Title
Mobility Support for 5G Networks and Beyond: New Challenges and Novel Solutions

Author
Li, Qianru

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Mobility Support for 5G Networks and Beyond: New Challenges and Novel Solutions

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science

by

Qianru Li

2022

ABSTRACT OF THE DISSERTATION

Mobility Support for 5G Networks and Beyond:
New Challenges and Novel Solutions

by

Qianru Li
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2022
Professor Songwu Lu, Chair

Cellular networks are the only large-scale system that supports mobility. Nowadays, 6.6 billion smartphone users [use] rely on cellular networks to stay connected with the Internet wherever they go. Mobility support is the substantial factor that makes the *anytime, anywhere* access happen. As a *cell* is a basic unit to provide mobile network services, a client should continuously camp on a nearby cell and get migrated to a new cell before leaving the limited coverage of the current one. So far, the existing infrastructure and practices have successfully ensured seamless connectivity.

Driven by emerging applications with the more stringent network services requirements, 5G/4G has been actively enhancing network capabilities (e.g., mmWave bands, carrier aggregation). On top of rich radio resources and advanced technologies, mobility support turns out to be a decisive factor in high reliability and high throughput. Specifically, mobile users rely on mobility support to enable seamless data transfer with negligible network failures and select good cells out of candidates with colossal throughput variance. More challenging, it should also work well in extreme mobility, which has become a norm in our daily life, like

high-speed rails and mmWave cells with a small coverage. Not limited to ubiquitous access, the 5G/4G mobility support is faced with new challenges: to achieve high reliability and high throughput in both extreme mobility and low mobility scenarios.

However, the existing mechanisms fail to address the challenges. For reliability, network failures increase vastly in extreme mobility. In low mobility, multi-carrier access is unstable and incurs persistent switching loops. To boost throughput, the network adopts carrier aggregation (CA) which serves one user with multiple cells and acquires a broader spectrum covering mmWave bands. Nonetheless, the mobility support largely misses cell groups with throughput and undermines the potential of rich radio resources. Our study indicates that all issues are attributed to the inherent design of mobility support: It is confined to the conventional goal – ubiquitous access, and can hardly match the upgrade in capability and demand. This dissertation proposes a revolutionary design to tackle the challenges above.

Unreliable 5G/4G under extreme mobility is rooted in *wireless-driven* mobility management. While reasonable in static and low mobility, it is vulnerable to drastic wireless dynamics from extreme mobility. Therefore, we devise REM, Reliable Extreme Mobility management. Based on the intuition that movement is far more stable than wireless, REM shifts to movement-based mobility management. The signaling overlay in the delay-Doppler domain exposes the client movement. REM exploits it to stabilize signaling and speed up feedback via cross-channel estimation and simplified cell selection logic. Our evaluation with operational high-speed rail datasets shows that, REM reduces failures comparable to static and low mobility, with low signaling and latency costs.

Unstable multi-carrier access is caused by conflicts between carrier switching policies and cell handover policies. We derive the conditions under which such oscillations occur and validate them with Google Fi to resolve the conflicts. Then, we manage to solve the problem without intruding on the internal policies of each mobile carrier. The critical insight is *to prioritize intra-carrier policy over inter-carrier policy.* Moreover, inconsistency among policies can be readily prevented by observing a few simple rules. We provide practical guidelines to

assure conflict-free interplay based on the findings above. The trace-driven emulations with Google Fi validate that our guidelines could eliminate all uncovered inconsistencies.

The current practice undermines CA and wide spectrum resources with mmWave cells. The problem is rooted in *sequential* operations that measure and aggregate cells sequentially. We devise CA++ to fulfill CA potentials over a broad frequency spectrum. The key idea is to *parallelize* cell selection with a group-based form. First, it enables parallel channel quality estimation to replace sequential measurement because aggregated cells share the same propagation paths. On top of that, it adopts group-based operations to evaluate cell groups and reach the best (or close) group. Real-world experiments and trace-driven emulation confirms that CA++ can greatly enhance CA performance.

We propose another design, RPerf, as a ready-to-launch solution to make better utilization of CA and rich radio resources in low mobility cases. The idea is to avoid cells with poor throughput by tuning parameters in the cell selection policies. Those parameters specify the criteria to determine whether and how to choose the next serving cell. Such reconfiguration enforces cell selection to consider throughput rather than target connectivity only. Furthermore, as a self-optimizing algorithm, RPerf can be directly adopted by the operational networks. The trace-driven evaluation over AT&T and T-Mobile shows promising throughput gains.

To prototype our design and provide a quick testing solution for other researchers, we set up a mobility testbed based on Flora [Flo], a flexible open-source 5G/4G platform. Our primary contribution is to support essential mobility functions, handovers and CA, on commodity phones. Therefore, it meets the desire to conduct realistic experiments in practice, which is unfortunately missed by existing solutions.

The dissertation of Qianru Li is approved.

Christina Panagio Fragouli

Lixia Zhang

George Varghese

Songwu Lu, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I am very grateful for my advisor, Prof. Songwu Lu. Throughout my Ph.D. career, he has always been supportive, trusting, and patient with my research. His passion for research, scientific attitude, and critical thinking continuously encourage me to overcome technical difficulties and produce high-quality works. From the beginning to the end, Prof. Lu is very generous in sharing his advice, insight, and vision. It has always been my guide in finding and studying new and appealing topics in computer science.

I would express my genuine appreciation for my dissertation committee: Prof. Lixia Zhang, Prof. George Varghese, and Prof. Christina Fragouli. They have provided constructive suggestions on my dissertation and precious advice on my future career. I am also thankful to Prof. Chunyi Peng from Purdue University and Prof. Lili Qiu from the University of Texas Austin. Their valuable suggestions have significantly benefited my research.

I am grateful to colleagues and collaborators from UCLA and other research groups or industry corporations for endorsing my research. During my Ph.D. study, I significantly enjoy working with my brilliant and diligent colleagues: Yuanjie Li, Zhehui Zhang, Zengwen Yuan, Zhaowei Tan, Muhammad Taqi Mehdi, Jinghao Zhao, Yunqi Guo, Boyan Ding, and Yifei Xu. I would not produce those works or earn happiness from research without you. In addition, I am indebted to my mentors and colleagues during my internship at Uber, Google, and HP Labs: Vinoth Chandar, Rajesh Mahindra, Yihua (Ethan) Guo, Sivabalan Narayanan, David Chu, Zengbin Zhang, Teng Wei, Connor Smith, and Kyu-Han Kim. I also consider myself fortunate to collaborate with these excellent researchers: Haotian Deng, Ghufran Baig, Yanbing Liu, Jingqi Huang, Kai Ling, and Jiacen Liu.

Lastly, I would thank my parents for their endless love and support. I would not complete my Ph.D. journey without their encouragement and trust. My deepest and most intimate heartfelt is reserved solely for my husband, Chen Liang. I have no fear of any past, present, or future challenges, with your company.

VITA

2012-2016 B.S. (Computer Science), Shanghai Jiao Tong University.

2016-Present Ph.D. Student, Computer Science Department, UCLA.

RELATED PUBLICATIONS

Qianru Li, Zhehui Zhang, Yanbing Liu, Zhaowei Tan, Chunyi Peng, Songwu Lu. *CA++: Enhancing Carrier Aggregation Beyond 5G*, Under review, **ACM MobiCom 2022**.

Zhehui Zhang, Yuanjie Li, Qianru Li, Jinghao Zhao, Ghufran Baig, Lili Qiu, Songwu Lu. *Movement-based Reliable Mobility Management for Beyond 5G Cellular Networks*, Under review, **IEEE/ACM Transactions on Networking (ToN) 2022**.

Qianru Li, Chunyi Peng. *Reconfiguring Cell Selection in 4G/5G Networks*, **IEEE ICNP 2021**.

Yuanjie Li, **Qianru Li**, Zhehui Zhang, Ghufran Baig, Lili Qiu, Songwu Lu. *Beyond 5G: Reliable Extreme Mobility Management*, **ACM SIGCOMM 2020**.

Haotian Deng, **Qianru Li**, Jingqi Huang, Chunyi Peng. *iCellSpeed: Increasing Cellular Data Speed with Device-AssistedCell Selection*, **ACM MobiCom 2020**.

Zengwen Yuan, **Qianru Li** (co-primary), Yuanjie Li, Songwu Lu, Chunyi Peng, George Varghese. *Resolving Policy Conflicts in Multi-Carrier Cellular Access*, **ACM MobiCom 2018**.

Zhaowei Tan, Yuanjie Li, **Qianru Li**, Zhehui Zhang, Zhehan Li, Songwu Lu. *Supporting Mobile VR in LTE Networks: How Close Are We?*, **ACM SIGMETRICS 2018**.

CHAPTER 1

Introduction

Cellular networks are the only large-scale system that supports mobility. They provide Internet access for billions of mobile users *anytime and anywhere*. The key enabler is the wide-area mobility support mechanisms. To put it simply, the mobility support switches the serving cell (i.e., the basic unit offering mobile network service) when a client moves out of its coverage, which is called *handover*. Mobility support has been a great success over the past decades by carrying out the promise of seamless connectivity.

As the mobile network is evolving to the next generation (5G) and beyond, mobile users are demanding better experiences even under thrilling use scenarios, including virtual/augmented reality, vehicle-to-everything communication (V2X), Internet-of-Things (IoT), to name a few. Therefore, mobile carriers have been actively adding more spectrum resources and developing advanced technologies to enhance network capabilities. For example, since late 2019 (the first 5G rollout), AT&T has increased its total downlink channel-width from 258 MHz to 4033 MHz by acquiring mmWave bands and mid-frequency 5G bands, repurposing 4G bands, etc. In addition, one advance is *carrier aggregation* (CA) which serves one client with more than one cell on different frequency channels simultaneously. CA combines "small trunks" scattered on the frequency spectrum to expand the channel for data transfer. Another technology is *multi-carrier access* which dynamically updates the mobile carrier for the client to improve coverage and radio performance.

Are those enhanced network capabilities *sufficient* to meet the high standard of network performance in 5G/4G? Not really. Mobility support is another decisive factor, especially

for network *reliability* and *throughput*. For reliability with negligible handover failures and thus seamless connectivity, we rely on mobility support to make it. In addition, mobility support is inherently cell selection. Considering diverse 5G/4G cells with a huge variance in throughput, making a good selection is critical to obtaining good performance.

This dissertation studies the challenges of achieving high reliability and high throughput with 5G/4G mobility support. We showcase the challenges in two use scenarios: extreme mobility and low/moderate mobility. Here, extreme mobility is characterized by the short residence time with each serving cell and thus could happen under fast speed or small cells (e.g., mmWave). Specifically, we present four challenges in §1.1 and contribute effective and novel solutions in §1.2.

1.1 Challenges of Mobility Support Today

1.1.1 High reliability

Is 5G/4G still reliable in extreme mobility? Extreme mobility becomes a norm rather than an exception. We have witnessed a boom in various extreme mobility scenarios, such as high-speed rails, vehicle-to-everything, drones, and many more. Compared to traditional static and low-mobility scenarios, extreme mobility involves much faster client movement speed (up to 350km/h [Wik19b]) or radio resources on a higher frequency (e.g., millimeter waves above 24 GHz) in the outdoor environment. While the existing mechanisms have successfully supported billions of mobile users, most of them are moving slowly or static. Despite notably increased moving speed and wireless dynamics, mobility support should remain reliable and maintain network failures as negligible as in the conventional low-mobility or static case.

Is multi-carrier access stable? Multi-carrier access selects a preferred mobile carrier from multiple choices (e.g., T-Mobile, Sprint) deployed at a location. It exploits the diversity of mobile carriers and radio access technologies (RATs) at any location to improve cover-

age and access speed. Google has deployed the first multi-carrier access system in *Google Fi* [Goo]. At the core of multi-carrier cellular access, the technology extends the legacy mobility support to a *two-tier switching* scheme. The top tier allows the device to select and switch to a preferred carrier network (aka inter-carrier switch). There are handovers within a carrier at the low tier to connect to the target cell (aka intra-carrier cell switch). While the handover (in)stability is well studied and practiced [zte, LXP16, LDL16, LMP14, CRR09], the inter-carrier switch is still largely unexplored. Switching decisions made by more parties requires careful scrutiny to avoid oscillation and make mobility support stable.

