compliant		✓	✗*	✓

*iCellSpeed bypasses standard 3GPP CM procedures.

releases introduce innovative CM techniques to improve user performance (*e.g.,* conditional and DAPS handovers [1]). This makes us rethink legacy CM practices.

Challenges. While a performance-driven CM framework promises significant gains for users, several challenges must be addressed. *First*, managing CM for 10s-100s of UEs is complex due to diverse performance requirements and heterogeneous cell deployments. *Second*, the solution must comply with operator policies. *Third*, accurate network performance profiling is challenging—active probing risks performance degradation, while passive techniques introduce approximation errors. *Fourth*, CM procedures like handovers risk causing data interruptions.

Opportunistic Performance-driven CM. We present OPCM to tackle these challenges. It operates opportunistically, ensuring at least the performance of legacy CM while seeking additional gains. A key design question is: *who* should be responsible for performance-driven CM? A user-side solution, like iCellSpeed [24], is immediately deployable on commercial 4G/5G networks; however, since each UE would individually handle CM, it cannot guarantee cross-user fairness or fulfill operator-imposed polices. In contrast, a centralized network-side solution can coordinate multiple UEs, enforce operator policies at various scopes, and facilitate fairness. OPCM sits on the base station, requiring no UE-side modifications. It remains 3GPP-compliant and fully backward-compatible with legacy CM. OPCM also exposes a modular custom metric registration API that reduces CM configuration complexity. Table 1 summarizes key differences between CM schemes, highlighting OPCM's broader metric support, policy compliance, and fairness guarantees.

We design three key components in OPCM. First, the **Smart Decision Framework (§4)** simplifies the multi-UE CM problem by decomposing it into independent single-UE decisions while pruning cell combinations that violate fairness or operator policies. It then opportunistically selects cell combinations from the pruned set to exploit missed performance gains. Second, the **Hybrid Profiling Engine (§5)** leverages cross-correlation among cell combinations to passively estimate performance, minimizing profiling overhead. Third, the **Robust Execution Module (§6)** employs queuing-aware delayed reconfiguration to reduce data interruptions. It also includes a fallback mechanism that reverts to legacy radio link quality criterion when gains are marginal.

Prototyping and Evaluation. We prototype OPCM on a programmable over-the-air testbed with open-source LTE/5G cellular suites [64, 65] in total 6.1K+ lines of code. Our experimental evaluation combines small-scale over-the-air experiments (for realism) with large-scale trace-driven simulations (for generality). We also test OPCM over multiple phone models, under diverse mobility patterns (walking, driving, *etc.*) from various locations (downtown, campus, suburban), and using real application workloads (*e.g.,* video-on-demand and video analytics). Key experimental takeaways are: ***(i)*** OPCM offers up to 65.2% and 28.1% higher average Quality-of-Experience (QoE) than the legacy CM and a UE-side CM solution iCellSpeed [24], respectively (§8.2). ***(ii)*** OPCM performs as well as the legacy CM under mobility, and can effectively find the highest performing cell combinations in the wild (§8.3). ***(iii)*** OPCM satisfies operator policies: cross-UE fairness is within 98-99% of the legacy CM, while the spectral efficiency improves by 2-3%. It incurs small (3.1-6.5%) system overhead, and efficiently manages advanced CA/DC settings (§8.4).

2 Background & Motivation

2.1 A 4G/5G Primer

Cells. At any location, a UE is under the coverage of one or more serving cells, each operating on a continuous frequency block called a component carrier (CC). A Base Station (BS)—eNodeB in 4G and gNodeB in 5G—houses the physical infrastructure for one or more cells. A key feature of modern cellular networks is CA, which combines multiple cells to boost data speeds. NSA-5G Dual Connectivity (DC) is essentially a form of CA, as it combines 4G and 5G cells. Serving cells include a mandatory Primary Cell (PCell), responsible for control signaling and data access, and optional Secondary Cells (SCells) that enhance data transmission. In NSA-5G, the Primary Secondary Cell (PSCell) acts as the anchor for 5G signaling in addition to the 4G PCell. We use the term *cell combination* to refer to a UE configured with a mandatory PCell and optional PSCell and SCells. A cell combination is considered *unique* if it differs in any of its serving cells (PCell, PSCell, or SCells).
CM Procedures. Connectivity Management (CM) procedures determine the serving cells to which a UE connects at any given moment. For instance, cell selection identifies the most suitable cell during initial access, while cell reselection transitions an idle UE to a new cell. Handovers ensure seamless continuity of service by transferring active connections between cells. Load balancing redistributes traffic among cells to prevent congestion and optimize network utilization. CA/DC combines multiple cells to enhance bandwidth and data throughput. Table 6 in Appendix A provides a detailed overview of legacy CM schemes. They perform similar actions (add, modify, or remove a UE's cells) based on UE's radio connection state (idle, inactive, or connected) and criteria (radio link quality, absolute priority, and cell accessibility).
Terminology. For brevity, we adopt a few rules to talk about cell combinations: *(i)* $61^5(3350)$ refers to a UE connected to a 5G PCell with Physical Cell ID (PCI) 61 operating at a frequency of 3350 MHz; *(ii)* $85^4(1850)/61^5(3350)$ is an NSA-5G UE with a 4G PCell 85 and a 5G PSCell 61; and *(iii)* $85^4(1850)/61^5(3350)/\{108^5(3370),111^5(3330)\}$ adds two 5G SCells 108 and 111 to the previous setting.

2.2 Measurement Setup

Operator, Band & Technology. We select three major commercial cellular operators (T-Mobile. AT&T & Verizon) in the US for our experiments. These operators have deployed their cellular services using 4G/LTE, NSA-5G, and SA-5G in the Low-Band, Mid-Band, and mmWave radio frequencies. To collect data in Europe, we use the countries' local cellular operators (Vodafone, Telekom, SFR, and Orange). European operators also support 4G/LTE and NSA-5G in Low-Band and Mid-Band range [39]; however, only Vodafone offered SA-5G at the time of our study. Since mmWave 5G coverage is still low across the US and Europe [7, 39, 70], we mainly focus on Low-Band and Mid-Band (600–3300 MHz) in our study.
Measurement App. We develop an Android app to test different application use cases in the wild. ***(i)*** *Live Video Streaming:* our streaming pipeline consists of three components: (a) an RTSP media server [3], (b) an ffmpeg-based encoder [2], and (c) a lightweight Android RTSP client [26]. We transmit pre-recorded video content at 30 FPS with a target bitrate of 34 Mbps, matching typical 4K streaming requirements. ***(ii)*** *Video Conferencing:* we implement a minimal Android client using a commercial WebRTC SDK [71], maintaining compatibility with standard real-time communication protocols. ***(iii)*** *Energy Profiling:* We employ pre-built energy consumption models to profile a UE's energy efficiency as a function of its network throughput, serving cell combination, and device model. This model-based approach offers practical advantages over deprecated Android PowerProfile APIs and external hardware monitors, avoiding both software limitations and the need to tether devices to measurement equipment during live, in-the-wild experiments. To construct models, we follow the methodology of Narayanan et al. [56]: we collect current draw traces using

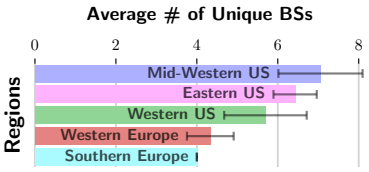


Fig. 2. Density of cell deployments across different regions.

Table 2. Cell combinations used in our experiments. # of carriers (SCells) vary over time.

Label	Cell Combination
S1	$113^4(1955)/39^5(626)$
S2	$107^4(1935)/278^5(2510)$
S3	$59^4(2145)$
S4	$46^5(636)$
S5	Legacy CM

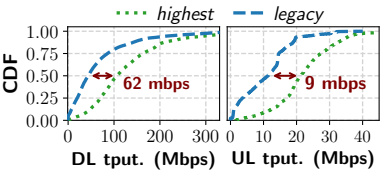


Fig. 3. Performance breakdown of cell combinations in our dataset.

the Monsoon Power Monitor [37] at varying sending rates, and then fit linear regression models to derive per-device power consumption profiles. We use a university-hosted server with 4 Gbps+ network bandwidth to host all the applications. Hence, the Internet was not a bottleneck. We also capture other key information using Android APIs, such as geolocation, mobility speed, signal strength, BS Id, and ping latency.

Methodology. We use multiple smartphone models to minimize the impact of smartphone diversity: Samsung Galaxy S10 (S10), S20 Ultra (S20U), S21 Ultra (S21U), and S22+ (S22+). These phones have diverse radio capabilities. We use a laptop – tethered to the phones via USB3 cables – to control and time-synchronize the experiments through ADB [23] commands. To extract lower-layer signaling messages on unrooted UEs, we rely on a professional tool Accuver XCAL [16]. Concurrent phone measurements can occasionally affect performance results. To mitigate this, we lock UEs to separate BSs when possible, benchmark devices before each test, and repeat experiments for consistency.

2.3 Wide Availability of Cell Deployments

To geographically map the footprint of cell deployments, we use band-locking code (∗#2263#) to explore cells a UE can access. ping is used to test Internet connectivity. We conducted experiments across 90+ locations in 12 cities in the US and Europe, repeating each experiment 3× per location.

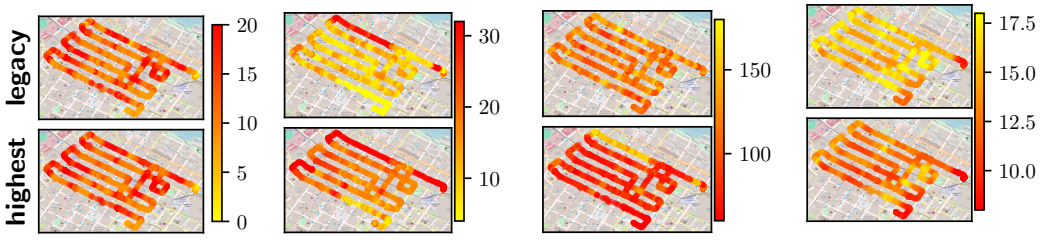
Our analysis reveals that *cell deployments are spatially stable and abundant*. Fig. 2 shows the number of unique BSs accessible to a UE in different regions. At any location, more than 3 BSs can be accessed approximately ∼94% of the time. The median case shows 3–6 BSs per location, with some locations reaching up to 8. With CA/DC considered, the number of unique cell combinations exceeds 7 in the median case. We notice denser cell deployments in the US compared to Europe. The disparity is caused by the limited availability of SA-5G and NSA-5G Low-Band in Europe during the study period: only one European operator offered SA-5G, and NSA-5G Low-Band was absent in our dataset. In contrast, US operators extensively used NSA-5G Low-Band for 5G services [34, 39].