1.1.2 High throughput

Advancing from 1 to N ($\geq$1) cells, CA offers an effective mechanism to combine different chunks of spectrum and thus adequately utilize existing and emerging spectrum resources (Figure 1.1). Wireless spectrum resources constantly grow with new bands acquired over time; the total channel width in use expands from several hundreds of MHz to several GHz. As a result, more and more cells are available to serve devices. CA and dense cell deployment have fundamentally reformed mobility support to selecting a group of cells out of massive candidates. Seemingly guaranteed, it is non-trivial to select a combination with high throughput anytime, anywhere, given vastly increased options (in terms of groups) and more significant variance. In general, we want to ask: **Can cell selection achieve the potential of CA and rich radio resources?**

In extreme mobility with mmWave cells, the challenge is upgraded: **Can cell selection be rapid and achieve high throughput simultaneously?**

1.2 Our Contribution

Unfortunately, our empirical study shows that the current practice *fails* to tackle reliability and throughput challenges. The problem is rooted in the *inherent design philosophy*

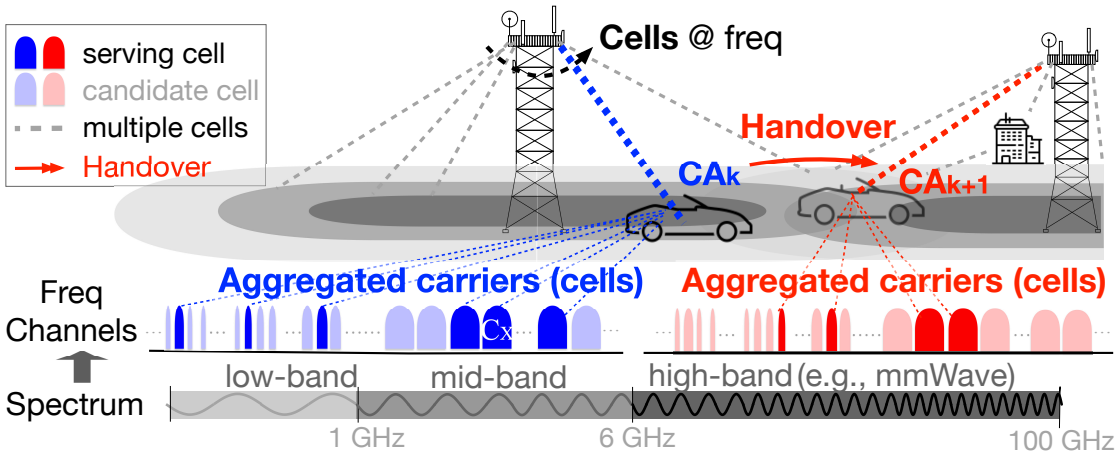


Figure 1.1: Carrier aggregation (CA) on the move.

of 5G/4G mobility support: designated for connectivity, low mobility, and homogeneous carrier access. However, it can hardly match the increased radio resources and advanced technologies. Therefore, we propose designs to address the challenges fundamentally.

1.2.1 REM: Reliable Extreme Mobility

(Reliability issue in extreme mobility) Our study on high-speed rails unveils that the current practice cannot achieve high quality in extreme mobility, as handover failures arise with alarming frequency: When the speed is above 200 km/h, the failure ratio goes beyond 10.6% (versus 4.3% for speed <100 km/h). Dramatic changes in wireless channel quality directly cause a surge in failures.

This dissertation challenges 5G/4G's *wireless-driven design*, which is very vulnerable to channel dynamics. The mobility support takes wireless signal strength as input, relies on the client-side feedback to trigger, and decides the target based on policies. While reasonable in static and low mobility, this design cannot tolerate dramatic wireless dynamics from the rapid multi-path fading and significant Doppler shift in extreme mobility.

We propose REM, **R**eliable **E**xtreme **M**obility management for 4G, 5G, and beyond. Our critical insight is that *client movement is more robust and predictable than wireless signal*

strength, thus suitable to drive mobility management. So REM shifts to *movement-based mobility management* and comes as the first such design for cellular networks. REM is a signaling overlay in the *delay-Doppler domain*, which reveals client movement and multi-path profile to help stabilize handover signaling, advance feedback, and simplify decision logic. REM is backward compatible with 5G/4G in static and low mobility without changing their designs or data transfer.

We prototype REM in commodity software-defined radio and evaluate it with high-speed rails datasets and 5G/4G standard channel models. Compared to solutions today, REM reduces failures by up to an order of magnitude (0.9×–12.7× depending on client speed). Moreover, REM controls the failure ratio to a level comparable to static and low mobility scenarios in extreme mobility. Meanwhile, REM retains marginal overhead of signaling traffic and latency without hurting data transfer.

1.2.2 Stable multi-carrier access

(**Reliability issue in low mobility**) We show that multi-carrier access is *unstable*, as the policy conflicts arise between inter-carrier selection and cell handovers within each carrier (akin to BGP loops). Such conflicts force the device to oscillate between carriers, disrupt the device's network service, slow down performance, and drain the device's battery.

We aim to resolve conflicts and make stable two-tier switching in the new context. In general, both inter-carrier switching and intra-carrier handover are *policy-based selections*. Reflecting the decision logic, a policy contains specific attributes in preference values and threshold-based criteria. Inspired by the methodology in the BGP study, we seek to analyze when such conflicts arise and derive analogous (but very different from [GR01]) conditions. We devise practical guidelines for stable multi-carrier access based on theoretical results. The fundamental challenge is coordinating inter-carrier and intra-carrier policies while protecting each mobile carrier's operation autonomy and privacy from others and the multi-carrier service provider (e.g., Google Fi).

Our critical insight is that *intra-carrier policy, which has been largely standardized and commonly practiced, should be prioritized over inter-carrier policy for conflict-free design.* In addition to stability, the top priority, our guidelines seek to retain policy confidentiality and flexibility. Specifically, we use them to regulate the inter-carrier policy only, without asking carriers to disclose their internal policies. Furthermore, they leave sufficient flexibility for carriers and MCSP to customize the selections. Finally, we assess the effectiveness of policy guidelines using trace-driven emulations. The results have validated that no loops would occur after this regulation.

1.2.3 CA++: Enhancing Carrier Aggregation

(**Throughput issue in extreme mobility**) The mobility support involving CA and mmWave cells should select good cell groups and quickly pick them up to boost throughput with broad-spectrum resources [Eri21, Ame19]. However, it is not the case in practice – And our empirical study (§6.1) indicates that it is not even close. In most cases (>75%), the aggregated channel width can go beyond 400MHz; But the reality is that such spectrum resources are not used in more than 50% of cases.

After examing the existing mechanisms and architecture of multi-cell selection, we unveil the root cause: *The sequential, cell-by-cell selection compromises the utilization under the tension of responsiveness and quality.* Responsiveness demands quick cell selections to keep connectivity especially given mmWave cells. At the same time, quality calls for exploring many candidate cells in groups as the foundation for making good decisions. In this dissertation, we seek to reach the best of both worlds and thus fulfill the potential of CA.

We devise CA++ with two enabling blocks (++) to achieve the goal. The first is the parallel estimation of channel quality. The first is parallel estimation of channel quality. It exploits the observation that the aggregated cells are located on the same base station and share common propagation paths. Instead of measuring all candidate cells, we only measure one per base station and estimate others, enhancing *responsiveness.* We transform

measurement in the conventional time-frequency domain into the delay-Doppler domain, which can easily map the multi-path profile extracted from one frequency onto another frequency. The parallel estimation quickly generates the accurate wireless quality of cells over a broad frequency spectrum. CA++ further replaces cell-by-cell operations with group-based forms. In this way, the network can evaluate, compare and select the best cell group, which unleashes higher *quality* CA.

We implement CA++ on an SDR-based testbed and evaluate it with real-world experiments and trace-driven emulation. Overall, CA++ brings channel width over 400MHz to 74.7% of cases, compared to 27.3% by the legacy mechanisms. The median throughput of all cases grows from 35.4 Mbps to 83.7 Mbps. Therefore, CA++ is promising to achieve responsiveness and quality over wide spectrum and even under high mobility.

1.2.4 RPerf: Reconfiguring cell selection for better throughput

(**Throughput issue in low mobility**) Our study uncovers that the enhanced 5G/4G capabilities do not guarantee to turn into throughput gains desired by mobile users. For example, a client only gets data speed below 1 Mbps, whereas several tens of Mbps are available for heavy traffic like video streaming. It thus under-utilizes the full potential of enhanced network capabilities.

In this work, we design RPerf to tackle this under-utilization problem by preventing improper cell selection. We believe that it is promising because cell selection follows standardized procedures and is controlled by operator-specific policies pre-configured by tunable parameters [3GP15, 3GP19h]. By adjusting those configuration parameters where under-utilization originated from, we should be able to avoid or reduce the likelihood of cell selections that under-utilize network capabilities. However, it is not easy because reconfiguration helps in some instances but may hurt others. A desired solution must statistically improve the overall gain by tuning parameters in a high-dimensional space. It is inherently complex due to the correlation among parameters of one cell, among co-located cells and among

all the cells in a geographical area. RPerf tackles the challenges above with heuristics learned from our empirical study. We evaluate the effectiveness and efficiency of RPerf using 5G datasets with AT&T and T-Mobile. Performance gains outweigh the losses on all the datasets. Specifically, 30.0% of cases would increase speed by 200% (median), while only 14.0% of cases would experience a loss of 45.9%.

1.2.5 Mobility testbed on Flora

To prototype our design and provide a 5G/4G platform for other researchers, we extend Flora [Flo], a flexible software-defined mobile network testbed, to support mobility-related functions. In particular, we enable handovers and CA for commodity phones. For handover, we first provide a basic version between two cells on the same base station, given the hardware limitation. Then, we develop *virtual* cells to bring up multiple cells and enable handovers among them. We also implement CA operations compatible with real phones' capability.

1.3 Organization of the Dissertation

This dissertation is organized into the following chapters. §2 introduces the primer of 5G/4G mobility support by answering *what* and *why it is crucial*. §3 overviews our findings that the current practice fails to tackle the above challenges (reliability and throughput in extreme or low mobility scenarios) and corresponding solutions. §4 uncovers unreliable extreme mobility management and proposes REM to solve the problem. §5 studies the instability of multi-carrier access and proposes practical guidelines to resolve underlying conflicts. §6 studies how the limitation in current cell selection impedes the potentials of CA and mmWave, and devises CA++ to make responsive and satisfactory selections. §7 adopts reconfiguration to make performance-aware cell selections and thus better utilize enhanced network resources. Finally, we summarize the related work in §9, draw conclusions, highlight insights, and discuss future work in §10.

CHAPTER 2

Background and State-of-the-art

This chapter introduces 5G/4G mobility support by answering the following questions: *How does 5G/4G support mobility* (§2.1), and *why is mobility support important* (§2.2)? In addition, we introduce the emergent extreme mobility and new challenges it brings to the provision of high-quality network service (§2.3).

2.1 How Does 5G/4G Support Mobility?

2.1.1 Handover: Basic instrument

In practice, 5G/4G services are provided by *mobile carriers* (or, say, *operators*), like AT&T and Verizon in the US. Mobile carriers have deployed many base stations to cover broad areas to enable ubiquitous network access. Each base station may run multiple *cells* on various *frequency channels* (interchangeable with *component carrier*). A cell is the primary logical unit to offer radio access to mobile clients in cellular networks. From the client's perspective, the cell currently connected to him is known as the *serving cell*. As a client leaves the serving cell's coverage, it will be migrated to another to retain the connectivity. This critical process to support fine-grained mobility is called *handover* [3GP15, 3GP19c, 3GP06, 3GP12a].

In 5G/4G networks, the serving cell controls handovers with the client's assistance. Figure 2.1 depicts the handover process, including steps ① through ⑤ [3GP15, 3GP19h].

① *Configuration.* Once the client connects to its serving cell, it is configured to monitor the serving and neighbor cells' signal strength. The serving cell also specifies the conditions

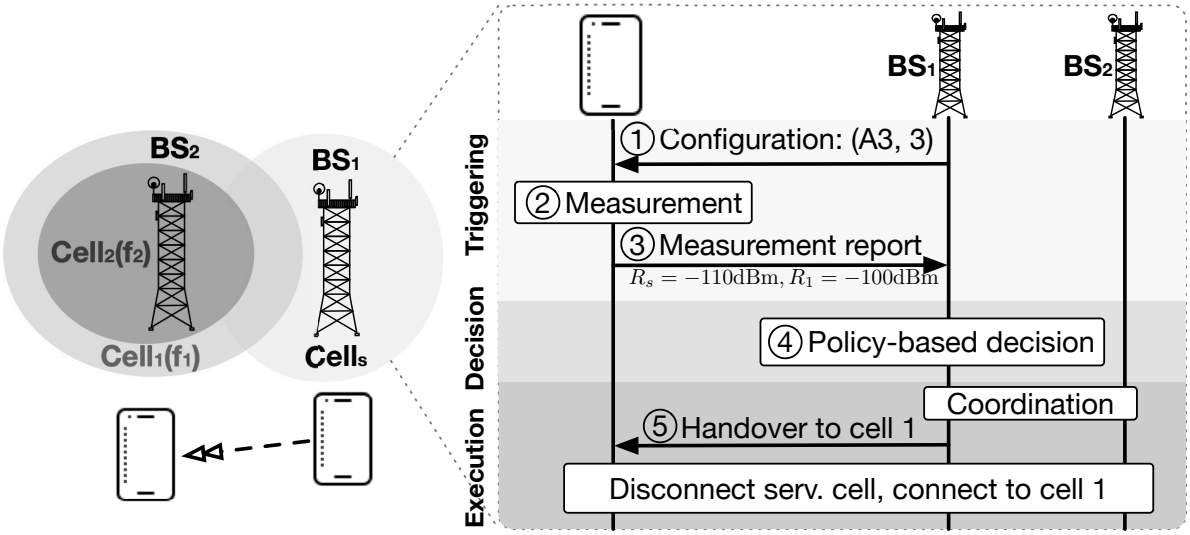


Figure 2.1: Procedure of handover: Switch one cell.

to report cells, as listed in Table 2.1.