To verify the *temporal stability of cell deployments*, we conducted a 13-month campaign at four fixed locations (university campus, suburban residential area, downtown plaza, and airport) in a major US city. Results show 5–6 cell combinations observed 95% of the time across all locations. The cell combinations always remained the same except when T-Mobile added SA-5G Mid-Band cells at two of the four locations, or Verizon and AT&T added NSA-5G (C-Band) cells at all locations.

2.4 Performance Diversity in Cell Deployments

With the widespread availability of cell deployments, a key question arises: do the cell combinations exhibit performance diversity both within and across metrics, and how large is the gap between the legacy CM schemes, optimized for seamless connectivity, and performance-driven approaches?

A case study. To evaluate performance diversity, we select a 740m × 510m rectangular loop on our university campus. Using XCAL, we confirm that a T-Mobile UE can connect to the same four PCells (and PSCells for NSA-5G), each served by a unique BS, at any location. Four S22+ UEs are locked to the four identified cell combinations, while a fifth S22+ uses cell combinations assigned by *legacy* CM. The cell combinations include LTE, SA-5G, and NSA-5G, with UEs free to perform CA.



(a) SINR (dB) (b) Video bitrate (Mbps) (c) Per-frame latency (ms) (d) Energy cons. (mW/Mbps)

Fig. 4. A case study to quantify the performance gap between legacy CM and the highest performing cell combinations. The color gradient ranges from red (highest performance) to yellow (lowest performance).

Table 2 lists the PCIs and frequencies for all five settings **S1-5**. The devices are placed side-by-side and collect data concurrently at walking speed. Each UE runs three apps: live video streaming, video conferencing, and energy profiling (§2.2). For video streaming, re-buffering time is negligible (<0.03%), so we only plot video bitrate. For video conferencing, we report per-frame latency, as the sending rate is low (~2 Mbps [53]). Additionally, we monitor Signal to Interference and Noise Ratio (SINR) for PCells/PSCells.

Fig. 4 plots SINR, live video streaming bitrate, video conferencing latency, and energy consumption. The top plots represent the legacy CM's performance (**S5**), while the bottom ones depict the highest performance across the other four settings (**S1–S4**), with darker red indicating better performance. Our experiments reveal four main insights. ***First***, legacy CM achieves SINR comparable to the best setting (Fig. 4a), as it is optimized for seamless connectivity. ***Second***, no single setting performs best across all metrics, and the gap between legacy CM and the best setting is significant (27.0-70.1%) for video bitrate, latency, and energy efficiency. This diversity arises from factors such as operator deployment preferences (*e.g.,* wider bandwidths improve throughput, while lower subcarrier spacings reduce latency), device capabilities, and CA priorities. ***Third***, even for a single metric, no setting consistently offers the highest performance. For example, **S2** achieves the highest video bitrate 43.2% of the time, but the lowest latency 11.3% of the time and is most energy-efficient 26.1% of the time. ***Fourth***, some operators aggressively camp users on 5G channels (*e.g.,* in C-band or mmWave band), or aggregate carriers for wider bandwidths. However, wider bandwidth does not guarantee higher throughput, and factors like channel variability and the disparate performance of CA combinations play crucial roles too [39, 77].

Next, we conduct large-scale experiments to characterize the network throughput gap across various mobility scenarios, including walking, driving, light rail, public bus, indoors, and stationary. The setup remains the same as before, with UEs band-locked to four diverse configurations (LTE Mid-Band, NSA Low-Band, NSA Mid-Band and SA Low-Band) rather than specific cell combinations. Each UE runs file transfer apps for both uplink and downlink directions.

Fig. 3 plots the downlink and uplink throughput CDF achieved by *legacy* CM and compares it with the *highest* (the UE with maximum throughput among all five devices). There are four key takeaways. ***(i)*** Although *legacy* CM provides the best link quality in 83.8% cases, it has the highest downlink throughput only 16.6% of the time. The uplink is even more drastic where *legacy* CM has maximum throughput a mere 3.6% of the time. ***(ii)*** The throughput gap between *highest* and *legacy* is huge: 143.1% (62 Mbps) median gap for downlink and 86.8% (9 Mbps) for uplink. ***(iii)*** In the median case, 2-3 cell combinations have better throughput than the cell combination configured by *legacy* CM. ***(iv)*** The uplink and downlink have the same *highest* cell combination only 42.1% of the time, implying that the uplink and downlink CM must be handled separately.

Overall, our findings reveal that *no single cell combination consistently outperforms across all metrics, underscoring the need for adaptive CM solutions.*

3 Opportunistic Performance-driven Connectivity Management (OPCM)

Here, we introduce OPCM (**O**pportunistic **P**erformance-driven **CM**) framework. Its design needs to overcome several challenges. **[C1]** Performing CM for 10s–100s of UEs presents significant complexity due to varying performance requirements and heterogeneous cell deployments. **[C2]** The framework must adhere to operator policies, such as ensuring cross-user fairness and balancing cell loads. **[C3]** Network performance profiling entails either actively switching to cells with potentially degraded performance or using passive approximation techniques, which are often error-prone. The challenge lies in determining the best approach to minimize overhead while ensuring accuracy. **[C4]** CM procedures, such as handovers, can cause data interruptions. Minimizing these disruptions is critical to maintaining seamless user experiences.

3.1 Design Overview

Fig. 5 illustrates how OPCM works compared to the legacy CM schemes. Typically, the serving BS configures ❶ a UE to continuously measure ❷ neighboring cells based on the radio link quality criteria. Once a configured criteria is met, the UE reports ❸ it to the serving BS. The serving BS then decides ❹ if it should initiate a CM procedure. If yes, the serving BS sends a command to the UE and helps execute ❺ the procedure. In the above process, ensuring that OPCM effectively selects the best serving cells according to the configured performance criteria without sacrificing other UEs' performance is the key to its adoption. Therefore, we design our system to be *opportunistic*: it strives to be as good as legacy CM while searching for additional performance gains. OPCM only replaces the decide ❹ and execute ❺ steps with its own *Decision Framework* and *Execution Module*, respectively. The underlying cellular procedures remain the same. It also introduces *Profiling Engine* to support diverse performance criteria. The *Data Manager* collects network, signal quality, and measurement report data readily available at the BS.

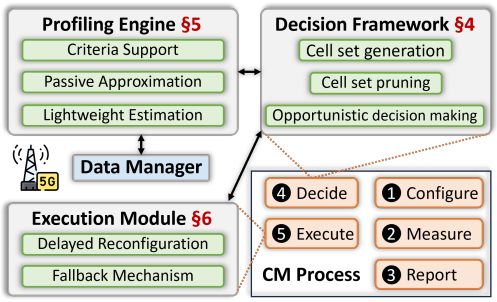


Fig. 5. The overall design workflow of OPCM.

The performance-driven CM involves solving an optimization problem that jointly considers all UEs, cell deployments, performance criteria, and RAN policies, leading to a scalability challenge (**C1**). To overcome it, the **Smart Decision Framework (§4)** strategically decomposes the monolithic multi-UE CM problem into multiple single-UE CM problems. It further reduces complexity by eliminating infeasible cell combinations. To achieve this, OPCM first generates a cell set containing all feasible cell combinations. Our experience from measuring public networks helps us reduce the number of cell combinations to consider. Then, we perform cell set pruning to remove cell combinations that may violate operator policies (**C2**), further reducing the cell set size. Finally, OPCM opportunistically decides which cell combination to use based on the performance estimates from *Profiling Engine*. It also balances the tradeoff between switching to a new cell combination (to profile its performance) and using the same cell combination (to maximize gains) (**C3**).

The **Hybrid Profiling Engine (§5)** estimates the performance of cell combinations in the cell set based on the appropriate criterion (radio link quality, network throughput, latency, and energy efficiency). OPCM leverages performance cross-correlation among cell combinations to opportunistically approximate performance without having to use a cell combination (**C3**). Finally, the **Robust Execution Module (§6)** executes the CM decisions. It employs delayed reconfiguration

to reduce data interruptions (**C4**). It also runs a fallback mechanism that reverts UE to the legacy radio link quality-based criterion if the performance gains are small.

Deployment Practicality. ***(i)*** OPCM operates as a lightweight service at the BS, requiring no changes to UEs. ***(ii)*** It adheres to standard 3GPP mechanisms and existing infrastructure; we do not replace the existing UE-side radio resource management (RRM) measurements which are essential for initial access, channel adaptation, *etc.* ***(iii)*** OPCM is backward-compatible; we implement legacy radio link quality-based CM on it and benchmark the performance (§8.3). ***(iv)*** Similar to standard X2/Xn handovers, it allows BSs to communicate on X2/Xn interfaces for shared performance profiling and coordinated CM decisions (§10). ***(v)*** To ensure that OPCM is opportunistic, i.e, it perform at least as well as *legacy* CM, we build mechanisms such as hysteresis, legacy fallback, and marginal-gain avoidance. ***(vi)*** Since the best cell combination for uplink and downlink may differ (§2.4), OPCM applies CM logic separately to each direction. For brevity, we only describe the downlink procedure in the remainder of this paper.

4 Smart Decision Framework

4.1 Cell Set Generation

At any moment, each UE is under the coverage of multiple cells with varying configurations. The serving BS configures all connected UEs to report measurements of neighboring cells (❸ in Fig. 5). Additionally, UEs can be configured with multiple serving cells, referred to as a cell combination in this paper (§2.1). OPCM leverages these radio link quality measurement reports, along with the UE's current cell combination, to construct a cell set $C = \{c_i | i \in \mathbb{N}\}$, where each cell combination c_i can have multiple serving cells. Furthermore, OPCM continuously updates C during runtime in case a new c_i is seen or an old one disappears.

UEs are typically surrounded by 3–6 BSs (§2.3). Each BS can have 3-4 PCells and up to 8 SCells, leading to over 100 possible cell combinations in the worst case. Profiling these many combinations is prohibitively expensive. Our measurements (Fig. 22 in Appendix B) reveal that operators limit cell combinations in practice by adding multiple SCells at once based on threshold-based policies [27]. This insight contrasts with the findings of CA++ [45], which assumed a sequential addition of carriers or SCells. Leveraging this observation, OPCM restricts cell combinations in C to those explicitly allowed by the operator, reducing the cell set size to under 9 in 95% of cases.

4.2 RAN Policy Compliance via Cell Set Pruning

A BS-side vantage point enables OPCM to respect Radio Access Network (RAN) operator policies. We next describe how the RAN policies are defined, calculated, and realized.