② *Measurement.* Following the configuration, the client measures the signal strength of specified cell(s). In 5G/4G networks, wireless signals are modulated over several Orthogonal Frequency Division Multiplexing (OFDM) subcarriers. Reference signals are placed at the specific subcarriers and time slots to measure radio quality [3GP19e]. RSRP, RSRQ, and SNR represent radio signal level and quality [3GP19e].

③ *Feedback.* The client reports the signal strengths of measured cells to the serving cell once the corresponding criterion is satisfied.

④ *Decision.* Upon receiving the client's feedback, the serving cell runs its local policy to decide if a handover should start; if yes, which cells the client should migrate to. It may also reconfigure the client to send more feedback. Typically, the policy consists of preference over neighbor cells and threshold-based criteria to judge whether a candidate is good enough as the handover target.

⑤ *Execution.* With a decision made, the serving cell performs handover by coordinating with the target cell and sending a command to the client. Then the client will disconnect from the serving cell and connect to the target.

Table 2.1: Wireless triggering criteria in 5G/4G [3GP15, 3GP19h]

Event	Criteria	Explanation
A1	$R_s > \Theta_{A1}$	Serving cell becomes better than a threshold
A2	$R_s < \Theta_{A2}$	Serving cell becomes worse than a threshold
A3 (A6)	$R_n > R_s + \Theta_{A3}$	Neighbor cell becomes offset better than serving cell
A4 (B1)	$R_n > \Theta_{A4}$	Neighbor cell becomes better than a threshold
A5 (B2)	$R_s < \Theta_{A5,1}, R_n > \Theta_{A5,2}$	Serving cell becomes worse than a threshold, and neighbor cell becomes better than a threshold

As 5G/4G evolves with advanced technologies to improve radio coverage and network performance, they may extend mobility support beyond the conventional handover. Next, we introduce two major extensions caused by *multi-mobile-carrier access* and *carrier aggregation*.

2.1.2 Multi-carrier access: Two-tier switching

Multi-carrier cellular access selects a preferred mobile carrier and its radio access technology (RAT, like 5G) from multiple options (e.g., T-Mobile, Sprint, AT&T, and others) deployed at a location. It is appealing in several aspects. First, it provides better coverage. No single mobile carrier can ensure complete coverage at any location [Ope17]. Given that the device has the flexibility to switch among multiple carriers, the obtained coverage is the union of all carriers. Second, it offers better access speed. One carrier may only support 4G at a given location, while another has 5G. The device thus benefits from access to higher quality. Third, it offers a device-based solution without changing the carrier infrastructure. Compared with the carrier-centric solution, this approach is easier to deploy.

Multi-carrier access in reality. The industry has initiated the deployments for the multi-carrier access, such as Google Fi [Goo], Apple SIM [App], Huawei Skytone [Arc15], and Samsung eSIM [GSM]. Most notably, since 2015, Google Fi has offered the first such service

for the Nexus/Pixel phone models. It supports runtime switch between three U.S. 4G/3G carriers (T-Mobile, Sprint, and U.S. Cellular). Some ongoing standards [NGM15, KVB17] seek to support multi-carrier access in 5G. Current multi-carrier access is realized with the readily-available mechanisms in commodity phones. A single re-configurable SIM card is used to support multi-carrier access. Given only one cellular hardware interface, only one carrier is selected and used each time. When a new carrier is selected, a system app directly reconfigures the SIM profile to the new carrier, so that the device can register to that carrier. Such an inter-carrier switch decision is policy-based (to be elaborated in §5.1). Afterward, it relies on the carrier's internal mechanism to select the cell, as elaborated below.

Two-tier switching. Mobility support is further extended to a *two-tier switching* with the inter-carrier switch on the top and legacy handover within a single carrier on the second tier. As §2.1 introduces, the intra-carrier handover determines whether the device should move from the serving cell to another one and which cell it should move to. The decision is based on the per-cell policy and runtime measures. Note that the local policies are configurable to meet diverse requirements, such as selecting the best radio quality, letting operators specify their priorities for cells, etc.. The interplay between inter-carrier and intra-carrier switch policies will be elaborated in §5.3.1.

2.1.3 Carrier aggregation: Selecting a cell group

Carrier aggregation (CA) is a technology that combines more than one cell to serve a device [3GP17a]. As Figure 1.1 illustrates, CA is to form virtual radio access over a much broader spectrum by combining multiple smaller "chunks", the frequency channels (a.k.a component carriers). From 4G to 5G, carrier aggregation (CA) combines more and more serving cells to widen the channel. Specifically, 5G acquires more spectrum resources on sub-6GHz (<6 GHz) and mmWave (>24 GHz) bands.

Typically, aggregated cells operate on the same cell tower: one primary cell (PCell) and several secondary cells (SCell). Only PCell is responsible for radio resource control, like

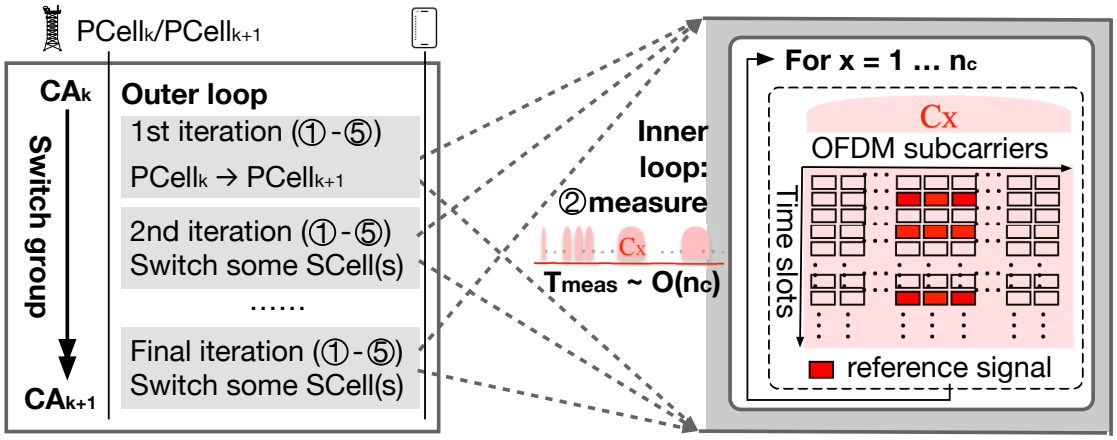


Figure 2.2: Procedure of CA on the move.

maintaining the radio connectivity on the move, while every cell is used for data transmission.

Selecting a group of cells. Considering CA, mobility support is extended to selecting a *group* of serving cells instead of a single one. Figure 2.2 illustrates the procedure to switch from an old group CA_k to a new group CA_{k+1}. There are two iteration loops. Every iteration updates some serving cells in the outer loop following the same five steps as handover (Figure 2.1). After an iteration is finished, the next one is invoked until the entire group converges to CA_{k+1}. Typically, the PCell is updated first, and then SCells are selected out of candidate cells by the new PCell.

The inner loop is to perform radio quality measurement on serving and candidate cells. First, the PCell configures a list of candidate cells (actually, their frequency channels) to measure. Then, if the frequency is the same as any serving one, the device can directly measure it. Otherwise, the device must first switch to that frequency to measure.

2.2 Why Is Mobility Support Important?

Mobility support is the fundamental instrument for providing ubiquitous network connectivity in cellular networks. Moreover, given users' stringent requirement of 5G/4G service quality, mobility support has become a decisive factor in achieving high reliability and

throughput.

High reliability. What does *reliability* mean? It has the following two requirements. First, handover failures should be negligible. Upon failures, the client cannot be migrated to the next serving cell seamlessly. Instead, the client must scan for a target and re-connect with it. A failure disrupts data transfer for hundreds of milliseconds to several seconds, stalling the service on the upper layer given TCP congestion control. The key to successful handover is timely decision-making before the client moves out of the current cell's coverage. Second, we desire stable handover without persistent loops, as mobility support inherently migrates the client among cells. Loops are harmful to both clients and the network, as they accumulate the cost of handover, including service disruption and excessive signaling. To avoid loops, the handover policies of neighbor cells should be well-coordinated, given that they make decisions in a distributed manner.

High throughput. Seemingly irrelevant, mobility support turns out to be a decisive factor of user-perceived throughput since it is inherently cell selection. 5G/4G should meet two requirements to boost data speed: (1) providing rich radio resources and (2) making them reachable by users. 5G/4G has made great efforts toward the first one by acquiring a much broader frequency spectrum (involving mmWave cells on bands >24GHz), deploying dense cells, and expanding the channel with CA (§2.1.3). While resources proliferate, however, every single cell selection is faced with more options and more considerable variance among them. Therefore, it is critical to make an optimal (or close) selection among dense cells and fulfill the potential of radio resources and advanced technologies.

2.3 Extreme Mobility

Nowadays, extreme mobility has become more and more common in our daily life. There are two typical use cases. First, clients move fast, like on high-speed rails. Second, clients are served by cells with small coverage, and the velocity is not necessarily very high. Like those

on mmWave bands, small cells are widely deployed by 5G. In both cases, mobile clients take *short residence time* with each serving cell and thus experience frequent switching.

In extreme mobility, the wireless dynamics become dramatic and will incur more uncertainty for mobility support. We briefly explain the dynamics based on the wireless quality and client mobility interplay. On the one hand, the wireless quality will decide the target cell for the mobile client (via handovers). On the other hand, as the client moves, the underlying signal propagation paths change accordingly and result in wireless variations (i.e., multi-path fading). The fast speed or high radio frequency also incurs a significant Doppler frequency shift, thus inter-carrier interference between cells and channel quality degradation. In 5G/4G OFDM/OFDMA[1], the channel remains approximately invariant in a short $T_c \propto 1/\nu_{max}$ [Wik19a], where T_c is the coherence time, $\nu_{max} \propto vf/c$ is the maximum Doppler frequency, f is the central frequency, v is client movement speed, and c is light speed. In static and low-mobility scenarios, the Doppler effect's impact is reasonably marginal (e.g., $T_c \approx$ 20ms for a vehicle at 60km/h under 900MHz 4G LTE band). Nevertheless, in extreme mobility, a fast-moving client (e.g., 200–350km/h in high-speed rails) under higher carrier frequency (e.g., mmWave) will experience fundamentally more dramatic channel dynamics (e.g., the coherent time $\approx$ 1ms). Despite the rapidly varying wireless channels, mobile users still demand high reliability and high throughput. In the following chapters, we show how to overcome the challenges.

[1]We use "OFDM" and "OFDMA" interchangeably since this dissertation focuses on wireless channel (not resource allocation), so they are equivalent.

CHAPTER 3

Overview: Challenges and Solutions

As §1 introduces, 5G/4G is faced with challenges in delivering high reliability and high throughput services in extreme and low mobility scenarios. This chapter showcases why the existing mobility support mechanisms fail to overcome those challenges (§3.1). Finally, we propose an innovative design to solve the problems (§3.2).

3.1 What is wrong the current mobility support?

Today's mobility support cannot overcome challenges to achieve high mobility or high throughput. Here, we briefly present the issues and root causes.

3.1.1 High reliability

Unreliable extreme mobility. Our large-scale empirical study on high-speed rails reveals vastly increasing failures and policy conflicts. Most failures are not caused by coverage holes in cell deployment. Instead, they are rooted in the limitation of the current design. In particular, the current wireless-driven mobility management is very vulnerable to dramatic wireless dynamics. Although it works well in the conventional static or low-mobility cases, the sharply deteriorating wireless qualities in extreme mobility corrupt signaling messages and incur handover failures. In that case, any tiny delay in any step of handover can amplify the possibility of failure.

Unstable multi-carrier access. 5G/4G mobility support is deemed reliable in the low

mobility, with negligible failures and attended instability. However, the recently integrated technology, multi-carrier access, incurs new problems. Though promising to exploit the diversity of mobile carriers and improve network performance, multi-carrier access is found with policy conflicts and incurs persistent switching loops. They continuously disrupt data service and drain the device battery. Here, conflicts exist between the inter-carrier switching policies and intra-carrier handover policies. Despite working well separately, their interplay is not well coordinated. In addition, the coordination is challenging as it should not violate the autonomy and privacy of each mobile carrier (from other carriers and Goggle Fi).