Defining RAN Policies: OPCM uses a simple framework to encode RAN policies. A δ-constraint $\delta^{\mathcal{P}}_{c_i}$ defines an operator's tolerance of dissatisfying a well-defined policy $\mathcal{P}$. For instance, $\delta^{FI}_{c_i}$ implies that if we were to move a UE to cell combination c_i, the fairness index FI could go down by at most δ from the best possible value $FI^{max}_{c_i}$. Operators can flexibly define policies over the scope of a UE (*e.g.,* a UE cannot access high-capacity mmWave cells), a cell (*e.g.,* CA channel priorities), or the whole BS, and define their calculation rules. Table 3 exemplifies two such policies. (i) *User Fairness* calculates the fairness index (FI) of the long-term average throughput ($\mathcal{T}^t_u$) among users ($\mathcal{U}$). (ii) *Load Balancing* measures how evenly the load should be distributed across cell combinations. The load distribution index (LDI) computes the degree of similarity of load (resource blocks $\mathcal{R}^t_c$) for C.

Table 3. OPCM RAN objective examples.

RAN Policy	$\mathcal{P}$	$\mathcal{P}^{max}$
User Fairness	$FI^t = \frac{(\sum_{u\in\mathcal{U}} \mathcal{T}^t_u)^2}{\lvert\mathcal{U}\rvert \sum_{u\in\mathcal{U}} (\mathcal{T}^t_u)^2}$	1.0
Band Load Balancing	$LDI^t = \frac{(\sum_{c\in C} \mathcal{R}^t_c)^2}{\lvert C\rvert \sum_{c\in C} (\mathcal{R}^t_c)^2}$	1.0

Realizing Policies: Given $\delta^{\mathcal{P}}_{c_i}$, OPCM evaluates a "what-if" scenario of moving to cell combination c_i

in the next time step, *i.e.,* calculating $\mathcal{P}_{c_i}^{t+1}$ if OPCM were to move the UE to c_i. Lastly, OPCM checks if the what-if scenario satisfies the constraint, *i.e.,* $\mathcal{P}_{c_i}^{max} - \mathcal{P}_{c_i}^{t+1} < \delta_{c_i}^{\mathcal{P}}$. If not, c_i is removed from the UE's available set C so c_i will not be considered in the subsequent decision-making stage.

4.3 Opportunistic Decision Making

Once OPCM prunes the cell set C and obtains performance estimated from *Profiling Engine* (§5) for each $c_i \in C$, the *Decision Framework* decides whether to retain the current cell combination or switch to another. The key challenge lies in balancing the *exploration vs. exploitation* tradeoff (**C3**). Exploitation means maximizing the immediate performance by using the best cell combination based on current estimates, while exploration updates current estimates of cell combinations' performance. Exploration may incur performance loss when the UE utilizes a suboptimal cell combination, or increase battery drain when done aggressively. However, not doing so leads to inaccurate performance estimates, hurting the efficacy of future exploitations. Notably, *legacy* radio link quality–based CM does not require exploration, as UEs can passively measure link quality of other cell combinations via radio resource measurements (❷ in Fig. 5).

Balancing Exploration and Exploitation. The above problem can be formulated as a non-stationary multi-armed bandit problem, *i.e.,* one where the reward distributions associated with each arm (representing a cell combination) can change over time. A common approach to solving such problems is the epsilon-greedy policy, where we explore with probability ϵ and exploit our current estimates with probability $1 - \epsilon$. Lines 1-2 in Algorithm 1 highlight this policy. OPCM employs exponential epsilon decay to balance exploration and exploitation over time. The idea is to start with a high exploration rate (ϵ) and gradually decrease it exponentially as our performance estimates get better. λ is the decay rate controlling how fast ϵ decreases. Using empirical analysis (details in Appendix B), we find that OPCM performs best when λ is close to $|C|/2I$, where I is the best cell combination change interval and it can be easily computed on the BS.

Gain-aware Greedy Exploitation. OPCM simply keeps using the cell combination with the highest performance according to our estimates $\mu_{c_i}^t$ (Line 3 in Algorithm 1). However, simply choosing the highest performing cell combination may not lead to better performance if the performance gains are small and data-plane interruption is high (**C4**). Therefore, OPCM selects a new cell combination $\hat{c}^{t+1}$ during exploitation only if its performance is greater than the current cell combination's performance $\mu_{\hat{c}^t}^t$ plus a small hysteresis H (Line 4-5). A small hysteresis, inspired by the hysteresis used in UE measurement reporting, ensures that greedy exploitation would not lead to performance loss. H is easily configurable (we use $H = 1/2 * sqrt(\mu_{\hat{c}^t}^t)$ in our evaluations).

Smart-random Exploration. As discussed in §3, OPCM explores suboptimal cell combinations to update our estimates in anticipation that their performance might have improved. If OPCM decides to explore, we do not simply choose any random cell combination c_i. Instead, the question of which c_i to explore depends on two rules: (i) avoid frequently exploring a c_i with low historical performance, and (ii) explore the ones that have not been used for a long time so that OPCM can update their estimates. On the basis of these rules, we combine normalized *estimated performance* ($\mu_{c_i}^t$) and *time since last usage* ($lu_{c_i}^t$) in equal proportions to calculate the *exploration weight* ($w_{c_i}^t$) for each $c_i \in C$ (Line 7). Lastly, we randomly select a cell combination based on these *exploration weights* (Line 8). This enables OPCM to traverse all available cell combinations (in a stochastic manner) based on the balanced priority between each cell combination's historical performance and its measurement staleness. If a new cell combination appears in C, its *exploration weight* will be higher than other cell combinations and it will be prioritized during exploration.

Algorithm 1: Pseudocode for deciding next cell combination to use in OPCM (§4.3).

Output: $\hat{c}^{t+1}$ // cell combination for next time step

$\epsilon \leftarrow e^{-\lambda t}$
if $\mathcal{U}(0,1) > \epsilon$ **then** // exploitation
  $c^* \leftarrow \{\arg\max_i, \arg\min_i\}(\mu^t_{c_i})$
  if $\mu^t_{c^*} > \mu^t_{\hat{c}^t} + H$ **then**
    $\hat{c}^{t+1} \leftarrow c^*$
else // exploration
  $w^t_{c_i} \leftarrow \frac{\mu^t_{c_i}}{\sum_i \mu^t_{c_i}} + \frac{lu^t_{c_i}}{\sum_i lu^t_{c_i}}$
  $\hat{c}^{t+1} \leftarrow RandChoice([1 \cdots |C|], w^t_{c_i})$

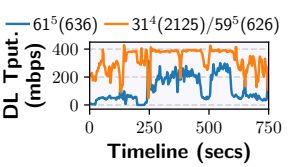


Fig. 6. An example of correlation b/w the performance of two cell combinations.

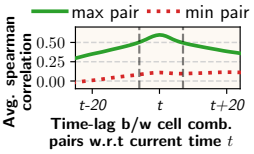


Fig. 7. Time Lagged Cross-Correlation (TLCC) across cell combinations.

5 Hybrid Profiling Engine

The *Profiling Engine* estimates the performance of cell combinations according to the configured criteria. A key challenge here is the absence of inactive cell combinations' network performance data (**C3**). The *Profiling Engine* addresses this via a novel passive approximation technique. This helps OPCM enhance the estimates obtained via exploration. Finally, the *Profiling Engine* produces the estimated performance metric $\mu^t_{c_i}$ for all $c_i \in C$. To compute $\mu^t_{c_i}$, the *Data Manager* collects the requisite raw data (*e.g.,* measured performance metric $m^t_{\hat{c}}$ of active cell combination $\hat{c}$ at time t).

5.1 Passive Performance Approximation

Our measurements suggest that the performance of (a subset of) cell combinations is often correlated. Therefore, historical correlations can help infer other cell combinations' performance. Recent works [18, 45, 48, 69] use cross-band link quality estimation on the physical layer. In contrast, we identify an opportunity where some cell combinations have correlated network performance.

To illustrate the correlation among cell combinations' network throughput, we present a representative trace in Fig. 6, which shows synchrony in the performance of two cell combinations. In particular, the troughs and stable performance regions are mostly correlated. This phenomenon is attributed to three reasons. First, cell combinations often share the same cells (or CCs) and radio infrastructure, which leads to correlated performance. Second, a blockage can cause sudden performance degradation for multiple cell combinations during mobility. Third, temporary user congestion can affect the BS-side radio resource scheduling and, henceforth, the throughput for more than one cell combination.

Inspired by the above observation, we use our large-scale dataset from §2.4 to calculate cross-correlations across cell combinations. We employ Time Lagged Cross-Correlation (TLCC) [62] to determine the lag period during which the performance is correlated. The rationale for TLCC is that a UE can use only one cell combination at any moment, so there is a time lag between when the data of two cell combinations is collected. In our analysis, we identify the pairs with the maximum and minimum (lagged) correlation at any given time and plot their averages. Fig. 7 shows that many cell combinations indeed exhibit non-trivial cross-correlation. Usually, the correlation is decent (> 0.5) within a short time window [t-5, t+5] secs, suggesting we use recent measurements of the active cell combination to improve the correlated cell combination's network estimates. Additionally, network latency and energy efficiency exhibited similar correlation patterns during our investigations.

$$m^t_{c_i} = \mu^t_{c_i} * [1 + Corr^t_{\hat{c},c_i} * (m^t_{\hat{c}} - m^{t-1}_{\hat{c}})/m^{t-1}_{\hat{c}}] \quad (1)$$

We detail the procedure of using performance measurements of the active cell combination $\hat{c}^t$ to improve the estimates of an inactive cell combination c^t_i. First, we calculate their TLCC to be $Corr^t_{\hat{c},c_i}$ and the lag to be δt. Let $a(t)$ and $b(t)$ be the measured throughput sample series of cell combination a (active) and b (inactive), respectively. $Corr^t_{a,b}$ is calculated as the maximum correlation between $a(t - \delta t)$ and $b(t)$ regarding δt, the lag. We empirically use a 2-minute worth of samples when calculating $Corr^t_{a,b}$. Next, if the TLCC is significant in a small lag period (empirically determined if

$Corr^t_{\hat{c},c_i} > 0.5$ and $\delta t < 5$ secs), we estimate the performance of c_i using Eqn. 1, which applies the (measured) gradient of $\hat{c}$'s performance to the (approximated) c_i's performance. The gradient is weighted by $Corr^t_{\hat{c},c_i}$ as our confidence level of the similarity between the two cell combinations' performance trends. In cases when TLCC-based correlation is weaker than 0.5 (*e.g.,* high mobility, or different BSs), OPCM uses c_i's most-recent estimate to avoid inaccurate predictions.

5.2 OPCM Performance Criteria

Lightweight Performance Estimation. Unlike applications requiring precise network performance estimates (*e.g.,* video streaming), OPCM only requires the comparative performance (ranking) of cell combinations. Since its goal is to perform an argmax operation, the system relies on the ordinal properties (ranking) of the estimates rather than their precise cardinal values. We thus adopt a highly lightweight filter-based approach with reasonable accuracy, without resorting to sophisticated approaches. Eqn. 2 shows how we use the traditional Exponential Weighted Moving Average (EWMA) to compute the performance estimate $\mu^t_{c_i}$. Through large-scale trace-driven simulations, we find that any alpha value in the range $\alpha \in [0.5, 0.8]$ works decently well (§8.3). Exploring more sophisticated prediction methods is left to future work.

$$\mu^t_{c_i} = \alpha * m^t_{c_i} + (1 - \alpha) * \mu^{t-1}_{c_i} \tag{2}$$

Criteria Support. OPCM can support diverse performance criteria and metrics. These metrics are transparent to applications and can be easily computed at the BS. By default, we use 5G QoS Identifier (5QI) to identify the performance criterion for a flow. Here, we list down four diverse criteria implemented in OPCM so far. ***(i)*** *Radio link quality* is the legacy criterion in cellular networks. OPCM can directly use the link-quality reports (❸ in Fig. 5) sent by UEs to the BS. ***(ii)*** *Network throughput* is a popular choice of performance criterion for CM. OPCM collects Radio Link Control (RLC)-layer throughput measurements to calculate network throughput. ***(iii)*** *Network latency* is the head-of-line delay experienced by packets in RLC queues. It represents the time spent by the first Service Data Unit (SDU) packet in the RLC queue. ***(iv)*** *UE energy efficiency* criterion is supported via offline energy models or preference lists, which can be constructed using measurements from external power monitor (§2.2) or Android's On Device Power Rails Monitor (ODPM) tool [17, 31]. ODPM allows segmentation of power consumption by hardware components, including the cellular modem. Offline models predict energy consumption based on UE's network throughput, cell combination, and device model [56]. Preference lists simply rank cell combinations by their mean energy efficiencies and are easier to construct but less robust than full models.

6 Robust Execution Module

Once the *Execution Module* receives the CM decision ($\hat{c}^{t+1}$) from the *Decision Framework*, it first checks UE's Radio Resource Configuration (RRC) state [12] and current cell combination ($\hat{c}^t$). Based on this, OPCM triggers the appropriate CM procedure (cell re/selection, handover, CA) to change UE's cell combination. The execution command is sent as an RRC reconfiguration message to the UE. OPCM also enables *two* mechanisms that improve its robustness and efficiency.

(i) Delayed Reconfiguration. Recall that CM procedures incur data-plane interruptions (**C4**). A close examination reveals that the interruption is high when data is waiting in the uplink or downlink transmission queues[1] while the BS executes the CM procedure. This observation suggests an optimization opportunity: OPCM can delay executing a CM procedure by up to ρ in anticipation that queues will drain up within the next ρ time units, thus minimizing interruptions.

[1]The uplink (downlink) data waits in UE-side (BS-side) queues before the BS assigns radio resources for data transmission. The BS does not assign resources to a UE during CM, leading to interruptions.

Inspired by the above, OPCM initializes a timer (counting down from ρ) and keeps monitoring the uplink/downlink queues. The downlink queue resides on the BS side so it can be directly monitored; meanwhile, UEs periodically send their uplink queue information (buffer status report [4]) to the BS. When the queues are empty, OPCM stops the timer and triggers the CM procedure. Otherwise, it exponentially increases UE's scheduling priority, attempting to drain the queues quickly. When the timer expires, OPCM executes the CM procedure regardless of queues' status. The selection of ρ incurs a tradeoff: if ρ is too short, there may not be enough time to drain the queues, whereas a large ρ may lead to inaccurate CM execution timing. To balance this tradeoff, we empirically set ρ to the median data-plane interruption (120 ms) due to CM procedures (§2.4). The exponential priority scheduling has a negligible impact on fairness since CM is infrequent and ρ is short.

(ii) Fallback Mechanism. OPCM includes an optimization to save BS computation and energy. When the UE experiences no data activity (idle state) or network activity is extremely low (*e.g.,* $\mu_{c_i}^t$ < 1 Mbps), OPCM temporarily falls back to the legacy radio link quality-based criterion and only passively listens to the traffic for any network activity.

7 Implementation

OPCM Prototype. OPCM is built on top of srsRAN [64, 65], an open-source 4G/5G software defined radio suite. We modified the cellular protocol stack (4G/5G Layer 2) in srsRAN to implement OPCM in over 6.1K lines of C/C++ code. First, we implemented necessary logging functionality for RLC, MAC and PHY layers to support data collection. Further, we developed a modular CM engine atop the RRC layer of the protocol stack. The engine abstracts the CM procedures inside RRC layer and adds support for diverse performance criteria. The *Data Manager* in OPCM collects logs (*e.g.,* UE measurement reports and performance metrics) at (configurable) periodic intervals. These logs are forwarded to: (i) the *Profiling Engine* that figures out the performance criterion using UE's 5QI information, and estimates the performance, and (ii) the *Decision Framework* that maintains the cell set C and makes strategic CM decisions. The system time step length ($\Delta t = t - (t-1)$) is 1 secs. Finally, the *Execution Module* receives CM decisions from the *Decision Framework* and triggers the appropriate CM procedure. To do so, it observes the UE state and the difference between cell combination under usage and the one we are switching to. Based on this, the *Execution Module* figures out which cells to add, remove, or modify. Finally, it sends the RRC reconfiguration message to the UE to initiate the CM procedure. Since the BS can assign absolute priorities to cells (Table 6), we make use of this 3GPP-defined feature to configure the desired cell combination on the UE. The *Execution Module* also oversees the delayed reconfiguration module and the fallback mechanism.

Custom Metric Registration. OPCM supports custom metrics through a lightweight C/C++ API. As shown in Fig. 8, metrics are registered via `register_metric()` with a name, callback, and interval. The *Profiling Engine* invokes each callback with raw RLC/MAC/PHY stats for all UE-cell combinations. Returned values are normalized and used by *Decision Framework*, enabling easy extensibility without altering core logic. This modular interface reduces the burden of manually configuring performance rules, directly addressing the CM management complexity (§1).

Trace-driven Simulator. We developed a 4G/5G network simulator based on the *ns-3* LTE and NR codebase [9, 57]. OPCM's simulator proof-of-concept is similar to our prototype implementation. Additionally, we integrated trace-driven channel simulations and implemented traffic generator for a file transfer application. Overall, we added or modified 4.9K+ lines of C/C++ and Python code.

8 Evaluation

We first build an in-lab end-to-end cellular network, given the lack of operator support and high cost of commercial BS deployment. Despite its limited scale, it provides a high physical-layer fidelity

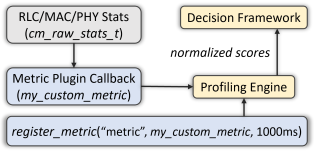


Fig. 8. Custom metric registration flow in OPCM.

Fig. 9. The over-the-air prototype testbed of OPCM.

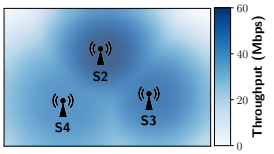


Fig. 10. DL tput. density in 60m×40m test area.

through real hardware and channel interaction. This is complemented by large-scale simulations to stress-test OPCM under a high density of users and cells, thereby ensuring reproducibility.

8.1 Experimental Setup

(i) Over-the-air Testbed. The setup is shown in Fig. 9. Each BS has two components: (i) a srsRAN-based eNodeB or gNodeB stack running on a laptop equipped with Intel Core i7 @ 3.00GHz CPU, and (ii) an RF frontend based on USRP B210 [68] software defined radio (SDR). The Open5GS Core Network (CN) [11] runs on a desktop machine. All apps are hosted locally with a 20 ms delay between the CN (packet gateway) and the remote server. We use ADB scripts to automate and time-synchronize experiments. All experiments are repeated at least 5×.

Experiment Settings. We set up three cell combinations on four B210s. NSA-5G needs two RF frontends, one for 4G and one for 5G. The frequencies for these cell combinations are according to **S2-4** in Table 2. Each one uses 20 MHz bandwidth with 64 QAM and 256 QAM MCS tables for uplink and downlink, respectively. All BSs use the proportional fair scheduler with default srsRAN parameters. We put the testbed (Fig. 9) in an empty open parking garage to eliminate environmental noise. Fig. 10 plots a density map showing maximum downlink throughput at any location. We freely walk at ∼3 mph during our experiments with UEs in hand.

Comparative Approaches. (i) The legacy approach uses the radio link quality-based CM to choose UE's cell combination. We use srsRAN's default criterion parameters. **(ii)** iCellSpeed [24] is a UE-side solution to increase UE's network throughput. For a fair comparison, we implement a network-side version of iCellSpeed. We modify its *iCustomize* module since CM procedures can be directly triggered from the BS. The *iprofile* module is set up as described in the paper.

COTS & Virtual UEs. We use a Google Pixel 7 (PX7) smartphone and apply a programmable sim card to register it with CN. The PX7 phone lacks the ability to configure 5QI, therefore, we fix the 5QI in OPCM and test one application use case at a time. For simulation experiments, each app sets up its data bearers with the appropriate 5QI value. We also use 30 (10 for each BS) ZeroMQ srsUEs [14] with virtual radios. These virtual UEs utilize real-world network traces to generate network traffic and channel traces to model realistic channel conditions.

Network and Channel Traces. We collect these traces with NG-Scope [73] and post-process them to match our BSs' numerologies (*e.g.,* cell bandwidth). We scale up/down traces to increase/decrease the BS load at the start of an experiment (call it *sload*). Note that the BS load can change during the experiment if a CM procedure transfers the UE from one cell to another. The average *sload* is set to be 67% unless otherwise stated. Additionally, we configure srsRAN to utilize real-world channel quality traces collected with XCAL. Our 14 hrs+ trace *corpus* is a 4-dimensional tensor (trace #,cell combination, channel, time), where each entry corresponds to a wideband Channel Quality Indicator (CQI) value. We chop these traces across time, with each trace spanning 350 secs. We randomly select 10 traces (for 10 virtual UEs) from the *corpus* for each BS. Since we have collected these traces in different mobility scenarios, the heterogeneity and randomization ensure that each BS has UEs with diverse channel conditions.

RAN Objectives. When OPCM is not running, the fairness index and load balancing index is in the range of 0.85–0.95 for our testbed (see Fig. 17). Therefore, we set OPCM delta tolerance (δ_b^O)

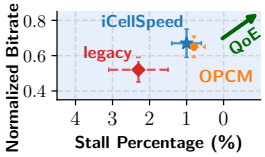


Fig. 11. Comparing OPCM VoD streaming performance across baselines.