3.1.2 High throughput

Undermined potential of CA with mmWave. Despite CA and mmWave, the current cell selection mechanisms cannot fulfill such good potential and achieve high throughput. Based on our 5G tests in a US city, carrier aggregation with mmWave cells can provide a median channel width of 400MHz and median throughput of 100Mbps. However, the actual numbers are only 60MHz and 40Mbps, respectively. Our analysis indicates that the problem originated from CA's sequential, cell-by-cell selection. It fails to address the dilemma between *responsiveness* and *quality*. On the one hand, the network has limited time to select cells before handover (responsiveness), otherwise hurting the connectivity. On the other hand, aggregation of multiple cells requires exploration of more candidates to make a good selection (*quality*), which could be slow due to the sequential operations.

Under-utilized rich radio resources. 5G/4G operators have been heavily adding network resources to provide a faster mobile broadband experience, like acquiring broad radio spectrums, deploying dense cells, etc. However, the enhanced capabilities may not necessarily turn into throughput gains desired by mobile clients. Our 5G tests show that the current cell selection chooses poorly-performed cells in the presence of good candidate cells, which can offer much higher throughput. As a result, it greatly under-utilizes the full potential of enhanced network capabilities. Today's practice is still connectivity-centric and

	High reliability	**High throughput**
Extreme mobility	REM (§4): Mobility support driven by wireless → movement	CA++ (§6): Sequential → parallel cell selection
Low mobility/Static	Resolving conflicts (§5): Conflicting → coordinated policies	RPerf (§7): Connectivity-driven → throughput-aware configuration

Table 3.1: Summary of challenges and solutions.

thus offers throughput at "best-effort".

3.2 Design and insights

How to address the above issues? Based on the root causes, we revisit the legacy mobility support design and propose innovative solutions (summarized in Table 3.1):

- REM: **R**eliable **E**xtreme **M**obility management (§4). We seek to *decouple* wireless from mobility management to combat dramatic wireless dynamics. We shift the wireless-driven mobility management to *movement-driven* design. The insight is that client movement evolves much slower than wireless, and it is the underlying factor that *decides* the signal propagation paths and thus the wireless channel.

- Resolving policy conflicts in multi-carrier access (§5). To resolve the inconsistency, we derive conditions of loop-freedom on the interplay between inter-carrier and intra-carrier policies. We make practical guidelines to regulate the interplay, and the critical rule is to *prioritize* intra-carrier over inter-carrier decisions. The insight is that the latter is easier to regulate and does not violate the autonomy or privacy of mobile carriers.

- CA++ (§6). The key to fulfilling CA potential is to relax the tension between responsiveness and quality. Therefore, CA++ replaces the sequential, cell-by-cell selection with a group-based design. The insight is to achieve both responsiveness and quality by *parallelizing* the operation (measurement, feedback, decision) over multiple individual cells.

- RPerf: Reconfiguring cell selection towards better throughput (§7). The key idea is to make the connectivity-centric design aware of throughput. As local policies with tunable parameters control handovers, our design reconfigures those parameters to avoid or reduce the likelihood of poor cell selection.

- Mobility testbed on Flora (§8). We set up a mobility testbed based on Flora to prototype our design and provide a quick solution for other researchers. It provides the essential functions, handover and CA, running on commodity phones.

CHAPTER 4

REM: Reliable Extreme Mobility Management

4.1 Unreliable Extreme Mobility

The 4G/5G mobility management is fundamentally a *wireless signal strength-based* design: It takes *wireless signal strength* as the main input, relies on client-side *wireless feedback* to trigger, and selects the target cell based on *wireless-driven policies*. While reasonable in static and low mobility, such design is sensitive to wireless dynamics in extreme mobility, and raises non-negligible network failures and policy conflicts in all phases of mobility management. We detail each phase (§4.1.1–4.1.3), analyze 5G's impact (§4.1.4), and define the problem (§4.1.5).

An overview of extreme mobility in reality: Table 4.1 compares two LTE datasets from high-speed rails (HSR, one from [WZN19] and another from us) with our low mobility dataset (all detailed in §4.5). We make four high-level observations:

(1) Frequent handovers in extreme mobility: On average, a client on HSR experiences a handover every 20.4s, 19.3s, and 11.3s at <200km/h, 200–300km/h and 300–350km/h, respectively. Handover is more frequent as the train moves faster.

(2) Non-negligible failures in extreme mobility: Different from static or low mobility, the client suffers from frequent network failures in extreme mobility. To detect the network failures from mobility, we extract the handover events from LTE signaling messages, and check if the client successfully connects to the target cell for each handover. If not, the client loses radio connectivity and network access. We then compute the percentage of these

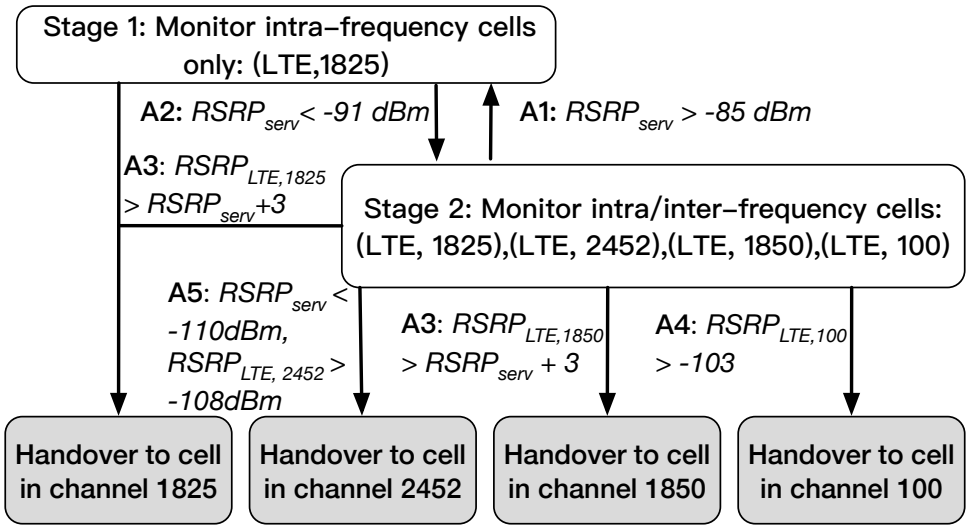


Figure 4.1: Policy state machine view

failures out of all handover events. Table 2.1 shows the failure becomes more frequent with faster speed, from 5.2% at $<$ 200km/h to 12.5% at 300–350km/h.

(3) Diverse failure causes: For each failure event in extreme mobility, we check its nearby wireless signal strength, signaling messages, and configurations in the LTE datasets to analyze its causes. Table 2.1 shows the failures arise from triggering (§4.1.1), decision (§4.1.2), and execution (§4.1.3). They can also *unavoidably* occur in a no-coverage area (e.g.,, caves). In LTE today, failures from coverage holes are not dominant (19.2%–33.3%). So we focus on failures *with* coverage.

(4) Policy conflicts from failures: To mitigate these failures, operators adopt proactive handover policies[1]. However, such practice incurs frequent policy conflicts (every 194.6–1090.0s on average) and voids operators' failure mitigation efforts (§4.1.2).

[1]We follow [LPY16] to model a serving cell's handover policy as a state machine, and infer it using the LTE signaling messages and configurations from the serving cell. Our inference is coherent with the policy from real 5G/4G vendors and operators [zte, Hua16].

Table 4.1: Network reliability in extreme mobility

		low mobility	high-speed rails (China)		
	Speed (km/h)	$0-100$	$100-200$	$200-300$	$300-350$
	Avg. handover interval	50.2 s	20.4 s	19.3 s	11.3 s
Failures (§4.1)	**Total network failure ratio**	4.3% (100%)	5.2% (100%)	10.6% (100%)	12.5% (100%)
	Feedback delay/loss (§4.1.1)	0.78% (18.0%)	1.7% (33.3%)	4.9% (46.3%)	6.9% (55.2%)
	Missed cell (§4.1.2)	1.8% (42.0%)	0.6% (11.1%)	0.4% (3.7%)	0.8% (6.4%)
	Handover cmd. loss (§4.1.3)	0.61% (14.0%)	1.1% (22.2%)	3.3% (31.5%)	2.4% (19.2%)
	Coverage holes	1.1% (26.0%)	1.7% (33.3%)	2.0% (18.5%)	2.4% (19.2%)
Conflicts (§4.1.2)	**Avg. loop frequency**	5,284.1s	410.1s	1,090.0s	194.6s
	Avg. # handovers/loop	2.2	3.9	3.0	3.3
	Avg. disruptions per loop	0.34 s	0.33 s	0.55 s	0.34 s
	Intra-frequency loops	**0%**	88.9%	100%	55.9%
	Inter-frequency loops	100%	11.1%	0%	44.1%

4.1.1 Triggering: Slow, Unreliable Feedback

5G/4G relies on client-side report to trigger handovers, and so triggering phase consists of the first three original steps: configuration, measurement and feedback (§2.1). The feedback tracks client-perceived *wireless quality* of cells based on standard criteria (Table 2.1). In extreme mobility, such wireless signal strength-based feedback can be sluggish and cause failures. It faces the fundamental dilemma between *exploration* (more measurements for proper decision) and *exploitation* (timely triggering for handover). This causes two reliability issues:

• **Slow feedback:** To avoid failures, the client should deliver feedback *before* it leaves serving cell's coverage. But existing feedback is slow for two reasons: *(1) Head-of-line blocking:* To decide an appropriate target cell, the client should detect *all* cells that meet the criteria. For wireless signal strength-based feedback, the client has to measure each

Table 4.2: Two-cell policy conflicts in HSR datasets.

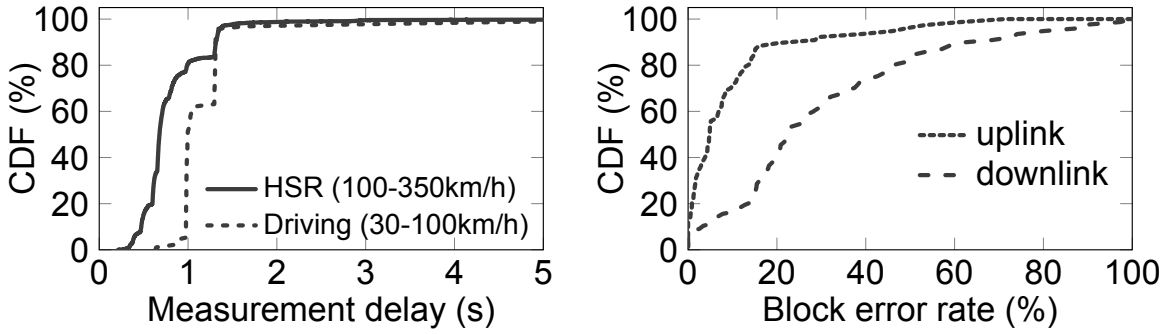
Conflicts	**Type**	**Beijing-Taiyuan**	**Beijing-Shanghai** [WZN19]
A3-A4	Inter-frequency	4 (2.4%)	316 (23.6%)
A3-A5	Inter-frequency	1 (0.6%)	24 (1.8%)
A4-A4	Inter-frequency	2 (1.2%)	200 (14.9%)
A4-A5	Inter-frequency	5 (3.0%)	49 (3.7%)
A5-A5	Inter-frequency	0	2 (0.1%)
A3-A3	**Intra-frequency**	**155 (92.8%)**	**749 (55.9%)**

cell *sequentially*, thus delaying later cells. Reducing the cells to measure can mitigate this delay, but at the risk of missing available cells (thus failures). *(2) Transient loop mitigation:* Instantaneous wireless measurement is dynamic and causes transient oscillations between base stations. To mitigate it, 4G/5G mandates the client to report a cell only if its criteria holds for a configurable triggering interval [3GP15, 3GP19h][2]. This delays feedback with late handovers. Moreover, wireless quality may have changed *before* measurements, thus causing sluggish feedback and misleading triggering. Shortening the triggering interval may help, but causes more transient loops and signaling.

• **Lost feedback:** With dramatic wireless dynamics, the feedback is prone to loss/corruption in delivery. Such loss can be amplified by feedback delay: The client may have left serving cell's coverage before measurement, thus losing more feedback.

Validation: Table 4.1 shows 33.3–55.2% failures in HSR are from feedback delay/loss. The loss is mostly caused by errors: Figure 4.2b shows 9.9% block error rate before the loss, which implies the feedback is corrupted in delivery. For the feedback delay, Figure 4.2a shows a client on HSR takes 800ms on average to generate feedback from different bands, during which it has moved 44.6–78.0m (200–350km/h) along the rails and is thus too late

[2]This configurable triggering interval is named as TimerToTrigger in 4G/5G.



(a) Slow feedback

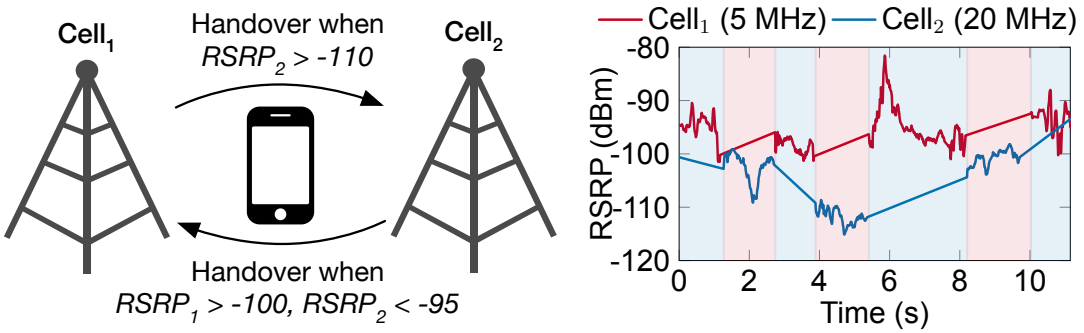
(b) Block errors in signaling loss

Figure 4.2: Unreliable handover triggering & execution.

for a viable handover. Moreover, the operator configures 40–80 ms the triggering interval for cells under same frequency as serving cell's (*intra-frequency cells*), and $128, 160, 256, 320$ or 640 ms for others (*inter-frequency cells*). These are 2 orders of magnitude longer than 5G/4G OFDM coherence time $T_c \approx c/fv \in [1.16ms, 6.18ms]$ (§2.3) given $f \in [874.2, 2665]$ MHz and $v \in [200, 350]$ km/h from our datasets. Note operators have shortened triggering interval for faster feedback than low mobility (mostly 640ms in our dataset), but at the cost of more transient loops and signaling.

Opportunity: Shared physical multipath It is possible to accelerate feedback *without* reducing the cells to be explored. In reality, a base station usually operates multiple cells under different bands to improve the radio coverage and performance. Our dataset shows 53.4% of cells share the same base station with another cell[3]. These cells' signals traverse the same paths from the base station to the client, thus experiencing similar channels. In §4.3, we will use this to relax the exploration-exploitation dilemma for reliable feedback.

[3]This is obtained by grouping the globally unique base station IDs from LTE cells' identifiers called ECIs [3GP11].



(a) Conflicts in load balancing (b) Real handover traces

Figure 4.3: Policy conflicts from load balancing in HSR.

4.1.2 Decision: Complex, Conflicting Policy

5G/4G handover decisions are policy-driven by design. To accommodate diverse demands (good radio coverage, fast data speed, load balancing, failure mitigation, etc.), each cell can customize its local policies with configurable criteria in Table 2.1. Figure 4.1 exemplifies a typical policy inferred from our HSR dataset[1]. Such policy is tightly coupled with wireless feedback (§4.1.1). It is too complicated for extreme mobility, and suffers from two deficiencies:

• **Multi-stage policy:** To tackle heterogeneous cells, most operators adopt *multi-stage* handover policies as exemplified in Figure 4.1. The neighbor cells under the same frequency as serving cell's are measured and chosen first. Only if no intra-frequency cells are available, the policy will consider inter-frequency cells via measurement reconfiguration. The reason is to reduce inter-frequency measurements, which consumes more radio resource and slows down the data transfer[4]. But if the client moves fast, this policy can miss candidate cells without sending its feedback to the serving cell. Even if no intra-frequency cells exist, extra

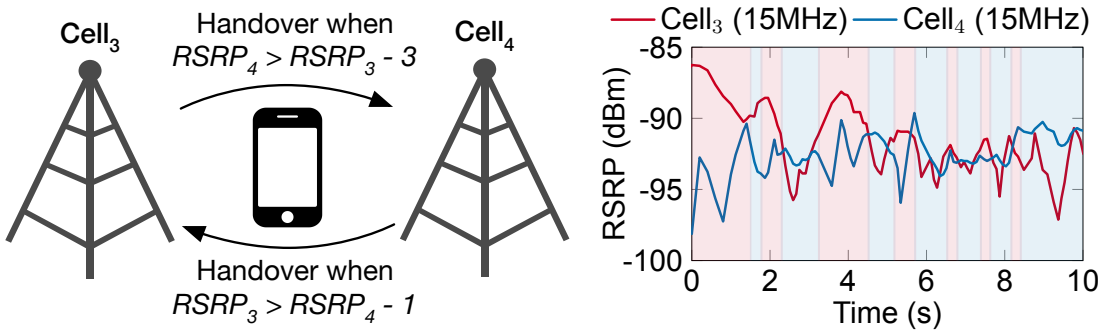
[4]To measure an inter-frequency cell, a client should synchronize to it and measure its signal strength. The serving cell pre-allocates MeasurementGaps [3GP17c, 3GP19e] for this, during which the client cannot send/receive data.

round trips (A2→reconfiguration→inter-frequency feedback) are needed for inter-frequency cells, during which the client may have missed the opportunity for handover and lost network access. The fundamental dilemma is that, inter-frequency measurements force existing policies to balance the spectral efficiency and decision delay.

• **Policy conflicts in extreme mobility:** It has been shown that [LDL16, YLL18], policies among cells can have conflicts and cause *persistent loops*. Figure 4.3a exemplifies a conflict from our dataset. Cell 1 and 2 have different bandwidths (5MHz v.s. 20MHz). For fast data speed, cell 1 moves a client to cell 2 if cell 2's signal strength $\text{RSRP}_2 > -110$dBm. But cell 2 adopts a different policy: It migrates a client to cell 1 if it is weak ($\text{RSRP}_2 < -95$dBm) and cell 2 is strong ($\text{RSRP}_1 > -100$dBm). Both policies can be *simultaneously* satisfied if $\text{RSRP}_1 > -100$dBm and $\text{RSRP}_2 \in (-110dBm, -95dBm)$. Then the client oscillates between cell 1 and 2 (8 handovers within 15s in Figure 4.3b). Such loop accumulates handover costs, disrupts client's service *and* incurs signaling storm for network.

Surprisingly, we note policy conflicts are amplified in extreme mobility, because of operators' desire for mitigating failures! This differs from [LDL16, YLL18] that focus on static scenarios, and has been frequently observed in our dataset (detailed in validation below). As shown in §4.1.1, a fast-moving client may miss the cells and lose service due to slow feedback and decisions. To mitigate it, the operators adopt proactive policies in Figure 4.4a, by running handovers *before* neighbor cell is better than serving cell's. However, this raises conflicts if neighbor cells use the same policy. Such policy will not mitigate failures; the client will move back with loops.

Validation: Our empirical study confirms both problems. First, multi-stage policy can miss inter-frequency cells and induce handover failures. It accounts for 3.7%–11.1% failures in HSR (Table 4.1). Even so, operators still prefer multi-stage policy due to its low spectral waste. Without multi-stage policy, our dataset shows MeasurementGap in HSR would consume 38.3%–61.7% spectrum in inter-frequency measurements (depending on cell configurations).



(a) Failure-induced conflicts (b) Real handover traces

Figure 4.4: Failure-induced policy conflicts in HSR.

Second, policy conflicts exist with alarming frequency in extreme mobility. Table 4.2 summarizes two-cell conflicts from our dataset. Note policy conflicts can also happen with >2 cells, so this result is a lower bound of conflicts in reality. On average, two-cell policy conflicts occur every 194.6–1090s in high-speed rails (3.8×–26.2× more than low mobility), each incurring 3.0–3.9 handovers on average. Surprisingly, intra-frequency policy conflicts (A3-A3) are much more than static or low-speed mobility [LDL16, YLL18], and dominate the policy conflicts in extreme mobility (55.9%–100%). To trigger handovers early with less failures, the operators configure a proactive policy among cells (Figure 4.4a with $\Theta_{A3} < 0$). Such policy causes oscillations and voids the efforts of failure mitigation.

4.1.3 Execution: Unreliable Signaling

5G/4G can also fail if the serving cell cannot deliver handover command to the client. Similar to feedback loss in §4.1.1, such unreliable signaling mainly arises from the wireless dynamics in extreme mobility. It can also come from failure propagation of slow feedback in triggering (§4.1.1) and multi-stage policy in decision (§4.1.2).

Validation: Table 4.1 shows 19.2%–31.5% of network failures arises from the handover command loss. We detect these failures by observing successful delivery of feedback that can trigger handovers based on inferred policy (e.g., Figure 4.1), but no handover command from

serving cell until the client loses network access. We also observe high physical-layer block errors when such failure occurs. Figure 4.2b shows block error rate within 5 seconds before network failures. The average block error rate is 30.3% for downlink (handover command) and 9.9% for uplink (measurement feedback). This implies the signaling is corrupted during the delivery, thus failing to execute the handovers and losing network access.

4.1.4 Implications for 5G

The emergent 5G standards [3GP20, 3GP19h, 3GP19e] offer various new features that 4G LTE lacks, such as the dense small cells, new radio bands (sub-6GHz and above-20GHz), renovated physical layer design, and advanced signaling protocols. Since 2019, 5G has been under active testing and deployment on the high-speed rails [ZTE19, Chi20]. While our empirical results in §4.1.1–§4.1.3 are from 4G LTE, we note reliable extreme mobility in 5G will be even more challenging because (1) 5G handovers [3GP19h] follow the same design as 4G [3GP15]; (2) 5G adopts small dense cells under high carrier frequency, which incurs more frequent handovers that are more prone to Doppler shifts (§2.3) and failures; (3) while 5G refines its physical layers (e.g., Polar code and more reference signals [3GP19e]) to improve the reliability, they are still based on OFDM and suffers from similar issues.

4.1.5 Problem Statement

This work aims at reliable extreme mobility management in 5G/4G and beyond. We seek a solution with significantly less network failures, verifiable conflict-free policies, and negligible latency/signaling/spectral overhead. The solution should be reliable with dramatic wireless dynamics in extreme mobility, during which it may experience errors, delays, and failures in all phases of mobility management. The solution should be backward-compatible with existing OFDM-based 5G/4G (especially data transfer) in static and low mobility, and retain flexible policy for the operators.

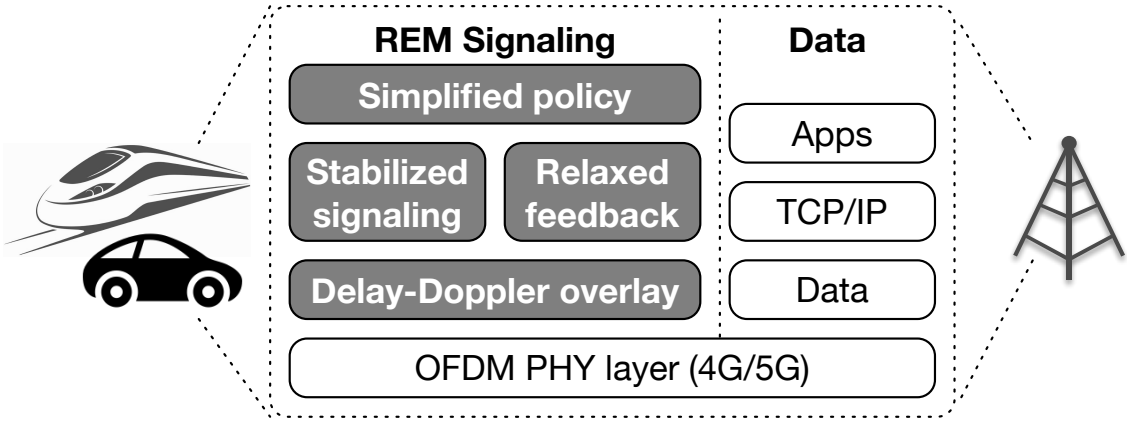


Figure 4.5: REM overview.

4.2 Intuitions Behind REM

We devise REM, **R**eliable **E**xtreme **M**obility management to achieve all the goals in §4.1.5. Our key insight is that, extreme mobility is unreliable because of *wireless signal strength-based* management today. In extreme mobility, wireless signal strength is unreliable with Doppler shift and multipath fading (§2.3). This propagates failures to all phases of mobility management, i.e., sluggish feedback in triggering (§4.1.1), policy conflicts in decision (§4.1.2), and signaling loss/error in execution (§4.1.3). To achieve reliable extreme mobility, a fundamental solution is to shift to more dependable criteria.