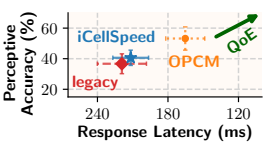


Fig. 12. Comp. OPCM video analytics performance across baselines.

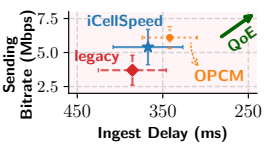


Fig. 13. Comp. OPCM video ingest performance across baselines.

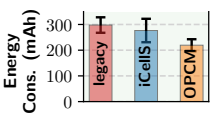


Fig. 14. Comparing OPCM energy efficiency across baselines.

for fairness and load balancing constraints to the higher limit 0.15 (*i.e.,* $1 - 0.85 = 0.15$). Later, we investigate the impact of δ on OPCM performance (§8.5).

(ii) Trace-driven Simulations. We use ns-3 [9, 57] to test OPCM with advanced 4G/5G settings and large number of users. The setup is identical to the over-the-air testbed, with a few exceptions. We increase the number of cell combinations to five by adding two new settings: **S1** in Table 2 and a 5G cell operating at 2155 MHz. Each BS can now aggregate up to 4 carriers, increasing the bandwidth from 20 MHz to 100 MHz. 150 UEs (30 for each BS) repeatedly download a 256 MB file from the remote server for 8 mins. Realistic channel conditions are still modeled with our *corpus* of traces.

8.2 OPCM QoE Improvement

To evaluate OPCM under our testbed, we develop a suite of four mobile apps with *diverse* workloads.

(i) VoD Streaming. Our VoD streaming experiments use a dash.js [28] player to stream a 4 min video. We mainly test buffer-based BOLA [63] and rate-based [49] adaptive bitrate (ABR) algorithms due to their popularity. The video is encoded at 6 unique quality levels (0.8-6.8 Mbps average bitrate). Fig. 11 plots the normalized bitrate and stall percentage for our experiments. The QoE improves in the top right direction as indicated by the arrow. The results show that, compared to the legacy, OPCM improves the average bitrate by 25.1%. Similarly, it reduces the average video stall percentage by 65.2%. iCellSpeed offers slightly (3.1%) higher bitrate than OPCM, but also has a 0.2% higher absolute stall rate. iCellSpeed performs well for downlink throughput-hungry applications but can lose performance for other application types and has a high memory footprint (details to follow).

(ii) Latency-critical Video Analytics. We select a popular video analytics task: Object detection (OD). OD app uses a state-of-the-art video analytics model (*i.e.,* YOLOv4 [19]) deployed locally. Instead of sending camera feeds, both phones stream the same video frames from the COCO dataset [6] at 30 FPS. The perceptive accuracy (defined in [30, 43]) captures mean average precision for sending frames, and replaces a frame's inference with the last feedback if a response is not received within 200 ms. Fig. 12 showcases that OPCM achieves 23.7% higher perceptive accuracy and 28.1% lower response latency than iCellSpeed on average. The performance difference can be attributed to two reasons: *(i)* iCellSpeed focuses on improving the throughput only while OPCM optimizes the performance criterion inferred from 5QI, and *(ii)* OPCM's queuing-aware delayed reconfiguration mechanism minimizes data-plane interruptions.

(iii) Uplink Video Ingest. We re-purpose Ant-Media's LiveVideoBroadcaster [8] to publish a pre-recorded video stream (1080p @ 30 FPS with 7.2 Mbps average bitrate). UEs send adaptive RTMP feeds [13] to a media server [5] deployed locally. We plot the sending bitrate and ingest delay for published video streams. Ingest delay, as defined in [80], is the time from when a video frame is generated at the source to when its quality variants are available at the server for client download. Fig. 13 depicts that OPCM achieves 12.9% higher average bitrate compared to iCellSpeed. Recall that uplink and downlink may have different highest performing cell combinations (§2.4). OPCM can infer which direction to optimize unlike iCellSpeed that only optimizes the downlink throughput. Moreover, the uplink-intensive video ingest sees comparatively higher improvement than other downlink-heavy apps because the gap between legacy and the highest performing cell combination is wider for uplink (§2.4).

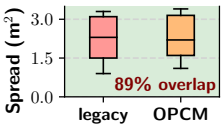


Fig. 15. Benchmarking OPCM performance.

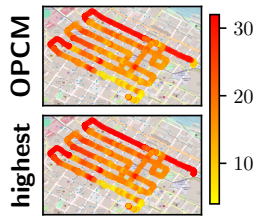


Fig. 16. Evaluating OPCM compliance with legacy link quality criterion.

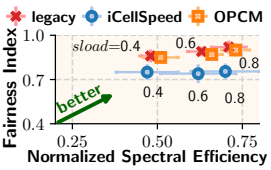


Fig. 17. Comparing RAN metrics across various load conditions.

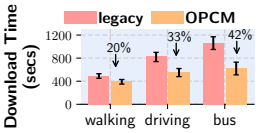


Fig. 18. OPCM performance under different mobility scenarios.

(iv) Lightweight Fitness Tracker. We build an app to transmit live health data (*e.g.,* heart rate) to the cloud. The app sends ~100B messages every 0.25 ms. In addition, we configure OPCM to use energy consumption models and optimize PX7's energy efficiency (§5.2). Fig. 14 plots the overall energy consumed by the UE during a 30 mins experiment. It shows that OPCM improves the average energy efficiency by up to 26.3% and 20.6% over legacy and iCellSpeed, respectively. While 3GPP has not defined energy-specific 5QI values, Releases 16 and 17 introduce UE Assistance Information (UAI), enabling UEs to share energy-related preferences with the BS [15]. This, along with our evaluation, highlights the potential of energy-aware CM for applications like IoT.

8.3 OPCM Benchmarking

OPCM is backward-compatible, and performs as well as the legacy CM under mobility. To test if OPCM achieves its desired goals, we use radio link quality as the performance criterion inside OPCM and compare it with legacy CM. The goal is to see if OPCM triggers CM procedures exactly the same way as legacy CM does if the CM parameters (hysteresis, Time-to-Trigger, *etc.*) are same. We only run ping on the PX7 phone to keep the UE radio in active state. We split the parking garage area (Fig. 10) into 6×4 lanes, mark the lanes with a tape, and walk on the tape vertically and horizontally to ensure reproducibility. We use WifiRttLocator [72] to position the UE with 1m accuracy. Overall, we collect 3 hrs+ of data with at least 65+ CM procedures triggered (mostly handovers and DC) for each setting (OPCM and legacy).

Given same mobility patterns, the CM procedures must be triggered at almost the same spot for both legacy CM and OPCM. Fig. 15 plots the spread of the areas, where CM procedures are triggered repeatedly. To get this spread, we compute the convex hull for each spot where legacy CM procedures are triggered, and use that as a reference. We also calculate the overlap percentage of legacy's spread with OPCM's spread. A high mean overlap value of 89.2% shows that OPCM indeed works like legacy. We also conduct benchmarking simulation experiments, ensuring full reproducibility. The results reinforce our over-the-air testbed findings, *i.e.,* OPCM triggers CM procedures at the same time and location as the legacy CM for radio link quality criterion.

OPCM's epsilon-greedy policy works effectively in the real world. To evaluate if OPCM's epsilon-greedy exploration can efficiently find the highest performing cell combination, we run a barebone version of our system on S22+ and test it under the same 740m × 510m rectangular loop as Fig. 4. We pre-configure the cell set C, turn off cell set pruning and delayed reconfiguration, and use a special code (∗#2263#) to switch cell combinations. Fig. 16 plots live video streaming's bitrate for OPCM (top plot) and compares it with the highest performance achieved by any of the other cell combinations **S1-4** (bottom plot). The results show that OPCM operates close to the highest performing setting: the median bitrate gap between the two is only 2.4 Mbps (10.1%). A small gap is because OPCM's epsilon-greedy policy starts with zero knowledge of cell combinations' performance. In comparison, the legacy CM had a median bitrate gap of 70.1% with the highest performing setting (see Fig. 4b).

8.4 End-to-end System Evaluation

OPCM satisfies all RAN objectives. Here, our setup only utilizes virtual UEs with *sload* between 40-80%. We plot user fairness (defined in §4.2) and spectral efficiency (bit/s/Hz), which indicates the amount of information sent through a network using the available bandwidth. We normalize spectral efficiency by dividing it with the highest value of the respective cell. Fig. 17 yields two key takeaways. ***First***, since iCellSpeed does not respect RAN objectives, it costs 12.7-17.9% in terms of fairness. Moreover, it improves the spectral efficiency of high-bandwidth cell combinations, while the efficiency of others degrades (see high variation of iCellSpeed's spectral efficiency). In other words, iCellSpeed overloads some cells while others are underused. ***Second***, OPCM's fairness is within 98-99% of legacy. In addition, OPCM improves spectral efficiency by 2-3% compared to the legacy. Although not shown, the load distribution index of OPCM is 0.91-0.94 for different *sload* values.

OPCM is particularly useful under high mobility. Remember that each virtual UE's channel trace (§8.1) belongs to a specific mobility scenario (*e.g.,* walking, driving, bus). To compare OPCM gains across mobility scenarios, we configure all virtual UEs to repeatedly download a 256 MB file from the remote server for 8 mins. To plot results, we form UE groups based on the mobility scenario of UEs' channel traces. Fig. 18 shows the average file download time of UE groups. Compared to the legacy case, OPCM reduces the average file download time for *bus* UEs by 41.5%. In contrast, *walking* UEs only see 20.4% reduction. When active cell combination's performance fluctuates rapidly (*i.e., bus*) and the throughput gap between the active and the highest performing cell combination widens, OPCM is more likely to select a better cell combination at the next time step due to its greedy policy (§4.3). This leads to higher OPCM gains under complex mobility scenarios.

OPCM is scalable and involves lightweight communication and system costs. ***First***, we evaluate OPCM under large-scale users. Fig. 19 plots the file download time of a 256 MB file as the total number of users (virtual UEs) grows. As users increase, the download time gradually increases due to the limited bandwidth. However, the increase is almost linear, and the standard deviations are small, suggesting that OPCM offers application-level fairness in the presence of multi-user competition. ***Second***, we record OPCM's CPU and memory consumption in Table 4. It shows that compared to legacy, OPCM increases CPU and memory utilization by 6.5% and 3.1%, respectively. Although not shown, the CPU utilization only increases slightly (2.2%) when users go from 30 to 90. Although not a fundamental limitation, iCellSpeed incurs higher memory usage than OPCM because its *iProfile* module tracks more per-UE state, reflecting both the frequency and performance of cell choices. ***Third***, OPCM slightly increases the signaling overhead between UE and BS due to exploration. In the average case, the number of signaling messages increases by 11.6% (from 68 to 76 per minute) compared to the setup where exploration is disabled on our testbed. Similar to legacy CM, this overhead is proportional to UE mobility: faster-moving UEs experience quicker channel variations and more frequent changes in the best cell combination, leading to higher signaling rates.

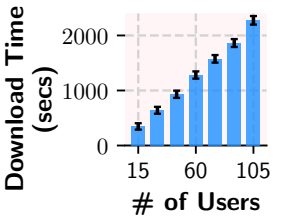


Fig. 19. OPCM scalability as # of users increase.

OPCM efficiently manages advanced CA/DC settings. Using large-scale ns-3 simulations, we evaluate OPCM with advanced 4G/5G numerologies and variable number of cell combinations. While not shown, our results yield two insights: *(i)* despite five cell combinations with advanced CA and DC settings, OPCM maintains fairness within the δ_b^{FI} range ($FI > 0.85$), evenly distributing load and achieving >90% normalized spectral efficiency across all cells; and *(ii)* increasing the number of cell combinations has diminishing returns, *e.g.,* increasing cell combinations from 3 to 5 only improves average file download time by 9.1%. For comparison, moving from one to three cell combinations significantly enhances download time by 26.0%.

Table 4. Comparing system overhead (30 users).

Metric	legacy	iCellSpeed	OPCM
CPU Utilization (%)	35.9 ± 4.8	41.1 ± 5.4	42.4 ± 5.6
Memory Utilization (%)	17.8 ± 3.1	33.7 ± 4.1	20.9 ± 3.6

Table 5. OPCM Profiling Engine *vs.* Baselines.

Metric	OPCM	OPCM w/o TLCC	iCellS	iCellS w/ TLCC
RMSE	0.15 ± 0.04	0.33 ± 0.06	0.25 ± 0.06	0.13 ± 0.03
MAE	0.12 ± 0.03	0.24 ± 0.06	0.18 ± 0.03	0.10 ± 0.02
RA	0.94 ± 0.02	0.74 ± 0.06	0.82 ± 0.04	0.94 ± 0.03

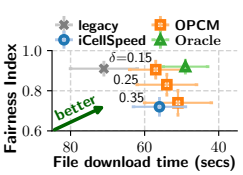


Fig. 20. Decision framework vs. Oracle.

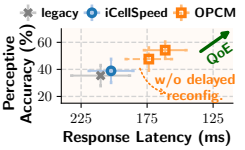


Fig. 21. Comparing OPCM execution module with legacy.

8.5 Micro-benchmarks

Decision Framework vs. Oracle. We use ns-3 simulations to compare OPCM *Decision Framework* with the *Oracle*, which has the complete knowledge of cell combinations' performance and does not require exploration. It thus represents a performance upper-bound. We turn off *Profiling Engine*, disable the delayed reconfiguration mechanism, and use ground-truth performance predictions. Fig. 20 yields three main insights. *(i)* OPCM's epsilon-greedy policy effectively balances the exploration and exploitation tradeoff. OPCM is within 98.4% and 83.7% of the *Oracle* in terms of fairness and file download time, respectively. This gap can be attributed to exploration, especially when OPCM starts building performance estimates (§8.3). *(ii)* The objective-aware *Decision Framework* offers adjustable knobs (δ) to balance the tradeoff between application performance and RAN metrics. *(iii)* When RAN polices are more tolerant (*i.e.,* large δ), OPCM behaves similar to iCellSpeed.

Data-plane Interruption Reduction in OPCM. Our results in Fig. 21 show that OPCM achieves 7.7% lower average response latency and 12.2% higher perceptive accuracy compared to the case when queuing-aware delayed reconfiguration is disabled on OPCM. It drains the uplink/downlink queues before triggering a CM procedure, reducing the queuing delay for video frames. This ultimately leads to a lower response latency and better perceptive accuracy.

OPCM Profiling Engine vs. Baselines. Using the full might of our *corpus*, we compare OPCM with three baselines: (i) OPCM *Profiling Engine* without time-lagged cross-correlation (TLCC), (ii) iCellSpeed, and (iii) iCellSpeed with TLCC. The root mean square error (RMSE) and mean absolute error (MAE) between estimated and ground-truth throughput is normalized by each UE's mean throughput. Apart from that, we also compute the ranking accuracy (RA) that measures how accurately an approach predicts the highest performing cell combination. Table 5 presents the summary of our results. There are four main takeaways: *(i)* performance estimation can tolerate small errors since we only need to know the comparative performance (ranking) of cell combinations. Notice that although the RMSE is 0.13-0.33 depending on the approach, RA is high (0.94-0.74); *(ii)* TLCC across cell combinations improves RA (even for iCellSpeed) by up to 27.0%; *(iii)* OPCM achieves 14.6% higher RA than iCellSpeed on average. This is because OPCM can utilize passive approximation to build more accurate performance estimates; and *(iv)* adding TLCC to iCellSpeed makes it achieve slightly (13.3%) lower RMSE than OPCM because of its *iProfile* module that builds more precise profiles but also results in higher memory usage (Table 4).

9 Related Work

Cell Re/selection and Handovers. Several prior studies have sought to enhance cell selection and reselection practices to achieve faster access speeds [24, 25, 44]. For instance, iCellSpeed [24] adopts proactive, device-side assistance to facilitate network-controlled cell selection, thereby improving network throughput. Similarly, Li *et al.* [44] propose reconfiguring operator-defined cell selection parameters to optimize network throughput dynamically. Many 4G/5G studies highlight suboptimal access speeds due to uncoordinated and inefficient mobility management practices and suggest

ways to unlock untapped potentials [25, 34, 46, 51, 58, 79]. Several prior studies devise techniques to distribute user load across cells under practical constraints [20, 36, 76].

Our work distinguishes itself in several key aspects. First, while previous approaches primarily focus on a single performance metric, mainly throughput, for a specific CM procedure, we advocate for a full reimagination of CM by *decoupling* performance metrics from individual procedures. Second, the existing solutions, particularly those implemented on the UE side, lack support for meeting operator policies. Third, when comparing the performance of different cells, our solution can reduce the overhead of online profiling via passive performance estimations.

Carrier Aggregation and Multi-operator Support. CA is a critical technology for boosting network capacity in 4G/5G networks. Wei *et al.* [77] analyze the deployment of CA in public 5G networks, shedding light on its operational characteristics and challenges. Some research focuses on radio access failures when adding secondary cells [50, 52]. To address limitations of CA's sequential, cell-by-cell operations, CA++ introduces novel algorithms for group-based CA [45]. However, our investigation reveals that operators already implement group-wise CA in practice (§4.1). Additionally, adaptive approaches have been proposed to enable efficient multi-carrier access for mobile devices [41, 47]. Our approach is complementary to these existing solutions and can be integrated with them to further enhance system performance.

Measurement Studies. Many recent studies investigate radio characteristics, network performance and energy efficiency of 5G deployments [54–56, 74]. Yang *et al.* [75] reveal how reusing 4G bands for newer 5G deployments reduces 4G performance. Hassan *et al.* [34] uncover mobility management overheads for 5G and contrast them with 4G. Some earlier studies [24, 42, 78] explore the performance of 3G/4G bands under limited settings. We go beyond these efforts by characterizing the wide availability and diversity of 4G/5G cell deployments, the interplay among them, and the limitations of operators' default CM strategies, to name a few.

10 Discussion & Conclusion

Deployment Scope. OPCM's gains are proportional to the diversity of available cells, making it most effective in dense urban deployments. In suburban or rural areas with fewer alternatives, OPCM converges to *legacy* CM decisions without added cost. The framework also adapts to permanent changes: new cells are incorporated as soon as UEs report them, while cell failures are handled similarly to existing CM mechanisms.

Mobility and Coordination. When a UE moves to a legacy BS, OPCM's backward compatibility ensures a seamless fallback to *legacy* CM. Performance can be improved by sharing profiles across BSs via the X2/Xn interface, while integration with O-RAN Radio Intelligent Controllers (RICs) [10] offers a natural future extension for richer coordination.

OPCM Time Step. OPCM uses a 1-second time step Δt, balancing responsiveness against signaling and computation overhead. Δt is operator-configurable: smaller values enable faster adaptation, while larger values reduce cost. Adaptive time-stepping remains an avenue for future work.

In summary, we reveal a new optimization dimension in the 5G/NextG ecosystem – performance-driven CM. Our multi-country measurement showcases the wide availability and heterogeneity of 4G/5G cell deployments, as well as the missed performance. OPCM demonstrates the feasibility of building intelligent CM services for 5G/xG networks.

Acknowledgments

This research is supported in part by the National Science Foundation (NSF) under grants 2106771, 2128489, 2212318, 2220286, 2220292, 2321531, 2321532, 2323174, 2333489, 2402991, 2409269, 2411625, and 2112562 (National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks), as well as a Cisco University Research grant.

References

[1] 2021. *NR and NG-RAN Overall description.* Retrieved January 2025 from https://www.etsi.org/deliver/etsi_ts/138300_138399/138300/16.04.00_60/ts_138300v160400p.pdf

[2] 2023. *ffmpeg Streaming Documentation.* Retrieved January 2024 from http://trac.ffmpeg.org/wiki/StreamingGuide

[3] 2023. *Simple RTSP Server.* Retrieved January 2024 from https://github.com/aler9/rtsp-simple-server

[4] 2024. *5G MAC BSR – Buffer Status Reporting.* Retrieved January 2024 from https://www.techplayon.com/5g-mac-bsr-buffer-status-reporting/

[5] 2024. *Ant Media.* Retrieved January 2024 from https://antmedia.io/

[6] 2024. *Coco Dataset.* Retrieved January 2024 from https://cocodataset.org/#home

[7] 2024. *EU and China lagging behind in mmWave spectrum.* Retrieved January 2024 from https://5gobservatory.eu/eu-and-china-lagging-behind-in-mmwave-spectrum/

[8] 2024. *LiveVideoBroadcaster.* Retrieved January 2024 from https://github.com/ant-media/LiveVideoBroadcaster

[9] 2024. *NS-3 LTE Module.* Retrieved January 2024 from https://www.nsnam.org/docs/models/html/lte.html

[10] 2024. *O-RAN RIC user guide.* Retrieved January 2024 from https://docs.o-ran-sc.org/projects/o-ran-sc-ric-plt-submgr/en/metrics/user-guide.html

[11] 2024. *Open5GS.* Retrieved January 2024 from https://open5gs.org/

[12] 2024. *Procedures for the 5G System (5GS).* Retrieved January 2024 from https://www.etsi.org/deliver/etsi_ts/123500_123599/123502/15.05.01_60/ts_123502v150501p.pdf

[13] 2024. *Real-Time Messaging Protocol.* Retrieved January 2024 from https://en.wikipedia.org/wiki/Real-Time_Messaging_Protocol

[14] 2024. *srsRAN 4G with ZMQ Virtual Radios.* Retrieved January 2024 from https://docs.srsran.com/projects/4g/en/latest/app_notes/source/zeromq/source/index.html

[15] 3GPP. 2023. *NR; User Equipment (UE) radio access capabilities.* Technical Report TS 38.306. 3rd Generation Partnership Project (3GPP). https://www.etsi.org/deliver/etsi_ts/138300_138399/138306/17.01.00_60/ts_138306v170100p.pdf Release 17.