Therefore, REM shifts from indirect wireless signal strength-based to direct *movement-based* mobility. Intuitively, the client movement decides its physical multi-paths and Doppler effect for each cell, thus impacting the wireless quality. Compared to wireless with short coherence and dramatic dynamics (§2.3), the client movement is slower and predicable by inertia, thus more reliable to drive the extreme mobility management. To this end, REM tracks the client movement in the *delay-Doppler domain*. With this knowledge, REM relaxes the feedback's exploration-exploitation dilemma in triggering phase, simplifies the policies in decision phase, and stabilizes the signaling traffic in execution phase.

Delay-Doppler domain: A wireless channel decides how radio signals from the sender

propagates along multiple physical paths, and combines at the receiver. A time-varying channel can be characterized in multiple ways. 5G/4G measures its OFDM channel in the time-frequency domain: An OFDM channel is defined as a function of time and carrier frequency $H(t, f)$. Equivalently, we can represent the same channel in the *delay-Doppler domain* [Bel63]:

$$h(\tau, \nu) = \sum_{p=1}^{P} h_p \delta(\tau - \tau_p)\delta(\nu - \nu_p) \tag{4.1}$$

where P is the number of paths (direct, reflected, and scattered ones), h_p, τ_p, ν_p are p-th path's complex attenuation, propagation delay (distance) and Doppler frequency shift, and δ is the Dirac delta function. Figure 4.6a exemplifies a channel with 3 paths. The delay-Doppler form reflects the *multi-path geometry* between cell and client in movement. Given $h(\tau, \nu)$ and a sent signal $s(t)$, the received signal $r(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau, \nu)s(t - \tau)e^{j2\pi\nu t}d\tau d\nu$. The OFDM channel $H(t, f)$ and delay-Doppler channel $h(\tau, \nu)$ are related by

$$H(t, f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau, \nu)e^{j2\pi(t\nu - f\tau)}d\tau d\nu = \sum_{p=1}^{P} h_p e^{j2\pi(t\nu_p - f\tau_p)}$$

Compared to $H(t, f)$, delay-Doppler representation $h(\tau, \nu)$ is more stable since its variance relates to slower path delay and Doppler change [HRT17, MHT16, HM18] (see Appendix A.1 for an analysis).

Why delay-Doppler domain: The delay-Doppler domain unveils client movement and multi-path propagation $\{h_p, \tau_p, \nu_p\}$. Mobility management on top of it can benefit in all its phases:

• *Triggering: Relaxed reliance on feedback.* Movement-based feedback allows fast and reliable triggering with relaxed exploration-exploitation (more measurements v.s. timely triggering) tradeoff. Cells from the same base station share the physical propagation paths to the client. Instead of measuring all cells sequentially, the client only measures one cell and performs *cross-band estimation* to others from the same location. This accelerates the feedback without reducing the cells to be explored.

• *Decision: Simplified, conflict-free policy.* The decision policy in the delay-Doppler domain can be simplified for two reasons. First, by replacing the inter-frequency measurement with cross-band estimation, the tradeoff between decision latency and spectral efficiency is bypassed. This eliminates the need for multi-stage policy (§4.1.2). Second, it reduces configurations (A1, A2, A4, A5) for heterogeneous cells that share the multipath, thus reducing the conflicts.

• *Execution: Stabilized signaling.* Similar to 5G/4G OFDM, we can represent, modulate, and transfer signals in the delay-Doppler domain. Compared to OFDM, the delay-Doppler signal transfer is directly coupled with the slowly-varying multi-path evolution. So it will exploit the full time-frequency diversity, and therefore experience more stable channels and less loss/corruption. This mitigates failures from signaling/feedback loss or corruption.

REM roadmap: REM devises a signaling overlay in delay-Doppler domain with the recently proposed OTFS modulation [HRT17]. REM further greatly extends OTFS to refine all phases of mobility management. Figure 4.5 overviews REM's main components.

• **Delay-Doppler signaling overlay (§4.3.1):** REM places the signaling traffic and reference signals in an delay-Doppler domain overlay. This overlay runs on top of existing OFDM, without changing 5G/4G designs or data traffic. It stabilizes the signaling in execution (§4.1.3), and exposes movement information to later phases.

• **Relaxed reliance on feedback (§4.3.2):** To mitigate the failures from slow and unreliable feedback (§4.1.1), REM devises cross-band estimation in the delay-Doppler domain. This approach accelerates the feedback without reducing the cells to be explored, and facilitate earlier handovers with less failures.

• **Simplified, conflict-free policy (§4.3.3):** To eliminate policy conflicts and failures from missed cells (§4.1.2), REM simplifies the policy in the delay-Doppler domain. It eliminates the multi-stage decision with cross-band estimation, reduces the configurations, and enables easy-to-satisfy conditions for the conflict-freedom.

4.3 The REM Design

We next elaborate each component in REM.

4.3.1 Delay-Doppler Signaling Overlay

REM runs its mobility management in delay-Doppler domain. To achieve so, REM should place its signaling traffic (e.g., measurement feedback, handover commands, reference signals) and modules (triggering, decision, execution) in this domain. We prefer to do so without changing existing 5G/4G designs or affecting OFDM-based data transfer. To this end, REM leverages recent advances in OTFS in delay-Doppler domain, builds a signaling overlay atop OFDM, extends OTFS with adaptive scheduling to enable the co-existence of OTFS signaling and OFDM data, and uses it to mitigates failures from signaling loss/corruption in execution (§4.1.3).

Delay-Doppler overlay with OTFS: OTFS is a modulation in the delay-Doppler domain. Intuitively, OTFS couples information with the multi-path geometry, modulates signals in the delay-Doppler domain, and multiplexes signals across all the available carrier frequencies and time slots. By exploiting full time-frequency diversity, signals enjoy similar channels with less variance, become robust to Doppler shifts and less vulnerable to loss and errors.

Figure 4.6a shows the OTFS modulation. It runs on top of OFDM. The OFDM time-frequency domain is discretized to a $M \times N$ grid (each being a 5G/4G radio resource element) by sampling time and frequency axes at intervals T and Δf[5], respectively. The modulated OFDM samples $X[n, m]$ are transmitted for a duration of NT and bandwidth of $M\Delta f$. Given a $M \times N$ time-frequency domain, the delay-Doppler domain is also a $M \times N$ grid

[5]In 4G OFDM, $T = 66.7\mu s$, $\Delta f = 15$KHz [3GP17c]. In 5G OFDM, T can be $4.2, 8.3, 16.7, 33.3$ or $66.7\mu s$ and Δf can be 15, 30, 60, 120 or 240KHz [3GP19e].

$(\frac{k}{M\Delta f}, \frac{l}{NT})$, where $k = 0..M-1, l = 0..N-1$ where $\frac{1}{M\Delta f}$ and $\frac{1}{NT}$ are the quantization steps of path delay and Doppler frequency, respectively. The OTFS modulator arranges MN data symbols in the delay-Doppler grid, denoted as $x[k,l]$. It then converts $x[k,l]$ to $X[n,m]$ in OFDM using the discrete Symplectic Fourier transform (SFFT)

$$X[n,m] = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} x[k,l]e^{-j2\pi(\frac{mk}{M}-\frac{nl}{N})} \quad \text{(SFFT)} \tag{4.2}$$

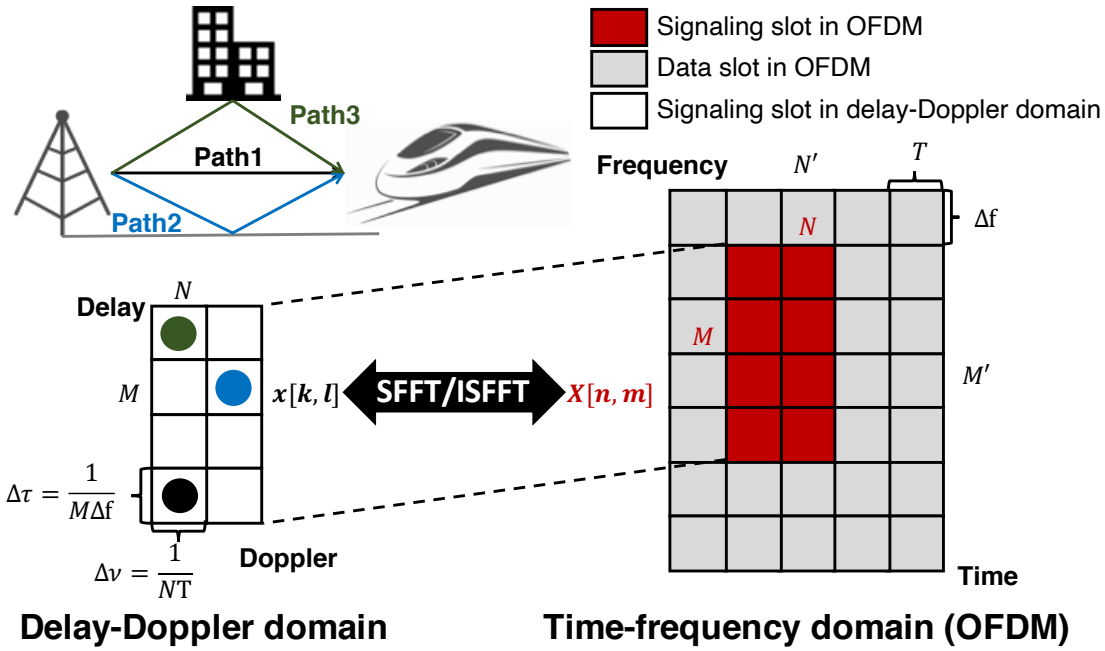
$$x[k,l] = \frac{1}{NM}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X[n,m]e^{j2\pi(\frac{mk}{M}-\frac{nl}{N})} \quad \text{(ISFFT)} \tag{4.3}$$

The OFDM signal $X[n,m]$ is transmitted via legacy 5G/4G radio. The received signal $Y[n,m]$ is in the time-frequency domain. Then inverse SFFT (ISFFT) in (4.3) is applied to $Y[n,m]$ and yields $y[k,l]$ in the delay-Doppler domain. With channel noises, we have [HRT17, RPH18]

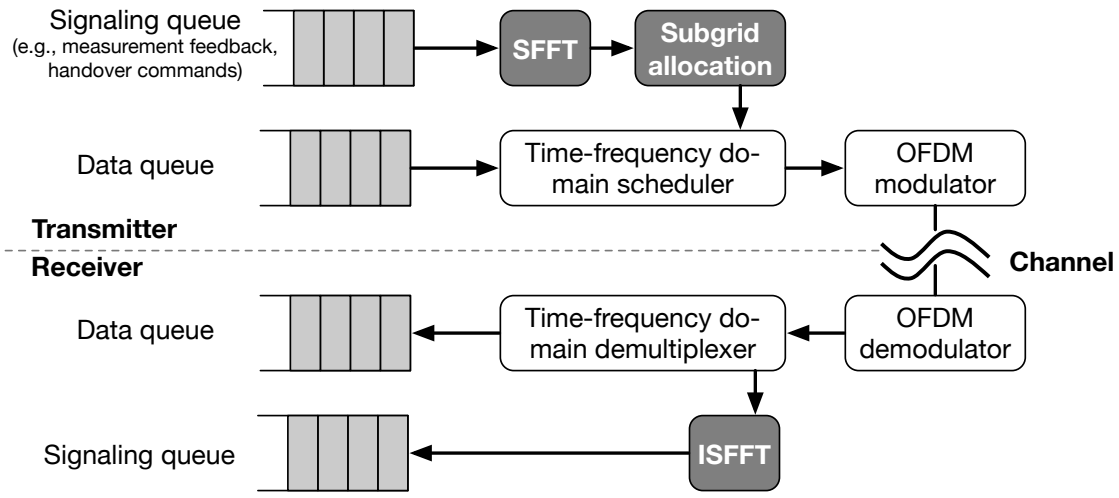
$$y[k,l] = \frac{1}{NM}\sum_{k'=0}^{M-1}\sum_{l'=0}^{N-1} h_w(k'\Delta\tau, l'\Delta\nu)x[k-k', l-l'] + n[k,l] \tag{4.4}$$

where $h_w(\tau,\nu) = \int\int e^{-j2\pi\tau'\nu'}h(\tau',\nu')w(\nu-\nu',\tau-\tau')d\tau' d\nu'$ is the convolution of channel $h(\tau',\nu')$ and rectangular signal window: $w(\tau,\nu) = \sum_{c=0}^{N-1}\sum_{d=0}^{M-1} e^{-j2\pi(\nu cT-\tau d\Delta f)}$, $n(k,l) = ISSFT(N[n,m])$ is ISFFT of time-frequency noises. Compared to OFDM channel $H(t,f)$ with short coherence T_c, the delay-Doppler channel $h_w(\tau,\nu)$ is invariant of multi-path fading or inter-carrier interference from Doppler shift, thus more stable and reliable in a longer period.

Challenge: Coexistence with OFDM data REM only adopts delay-Doppler domain for its signaling traffic. We are neutral to if data traffic should also use OTFS. While OTFS can help data combat Doppler shifts, it also incurs more data processing delays and may not be preferred by latency-sensitive scenarios. Instead, REM supports hybrid mode between OTFS-based signaling and OFDM/OTFS-based data. It offers flexibility for operators with both choices.



(a) Subgrid allocation for signaling in delay-Doppler domain



(b) Realization (gray modules) on top of OFDM

Figure 4.6: Signaling overlay in delay-Doppler domain.

The challenge for this hybrid mode is that, to function correctly, OTFS requires a continuous $M \times N$ OFDM grid. But in 5G/4G, the signaling and data traffic are multiplexed in the OFDM grid. In case data still uses OFDM, the signaling traffic may span on *disjoint* OFDM slots, and cannot run OTFS directly. A possible solution is to define separated data and signaling grids, which however may waste the radio resource and needs 5G/4G physical layer redesign.

Our solution: Scheduling-based OTFS To address this, we note *the 5G/4G signaling traffic is always prioritized in scheduling and delivery by design* [3GP15, 3GP19h]. Before successful signaling procedures, the data traffic may not be correctly delivered or processed. So given pending signaling traffic, the base station will always schedule the radio resource and deliver the signaling traffic first, regardless of if any data is waiting. REM leverages this readily-available feature to allocate a *sub-grid* for OTFS-based signaling traffic first. It decouples OTFS-based signaling and OFDM-based data for co-existance, without changing the 5G/4G design or adding delay/spectral cost.

Figure 4.6b illustrates REM's ultimate signaling overlay. At the transmitter (base station for downlink and client for uplink), the overlay modulates the signaling traffic and reference signals with SFFT, and forwards them to the signaling radio bearer for traffic scheduling. Given the signaling traffic, the scheduler will always process them first by design. To ensure the applicability of OTFS, REM adapts the scheduler to guarantee that, all signaling traffic is always placed in a $M \times N$ subgrid of the 5G/4G resource grid ($M \leq M', N \leq N'$). On receiving these signaling, the receiver demodulates them in OFDM, runs REM's overlay to further demodulate in OTFS, and then forwards to upper layer for further mobility actions.

Overhead for signaling: REM adds the SFFT/ISFFT to pre/post-process the signaling traffic, with the complexity of $O(MNlog(MN))$. Such complexity is similar to 5G/4G uplink's SC-FDMA on top of OFDM (with additional fast Fourier transform). No additional delays, spectral waste or other overhead is incurred for the data traffic.

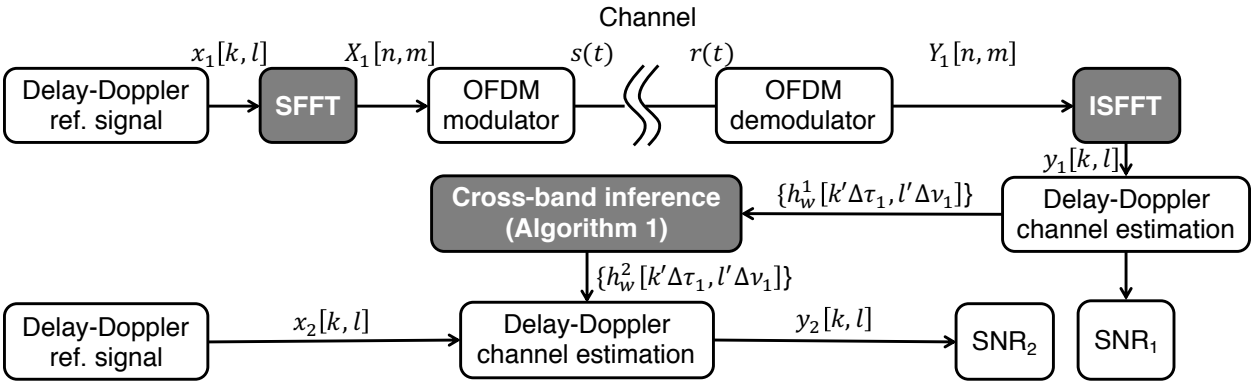


Figure 4.7: REM's cross-band channel estimation. Gray boxes are additional modules to OFDM today.

4.3.2 Relaxed Reliance on Feedback

With the delay-Doppler overlay, REM relaxes the handover's reliance on the feedback for fast and satisfactory triggering (§4.1.1). To achieve so, the key is to relax the unique dilemma in extreme mobility, between *exploration* of more measurements for satisfactory triggering and *exploitation* for fast triggering. We observe that, cells from the same base station share the multi-paths to the client and thus similar channels in the delay-Doppler domain (§4.1.1). To this end, REM devises *cross-band estimation* to parallelize the feedback: It measures *one only* cell per base station, extracts the multi-path profile from this measurement, maps it to other cells from the same base station, and estimates these cells' qualities *without* measurements. This allows the serving cell to make decisions *without* waiting for all feedback and triggering intervals in §4.1.1.

Existing cross-band estimations: Cross-band estimation is recently proposed in [Kal10, Vas16, Bak19] to save the channel feedback overhead. Existing solutions are designed in the time-frequency domain and primarily for static scenarios. The idea is to extract the multi-path profiles (path delay, attenuation, phase, etc.) from one band's channel estimation, and map it to another band traversing the same paths. In the time-frequency domain,

the channel $H(f,t)$ differs among frequency bands, and does not reveal path parameters. So [Vas16, Bak19] estimate the multi-path profile with non-linear optimization or machine learning. Unfortunately, these approaches face two fundamental limitations in extreme mobility. First, they do not consider the Doppler effect in mobility. Second, their optimization and machine learning are too slow to track the fast-varying channel dynamics (§4.5.2). The hardware acceleration with GPU, FPGA, or multi-core CPU could help. But such hardware is too expensive for the resource and energy-constrained mobile devices.

REM's intuition: To overcome these limitations, REM generalizes and simplifies the cross-band estimation in the delay-Doppler domain. Compared to the time-frequency domain representation $H(t,f)$, the delay-Doppler domain representation $h(\tau,\nu)$ in Equation (4.1) directly unveils the multi-path profiles $\{h_p,\tau_p,\nu_p\}$ and is more feasible for cross-band estimation. Besides, $h(\tau,\nu)$ evolves slower than $H(t,f)$ (§4.2), thus reducing frequent feedback and facilitating shorter triggering interval. With the delay-Doppler domain, REM can tackle the Doppler shift in extreme mobility, and eliminates the optimization and machine learning in existing solutions.

Specifically, consider two cells from the same base station. Given cell 1's channel estimation $\{h_w^1(k\Delta\tau,l\Delta\nu)\}_{k,l}$, REM estimates cell 2's channel $\{h_w^2(k\Delta\tau,l\Delta\nu)\}_{k,l}$ *without* measuring it. To do so, REM first extracts multi-path profile $\{h_p,\tau_p,\nu_p^1\}$ from cell 1 $\{h_w^1(k\Delta\tau,l\Delta\nu)\}_{k,l}$. Note that the path delays τ_p and attenuations h_p are *frequency-independent*, thus identical for cell 1 and 2. The Doppler shifts of cell 1 ν_p^1 and cell 2 ν_p^2 are *frequency-dependent* and $\nu_p^1 \neq \nu_p^2$. But they are correlated by $\nu_p^1/\nu_p^2 = f_1/f_2$ (§2.3). So with cell 1's multi-path profile, we can estimate cell 2 by reusing $\{h_p,\tau_p\}$ and deriving $\{\nu_p^2\}$ from ν_p^1.

REM's cross-band estimation: REM first estimates cell 1's channel in the delay-Doppler domain. With its signaling overlay (§4.3.1), REM reuses 5G/4G's reference signals[6] but pre/post-process them in the delay-Doppler domain (Figure 4.7). By comparing re-

[6]The cell-specific reference signals in 4G LTE, and CSI-RS in 5G NR [3GP19e]. Both are decoupled from demodulation reference signals for data transfer.

ceived and constant sent reference signal $(y(k,l), x(k,l))$, we can estimate the delay-Doppler channel $\{h_w(k\Delta\tau, l\Delta\nu)\}_{k,l}$ by applying standard channel estimation [MAT18] to OTFS's input-output relation in (4.4).

Now consider two cells from the same base station. Given cell 1's channel estimation $\{h_w^1(k\Delta\tau, l\Delta\nu)\}_{k,l}$, REM estimates cell 2's channel $\{h_w^2(k\Delta\tau, l\Delta\nu)\}_{k,l}$. We note channel estimation in (4.4) has

$$\frac{1}{MN} h_w(k\Delta\tau, l\Delta\nu) = \sum_{p=1}^{P} \frac{\Gamma(k\Delta\tau, \tau_p)}{M} \cdot h_p e^{-j2\pi\tau_p\nu_p} \cdot \frac{\Phi(l\Delta\nu, \nu_p)}{N} \tag{4.5}$$

where we have $\Gamma(k\Delta\tau, \tau_p) = \sum_{d=0}^{M-1} e^{j2\pi(k\Delta\tau - \tau_p)d\Delta f}$, $\Phi(l\Delta\nu, \nu_p) = \sum_{c=0}^{N-1} e^{-j2\pi(l\Delta\nu - \nu_p)cT}$. We can rewrite it in a matrix form:

$$\mathbf{H} = \mathbf{\Gamma P \Phi} \tag{4.6}$$

where $\mathbf{H} \in \mathbb{C}^{M\times N}$ is the channel estimation matrix from (4.4): $H(k,l) = \frac{1}{MN} h_w(k\Delta\tau, l\Delta\nu)$.

$$\mathbf{H} = \frac{1}{MN} \begin{bmatrix} h_w(0,0) & \cdots & h_w(0,(N-1)\Delta\nu) \\ h_w(\Delta\tau,0) & \cdots & h_w(\Delta\tau,(N-1)\Delta\nu) \\ \cdots & \cdots & \cdots \\ h_w((M-1)\Delta\tau,0) & \cdots & h_w((M-1)\Delta\tau,(N-1)\Delta\nu) \end{bmatrix}$$

$\mathbf{\Gamma} \in \mathbb{C}^{M\times P}$ is the *frequency-independent* path delay spread matrix from Equation 4.5: $\Gamma(k,p) = \frac{\Gamma(k\Delta\tau,\tau_p)}{M}$,

$$\mathbf{\Gamma} = \frac{1}{M} \begin{bmatrix} \Gamma(0,\tau_1) & \Gamma(0,\tau_2) & \cdots & \Gamma(0,\tau_P) \\ \Gamma(\Delta\tau,\tau_1) & \Gamma(\Delta\tau,\tau_2) & \cdots & \Gamma(\Delta\tau,\tau_P) \\ \cdots & \cdots & \cdots & \cdots \\ \Gamma((M-1)\Delta\tau,\tau_1) & \Gamma((M-1)\Delta\tau,\tau_2) & \cdots & \Gamma((M-1)\Delta\tau,\tau_P) \end{bmatrix}$$

$\mathbf{\Phi} \in \mathbb{C}^{P\times N}$ is the *frequency-dependent* path Doppler spread matrix with $\Phi(p,l) = \frac{\Phi(l\Delta\nu,\nu_p)e^{-j(\theta_p + 2\pi\tau_P\nu_P)}}{N}$,

θ_p is the frequency-independent path phase: $h_p = |h_p|e^{-j\theta_p}$.

$$\mathbf{\Phi} = \frac{1}{N}\begin{bmatrix} \Phi(0,\nu_1)e^{-j(\theta_1+2\pi\tau_1\nu_1)} & \cdots & \Phi((N-1)\Delta\nu,\nu_1)e^{-j(\theta_1+2\pi\tau_1\nu_1)} \\ \Phi(0,\nu_2)e^{-j(\theta_2+2\pi\tau_2\nu_2)} & \cdots & \Phi((N-1)\Delta\nu,\nu_2)e^{-j(\theta_2+2\pi\tau_2\nu_2)} \\ \cdots & \cdots & \cdots \\ \Phi(0,\nu_P)e^{-j(\theta_P+2\pi\tau_P\nu_P)} & \cdots & \Phi((N-1)\Delta\nu,\nu_P)e^{-j(\theta_P+2\pi\tau_P\nu_P)} \end{bmatrix}$$

and $\mathbf{P} \in \mathbb{R}_{\geq 0}^{P\times P}$ is the *multi-path attenuation* diagonal matrix:

$$\mathbf{P} = \begin{bmatrix} |h_1| & 0 & \cdots & 0 \\ 0 & |h_2| & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & |h_P| \end{bmatrix}$$

Given the cell 1's channel estimation matrix $\mathbf{H_1}$, if we can decompose it as $\mathbf{H_1} = \mathbf{\Gamma P \Phi_1}$, then the frequency-independent path delay $\mathbf{\Gamma}$ and attenuation $\mathbf{P}$ can be directly reused by cell 2, while the frequency-dependent Doppler shift $\mathbf{\Phi_2}$ can be derived from $\mathbf{\Phi_1}$ since $\frac{\nu_p^1}{\nu_p^2} = \frac{f_1}{f_2}$. Then we can obtain cell 2's channel $\mathbf{H_2} = \mathbf{\Gamma P \Phi_2}$.