[16] Accuver. 2022. *Accuver XCAL.* Retrieved January 2024 from https://www.accuver.com/sub/products/view.php?idx=6

[17] Android Developers. 2024. Power Profiler - Android Studio. https://developer.android.com/studio/profile/power-profiler. Accessed: 2025-05-28.

[18] Arjun Bakshi, Yifan Mao, Kannan Srinivasan, and Srinivasan Parthasarathy. 2019. Fast and Efficient Cross Band Channel Prediction Using Machine Learning. In *The 25th Annual International Conference on Mobile Computing and Networking* (Los Cabos, Mexico) *(MobiCom '19)*. Association for Computing Machinery, New York, NY, USA, Article 37, 16 pages. doi:10.1145/3300061.3345438

[19] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[20] Pablo Caballero, Albert Banchs, Gustavo de Veciana, Xavier Costa-Perez, Pablo Caballero, Albert Banchs, Gustavo de Veciana, and Xavier Costa-Perez. 2017. Multi-Tenant Radio Access Network Slicing: Statistical Multiplexing of Spatial Loads. *IEEE/ACM Trans. Netw.* 25, 5 (2017). doi:10.1109/TNET.2017.2720668

[21] Yongzhou Chen, Ruihao Yao, Haitham Hassanieh, and Radhika Mittal. 2023. Channel-Aware 5G RAN Slicing with Customizable Schedulers. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, Boston, MA, 1767–1782. https://www.usenix.org/conference/nsdi23/presentation/chen-yongzhou

[22] Nokia Communications. 2023. 5G Carrier Aggregation explained. https://www.nokia.com/about-us/newsroom/articles/5g-carrier-aggregation-explained/#:~:text=Carrier%20Aggregation%20is%20a%20software,enhance%20the%20end%20user%20experience. Accessed: 2023-06-09.

[23] Android Developers community. 2023. *Android Debug Bridge (ADB).* Retrieved January 2024 from https://developer.android.com/studio/command-line/adb

[24] Haotian Deng, Qianru Li, Jingqi Huang, and Chunyi Peng. 2020. Icellspeed: Increasing cellular data speed with device-assisted cell selection. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[25] Haotian Deng, Chunyi Peng, Ans Fida, Jiayi Meng, and Y Charlie Hu. 2018. Mobility support in cellular networks: A measurement study on its configurations and implications. In *Proceedings of the Internet Measurement Conference 2018*. 147–160.

[26] Android Developers. 2023. *Media3 Exoplayer RTSP.* Retrieved January 2025 from https://developer.android.com/media/media3/exoplayer/rtsp

[27] Phuc Dinh, Moinak Ghoshal, Dimitrios Koutsonikolas, and Joerg Widmer. 2022. Demystifying Resource Allocation Policies in Operational 5G mmWave Networks. In *2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. 1–10. doi:10.1109/WoWMoM54355.2022.00016

[28] dsilhavy. 2023. *DASH.js framework.* Retrieved January 2024 from https://github.com/Dash-Industry-Forum/dash.js/

[29] Data Center Dynamics. 2025. *Syniverse blames US carrier roaming outage on "signaling storm".* Retrieved January 2025 from https://www.datacenterdynamics.com/en/news/syniverse-blames-us-carrier-roaming-outage-on-signaling-storm

[30] Ionel Gog, Sukrit Kalra, Peter Schafhalter, Matthew A. Wright, Joseph E. Gonzalez, and Ion Stoica. 2021. Pylot: A Modular Platform for Exploring Latency-Accuracy Tradeoffs in Autonomous Vehicles. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. doi:10.1109/icra48506.2021.9561747

[31] Agrim Gupta, Adel Heidari, Avyakta Kalipattapu, Ish Kumar Jain, and Dinesh Bharadia. 2024. 3 W's of smartphone power consumption: Who, Where and How much is draining my battery?. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 2248–2250. doi:10.1145/3636534.3695905

[32] Ronny Hadani and Anton Monk. 2018. OTFS: A New Generation of Modulation Addressing the Challenges of 5G. arXiv:1802.02623 [cs.IT] https://arxiv.org/abs/1802.02623

[33] Ahmad Hassan, Shivang Aggarwal, Mohamed Ibrahim, Puneet Sharma, and Feng Qian. 2024. Wixor: Dynamic TDD Policy Adaptation for 5G/xG Networks. *Proc. ACM Netw.* 2, CoNEXT4, Article 38 (Nov. 2024), 24 pages. doi:10.1145/3696395

[34] Ahmad Hassan, Arvind Narayanan, Anlan Zhang, Wei Ye, Ruiyang Zhu, Shuowei Jin, Jason Carpenter, Z Morley Mao, Feng Qian, and Zhi-Li Zhang. 2022. Vivisecting mobility management in 5G cellular networks. In *Proceedings of the ACM SIGCOMM 2022 Conference*. 86–100.

[35] Hongji Huang, Song Guo, Guan Gui, Zhen Yang, Jianhua Zhang, Hikmet Sari, and Fumiyuki Adachi. 2019. Deep Learning for Physical-Layer 5G Wireless Techniques: Opportunities, Challenges and Solutions. arXiv:1904.09673 [eess.SP] https://arxiv.org/abs/1904.09673

[36] Miaona Huang and Jun Chen. 2022. Joint Load balancing and Spatial-temporal Prediction Optimization for Ultra-Dense Network. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)* (Austin, TX, USA). IEEE Press. doi:10.1109/WCNC51071.2022.9771749

[37] Monsoon Solutions Inc. 2022. *Monsoon power monitor.* Retrieved January 2024 from https://www.msoon.com/online-store

[38] iperf3 community. 2023. *iPerf3.* Retrieved January 2024 from https://iperf.fr/iperf-download.php

[39] Rostand A K. Fezeu, Claudio Fiandrino, Eman Ramadan, Jason Carpenter, Lilian Coelho de Freitas, Faaiq Bilal, Wei Ye, Joerg Widmer, Feng Qian, and Zhi-Li Zhang. 2024. Unveiling the 5G Mid-Band Landscape: From Network Deployment to Performance and Application QoE. In *Proceedings of the ACM SIGCOMM 2024 Conference*. 358–372.

[40] Volodymyr Kuleshov and Doina Precup. 2014. Algorithms for multi-armed bandit problems. arXiv:1402.6028 [cs.AI] https://arxiv.org/abs/1402.6028

[41] Li Li, Ke Xu, Tong Li, Kai Zheng, Chunyi Peng, Dan Wang, Xiangxiang Wang, Meng Shen, and Rashid Mijumbi. 2018. A measurement study on multi-path TCP with multiple cellular carriers on high speed rails. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 161–175.

[42] Li Li, Ke Xu, Dan Wang, Chunyi Peng, Kai Zheng, Rashid Mijumbi, and Qingyang Xiao. 2017. A longitudinal measurement study of TCP performance and behavior in 3G/4G networks over high speed rails. *IEEE/ACM transactions on networking* 25, 4 (2017), 2195–2208.

[43] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. 2020. Towards Streaming Perception. arXiv:2005.10420 [cs.CV]

[44] Qianru Li and Chunyi Peng. 2021. Reconfiguring Cell Selection in 4G/5G Networks. In *2021 IEEE 29th International Conference on Network Protocols (ICNP)*. IEEE, 1–11.

[45] Qianru Li, Zhehui Zhang, Yanbing Liu, Zhaowei Tan, Chunyi Peng, and Songwu Lu. 2023. *CA++: Enhancing Carrier Aggregation Beyond 5G.* Association for Computing Machinery, New York, NY, USA.

[46] Yuanjie Li, Haotian Deng, Jiayao Li, Chunyi Peng, and Songwu Lu. 2016. Instability in Distributed Mobility Management: Revisiting Configuration Management in 3G/4G Mobile Networks. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science* (Antibes Juan-les-Pins, France) *(SIGMETRICS '16)*. Association for Computing Machinery, New York, NY, USA, 261–272. doi:10.1145/2896377.2901457

[47] Yuanjie Li, Haotian Deng, Chunyi Peng, Zengwen Yuan, Guan-Hua Tu, Jiayao Li, and Songwu Lu. 2016. iCellular: Device-Customized Cellular Network Access on Commodity Smartphones. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. USENIX Association, Santa Clara, CA, 643–656. https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/li-yuanjie

[48] Yuanjie Li, Qianru Li, Zhehui Zhang, Ghufran Baig, Lili Qiu, and Songwu Lu. 2020. Beyond 5G: Reliable Extreme Mobility Management. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication* (Virtual Event, USA) *(SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA, 344–358. doi:10.1145/3387514.3405873

[49] Zhi Li, Xiaoqing Zhu, Joshua Gahm, Rong Pan, Hao Hu, Ali C Begen, and David Oran. 2014. Probe and adapt: Rate adaptation for HTTP video streaming at scale. *IEEE Journal on Selected Areas in Communications* 32, 4 (2014), 719–733.

[50] Yanbing Liu, Junpeng Guo, and Chunyi Peng. 2024. Demystifying Secondary Radio Access Failures in 5G. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications* (San Diego, CA, USA) *(HOTMOBILE '24)*. Association for Computing Machinery, New York, NY, USA, 114–120. doi:10.1145/3638550.3641125

[51] Yanbing Liu and Chunyi Peng. 2023. A Close Look at 5G in the Wild: Unrealized Potentials and Implications. In *IEEE International Conference on Computer Communications (INFOCOM'23)*.

[52] Yanbing Liu and Chunyi Peng. 5555. Handling Failures in Secondary Radio Access Failure Handling in Operational 5G Networks . *IEEE Transactions on Mobile Computing* 01 (Oct. 5555), 1–14. doi:10.1109/TMC.2024.3477462

[53] Kyle MacMillan, Tarun Mangla, James Saxon, and Nick Feamster. 2021. Measuring the performance and network utilization of popular video conferencing applications *(IMC '21)*. Association for Computing Machinery, New York, NY, USA, 229–244. doi:10.1145/3487552.3487842

[54] Arvind Narayanan, Eman Ramadan, Rishabh Mehta, Xinyue Hu, Qingxu Liu, Rostand AK Fezeu, Udhaya Kumar Dayalan, Saurabh Verma, Peiqi Ji, Tao Li, et al. 2020. Lumos5G: Mapping and predicting commercial mmWave 5G throughput. In *Proceedings of the ACM Internet Measurement Conference*. 176–193.

[55] Arvind Narayanan, Eman Ramadan, Jacob Quant, Peiqi Ji, Feng Qian, and Zhi-Li Zhang. 2020. 5G tracker: a crowd-sourced platform to enable research using commercial 5g services. In *Proceedings of the SIGCOMM'20 Poster and Demo Sessions*. 65–67.

[56] Arvind Narayanan, Xumiao Zhang, Ruiyang Zhu, Ahmad Hassan, Shuowei Jin, Xiao Zhu, Xiaoxuan Zhang, Denis Rybkin, Zhengxuan Yang, Zhuoqing Morley Mao, et al. 2021. A variegated look at 5G in the wild: performance, power, and QoE implications. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. 610–625.

[57] Natale Patriciello, Sandra Lagen, Biljana Bojovic, and Lorenza Giupponi. 2019. An E2E simulator for 5G NR networks. *Simulation Modelling Practice and Theory* 96 (2019), 101933.

[58] Chunyi Peng and Yuanjie Li. 2016. Demystify Undesired Handoff in Cellular Networks. In *2016 25th International Conference on Computer Communication and Networks (ICCCN)*. 1–9. doi:10.1109/ICCCN.2016.7568506

[59] Tech Radar. 2025. *T-Mobile went down – everything we know about this network outage.* Retrieved January 2025 from https://www.techradar.com/news/live/tmobile-november-outage

[60] The Register. 2025. *Failure to follow proper procedures caused US-wide AT&T outage, FCC says.* Retrieved January 2025 from https://www.theregister.com/2024/07/23/atandt_outage_fcc_report/

[61] Muhammad Iqbal Rochman, Wei Ye, Zhi-Li Zhang, and Monisha Ghosh. 2024. A Comprehensive Real-World Evaluation of 5G Improvements over 4G in Low-and Mid-Bands. In *2024 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 257–266.

[62] Chenhua Shen. 2015. Analysis of detrended time-lagged cross-correlation between two nonstationary time series. *Physics Letters A* 379, 7 (2015), 680–687.

[63] Kevin Spiteri, Rahul Urgaonkar, and Ramesh K Sitaraman. 2020. BOLA: Near-optimal bitrate adaptation for online videos. *IEEE/ACM Transactions On Networking* 28, 4 (2020), 1698–1711.

[64] srslte. 2024. *srsRan 4G.* Retrieved January 2024 from https://github.com/srsran/srsRAN_4G

[65] srslte. 2024. *srsRan Project.* Retrieved January 2024 from https://www.srslte.com/

[66] FoneArena News Team. 2025. Qualcomm and Ericsson showcase AI-powered wireless and 6G advances at MWC 2025. https://www.fonearena.com/blog/447227/qualcomm-ai-powered-wireless-6g-advances-mwc-2025.html. Accessed: 2025-05-31.

[67] Michael ThelanderMichael Thelander. 2023. This is why I TURNED OFF 5G. https://www.linkedin.com/posts/michaelthelander_this-is-why-i-turned-off-5gfor-someone-activity-7115871320852103168-I7_L/. Accessed: 2024-09-05.

[68] USRP. 2024. *USRP B210 Universal Software Radio Peripheral.* Retrieved January 2024 from https://www.ettus.com/all-products/ub210-kit/

[69] Deepak Vasisht, Swarun Kumar, Hariharan Rahul, and Dina Katabi. 2016. Eliminating Channel Feedback in Next-Generation Cellular Networks. In *Proceedings of the 2016 ACM SIGCOMM Conference* (Florianopolis, Brazil) *(SIGCOMM '16)*. Association for Computing Machinery, New York, NY, USA, 398–411. doi:10.1145/2934872.2934895

[70] Verizon. 2024. *Verizon mmWave Coverage.* Retrieved January 2024 from https://www.verizon.com/coverage-map/

[71] videosdk live. 2025. *WebRTC based video conferencing SDK for Android.* Retrieved January 2025 from https://github.com/videosdk-live/videosdk-rtc-android-kotlin-sdk-example

[72] Developed with Google. 2023. *WifiRttLocator App.* Retrieved January 2025 from https://play.google.com/store/apps/details?id=com.google.android.apps.location.rtt.wifirttlocator

[73] Yaxiong Xie. 2024. *NG-scope.* Retrieved January 2024 from https://github.com/YaxiongXiePrinceton/NG-Scope

[74] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. 2020. Understanding operational 5G: A first measurement study on its coverage, performance and energy consumption. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 479–494.

[75] Xinlei Yang, Hao Lin, Zhenhua Li, Feng Qian, Xingyao Li, Zhiming He, Xudong Wu, Xianlong Wang, Yunhao Liu, Zhi Liao, et al. 2022. Mobile access bandwidth in practice: Measurement, analysis, and implications. In *Proceedings of the ACM SIGCOMM 2022 Conference*. 114–128.

[76] Qiaoyang Ye, Beiyu Rong, Yudong Chen, Mazin Al-Shalash, Constantine Caramanis, and Jeffrey G. Andrews. 2013. User Association for Load Balancing in Heterogeneous Cellular Networks. *IEEE Transactions on Wireless Communications* 12, 6 (2013), 2706–2716. doi:10.1109/TWC.2013.040413.120676

[77] Wei Ye, Xinyue Hu, Steven Sleder, Anlan Zhang, Udhaya Kumar Dayalan, Ahmad Hassan, Rostand A. K. Fezeu, Akshay Jajoo, Myungjin Lee, Eman Ramadan, Feng Qian, and Zhi-Li Zhang. 2024. Dissecting Carrier Aggregation in 5G Networks: Measurement, QoE Implications and Prediction. In *Proceedings of the ACM SIGCOMM 2024 Conference* (Sydney, NSW, Australia) *(ACM SIGCOMM '24)*. Association for Computing Machinery, New York, NY, USA, 340–357. doi:10.1145/3651890.3672250

[78] Zengwen Yuan, Qianru Li, Yuanjie Li, Songwu Lu, Chunyi Peng, and George Varghese. 2018. Resolving policy conflicts in multi-carrier cellular access. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 147–162.

[79] Zhehui Zhang, Yanbing Liu, Qianru Li, Zizheng Liu, Chunyi Peng, and Songwu Lu. 2023. Dependent Misconfigurations in 5G/4.5G Radio Resource Control. *Proc. ACM Netw.* 1, CoNEXT1, Article 2 (July 2023), 20 pages. doi:10.1145/3595288

[80] Xiao Zhu, Subhabrata Sen, and Z Morley Mao. 2021. Livelyzer: analyzing the first-Mile ingest performance of live video streaming. In *Proceedings of the 12th ACM Multimedia Systems Conference*. 36–50.

Table 6. An overview of common CM procedures and their UE data transmission modes (idle, inactive, connected), actions (add, modify, remove cells), and criteria (cell accessibility, link quality, absolute priority).

Procedure	Action on UE's Cell Set	Mode	Cell Accessibility	Radio Link Quality	Absolute Priority
Cell Selection	Adds PCell	Idle/Inactive	Mandatory	Irrelevant	Primary
Cell Reselection	Modify PCell	Idle/Inactive	Mandatory	Primary	Primary
Handover	Modify PCell/PSCell	Connected	Mandatory	Primary	Secondary
Carrier Aggregation	Add/Remove SCells	Connected	Mandatory	Primary	Primary*
Dual Connectivity	Add/Remove PSCell	Connected	Mandatory	Primary	Secondary
Load Balancing	Modify PCell/PSCell	Connected	Mandatory	Irrelevant	Primary

*Measurements below suggest that operators often add SCells in groups, not sequentially.

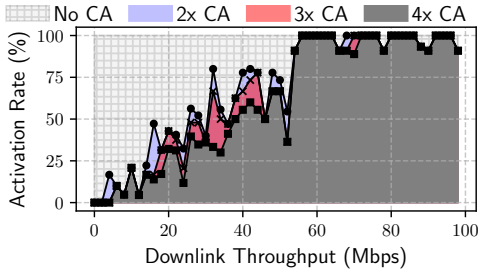


Fig. 22. Group-based addition of carriers during CA.

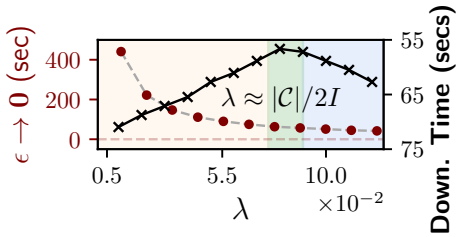


Fig. 23. Impact of λ on **OPCM** performance.

A Additional Measurement Details

Cell Management (CM) procedures. Table 6 gives an overview of CM procedures and describes if a certain criterion is mandatory, primary, or secondary for a procedure.

Group-based carrier aggregation. We perform stationary tests under ideal line-of-sight conditions to analyze how operators add carriers or SCells during CA. Using iperf3 [38], we generate downlink traffic with sending rates between 1-100 Mbps and collect lower-layer CA data with XCAL. Our analysis in Fig. 22 reveals two key findings. *(i)* Operators employ threshold-based policies to trigger CA, with relatively low activation thresholds. Additional carriers are activated at sending rates of 4.4 Mbps, 7.8 Mbps, and 11.4 Mbps for the second, third, and fourth carrier, respectively. *(ii)* The prevalence of 4x CA even at low sending rates suggests a group-based CA strategy, where BSs add multiple carriers simultaneously rather than sequentially.

B Design Details

Decay rate for ϵ-greedy policy. The decay rate λ controls how fast exploration rate ϵ decreases at startup. We run ns-3 simulations using our dataset from §2.4. Fig. 23 plots how quickly ϵ decays to 0 for different λ values (left y-axis). It also shows the average file download time (right y-axis inverted). When λ is too small (yellow region), OPCM does not explore enough at startup to find the best performing cell combination. In contrast, when λ is large (blue region), OPCM has lower network throughput due to frequent exploration. OPCM performs best when λ is close to $|C|/2I$=0.08 (green region). This value is not surprising as the optimal λ value [40] for an epsilon-greedy policy is: (i) proportional to the number of cell combinations to explore $|C|$, and (ii) inversely proportional to the best cell combination change frequency, which is I=32 secs on average for our dataset.

Received June 2025; accepted September 2025