So how to decompose the delay-Doppler channel matrix $\mathbf{H_1} = \mathbf{\Gamma P \Phi_1}$? It turns out that, such decomposition can be approximated by the classical singular value decomposition (SVD) [Sin19]. Recall that SVD can factorize *any* matrix $\mathbf{H} \in \mathbb{C}^{M\times N}$ into two unitary matrices and a diagonal matrix: $\mathbf{H} = \mathbf{U\Sigma V}$, where $\mathbf{U} \in \mathbb{C}^{M\times M}$ is a unitary matrix with $\mathbf{UU}^* = \mathbf{I_M}$, $\mathbf{V} \in \mathbb{C}^{N\times N}$ is a unitary matrix with $\mathbf{VV}^* = \mathbf{I_N}$, and $\mathbf{\Sigma} \in \mathbb{R}_{\geq 0}^{M\times N}$ is a diagonal matrix with non-negative real numbers on the diagonal (i.e., singular values). Intuitively, SVD factorizes a matrix into two orthonormal bases $\mathbf{U}$ (for each row) and $\mathbf{V}$ (for each column), and attenuation $\mathbf{\Sigma}$. In practice, to reduce matrix dimensionality, SVD typically keeps the major singular values ("principle components") and truncate negligible ones. In this way, SVD approximates a matrix as $\mathbf{H} \approx \mathbf{U'\Sigma'V'}$ where $\mathbf{U'} \in \mathbb{C}^{M\times P}$, $\mathbf{\Sigma'} \in \mathbb{C}^{P\times P}$, and $\mathbf{V'} \in \mathbb{C}^{P\times N}$ with smaller matrix dimension $P \leq \min(M,N)$. This form is the same as our delay-Doppler channel decomposition $\mathbf{H} = \mathbf{\Gamma P \Phi}$. In fact, we can prove their relation as follows (proved in Appendix A.2):

Theorem 4.3.1 (Cross-band estimation with SVD)**.** *A delay-Doppler decomposition* $\mathbf{H} = \mathbf{\Gamma P \Phi}$ *is also a singular value decomposition if (i) the number of physical paths* $P \leq \min(M, N)$*; and (ii) for any two paths* $p \neq p'$*, we always have* $\tau_p - \tau_{p'} = k\Delta\tau$ *and* $\nu_p - \nu_{p'} = l\Delta\nu$ *for some non-zero integer* k, l*.*

In reality, we note condition (i) almost always holds. It has been shown the real 4G/5G channels have sparse multi-paths[JBK15, WZZ17, CHO04][7]. Condition (ii) also approximately holds in reality: With 40ms triggering interval for a 20MHz channel (§4.1.1), $(M, N) = (1200, 560)$ and the wavelength is $c/f \approx 15$m. In the high-speed rails, the line-of-sight distance between the base station and the train is approximately multiple times of 15m (typically between 80m and 550m [Tan14]). The non-line-of-sight reflection/scattering propagation paths are even longer. So such (M, N) results in fine-grained delay/Doppler sampling $(\Delta\tau, \Delta\nu)$ and approximates condition (ii).

Algorithm 1 shows REM's cross-band estimation via SVD. Given cell 1's channel estimation matrix $\mathbf{H_1}$, we run SVD and use it as an approximation of $\mathbf{H_1} = \mathbf{\Gamma P \Phi_1}$ (line 1). Note cell 1's $\mathbf{\Gamma P}$ is frequency-independent and can be reused by cell 2. To estimate cell 2, we need to infer $\mathbf{\Phi_2}$ from $\mathbf{\Phi_1}$. To this end, Algorithm 1 estimates multi-path profile $\{h_p, \tau_p, \nu_p^2\}_{p=1}^{P_{max}}$ (line 2–8) based on the derivations in Appendix A.3. Then Algorithm 1 re-constructs $\mathbf{\Phi_2}$ and estimates cell 2 as $\mathbf{H_2} = \mathbf{\Gamma P \Phi_2}$. Algorithm 1 supports multi-antenna systems such as MIMO and beamforming, by running it on each antenna.

Complexity: REM's runs SFFT/ISFFT to process the reference signals and Algorithm 1 for cross-band estimation. Both have polynomial complexity: The SFFT/ISFFT complexity is $O(MN \log MN)$, and Algorithm 1's complexity is $O(\min(M, N) \max(M, N)^2)$. It is faster than [Vas16, Bak19] that rely on optimization or machine learning, thus suitable to track the fast-varying channel in extreme mobility.

[7]In 5G/4G, even the smallest OFDM resource block has $M = 12, N = 14$ and thus can support up to 12 paths. This suffices for standard reference multi-path models in 4G (7–9 paths depending on the scenario [3GP19b]) and 5G (12 paths [3GP19d]).

Algorithm 1 REM's cross-band channel estimation

Input: Band 1's channel estimation matrix $\mathbf{H_1}$, $H_1(k,l) = h_w^1(k\Delta\tau, l\Delta\nu)$ from (4.4)

Output: Band 2's channel estimation matrix $\mathbf{H_2}$

1: Decompose $\mathbf{H_1} = \mathbf{\Gamma P \Phi_1}$ using SVD matrix factorization;

2: **for** each path $p = 1, 2, ... \min(M, N)$ **do**

3: For any $\forall l, l' \neq l \in [0, N-1]$ and $\forall k, k' \neq k \in [0, M-1]$;

4: $\nu_p^1 \leftarrow e^{-j2\pi\nu_p^1 T} = \frac{1}{N(N-1)} \sum_{l,l'} \frac{\Phi_1(p,l) - \Phi_1(p,l')}{\Phi_1(p,l)e^{j2\pi l\Delta\nu T} - \Phi_1(p,l')e^{j2\pi l'\Delta\nu T}}$;

5: $\tau_p \leftarrow e^{j2\pi\tau_p\Delta f} = \frac{1}{M(M-1)} \sum_{k,k'} \frac{\Gamma(k,p) - \Gamma(k',p)}{\Gamma(k,p)e^{-j2\pi k\Delta\tau\Delta f} - \Gamma(k',p)e^{-j2\pi k'\Delta\tau\Delta f}}$;

6: $\nu_p^2 \leftarrow \nu_p^1 \frac{f_2}{f_1}$; ▷ *Transfer to band 2's Doppler frequency*

7: $e^{-j\theta_p} \leftarrow \frac{1}{N} \sum_l \frac{\Phi(p,l)N}{\Phi(l\Delta\nu, \nu_p)e^{-j2\pi\tau_p\nu_p}}$;

8: **end for**

9: Compute $\mathbf{\Phi_2}$ with $\{h_p, \tau_p, \nu_p^2\}_p$;

10: $\mathbf{H_2} \leftarrow \mathbf{\Gamma P \Phi_2}$;

The impact of channel noises: The noises impacts channel estimation accuracy and indirectly affects cross-band estimation. REM is robust to noises since it runs in the delay-Doppler domain. According to (4.4), the noise in the time-frequency domain $N[n,m]$ is smoothed to $n[k,l]$ in the delay-Doppler domain via IFFT. For typical 5G/4G noises, this results in more robust channel estimation for h_w and thus decomposition. REM may be less robust if the OFDM noises are carefully crafted (e.g., spamming attack), so that the channel estimation is inaccurate. Both OFDM and REM would be affected then, and REM is no worse than OFDM in terms of reliability.

4.3.3 Simplified, Conflict-Free Policy

REM last simplifies the handover policy for high reliability and verifiable correctness (§4.1.2). Our goal is to: (1) avoid multi-stage policy whenever possible, without missing cells or delaying handovers; and (2) eliminate policy conflicts in extreme mobility. Meanwhile, REM still retains flexibility for operators to customize their policies.

Extreme mobility policy in delay-Doppler domain: Compared to the complex policy today, extreme mobility policy in delay-Doppler domain can be simplified for three reasons:

(1) Bypassed the latency-spectral efficiency tradeoff: As shown in §4.1.2, multi-stage policy is common today to balance the spectral efficiency and decision latency for inter-frequency cells. This is mostly unnecessary with REM's cross-band estimation in §4.3.2. Inter-frequency cells can be inferred from intra-frequency cells at the location, *without* extra round trips or allocating radio resource.

(2) Coherent, stable decision metric: Delay-Doppler domain enables more stable channel and signal-noise-ratio (SNR), and makes SNR-based handover feasible[8]. This benefits policy simplification with less events (Table 2.1). In signal strength-based 5G/4G mobility, A4 is used for load balancing and A5 is for indirect signal strength comparison between heterogeneous cells (§4.1.2). These events are not "must-haves" if SNR is used, since SNRs between cells are directly comparable and decide the capacity $C = B\log(\mathrm{SNR}+1)$ (B is the bandwidth) based on information theory.

(3) Reduced demand for proactive policies: In extreme mobility, the policy conflicts are amplified by operators' demand for proactive failure mitigation (§4.1.2). In delay-Doppler domain, this demand can be satisfied by REM instead (§4.3.1–§4.3.2), thus eliminating the need for conflict-prone proactive policies.

REM's simplification approach: Figure 4.8 exemplifies how REM simplifies an extreme mobility policy today in four steps:

(1) Replace received signal strength with delay-Doppler SNR. This helps stabilize the input and simplifies events needed. Note SNR should always be evaluated in handover, regardless of other metrics to be used. Otherwise, "blind handovers" will *always* happen with loops

[8]In theory, 5G/4G OFDM could also use SNR for handover. But this is rare (if not non-existent) since OFDM SNR fluctuates rapidly and causes frequent oscillations (§4.1.1). Instead, 5G/4G decides handover using stabler signal strength [3GP15, 3GP19h, DPF18b, LPY16, Hua16, zte].

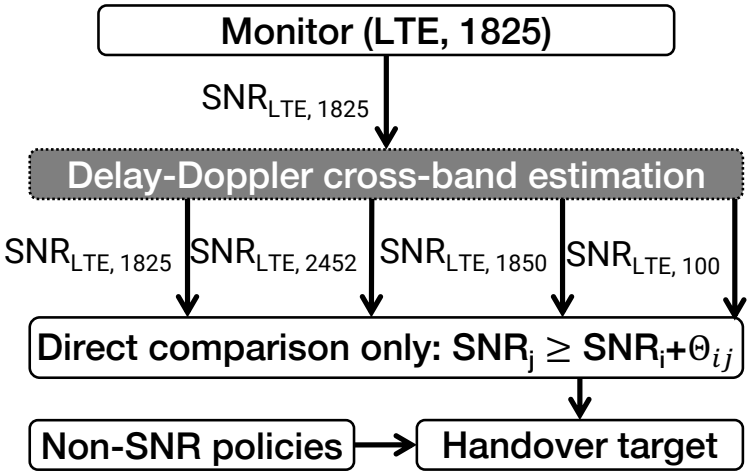


Figure 4.8: REM's policy simplification for Figure 4.1.

[LDL16], and lose network access if target cell's coverage is weak;

(2) Replace multi-stage policy with cross-band estimation. If inter-frequency cells are co-located with intra-frequency ones, REM replaces A1/A2-based multi-stage policy with cross-band estimation in §4.3.2. This avoids missing cells and bypasses the tradeoff between latency and spectral cost for inter-frequency cells. Otherwise, REM retains the multi-stage policy and moves to next step (but still with the same conflict-freedom guarantees below).

(3) Remove unnecessary events in policy. By removing the multi-stage decision, A1/A2 events are removed. For other events, REM replaces them with A3. For each A5 event today, REM replaces it with an equivalent A3 with $\Theta_{A3} = \Theta_{A5,2} - \Theta_{A5,1}$, since A5 $R_s < \Theta_{A5,1}, R_n > \Theta_{A5,2}$ implies $R_n > R_s + \Theta_{A5,2} - \Theta_{A5,1}$. To remove A4, there are two cases in extreme mobility. First, due to multi-stage policy, most A4 events occur after A2 is triggered. They are equivalent to A5 with $\Theta_{A5,1} = \Theta_{A2}, \Theta_{A5,2} = \Theta_{A4}$ and replaced by A3 with above procedure. Second, for load balancing or adding capacity [DPF18b, LPY16, Hua16], a small amount of A4 events are directly triggered without A2 (§4.1.2). They can also be replaced by A3: The serving cell can equally find a cell with less load or more capacity using A3 comparison on $C = Blog(SNR + 1)$, where Θ_{A3} decides capacity difference. Afterwards, REM only regulates A3 for conflict freedom as detailed below.

(4) Retain remaining policies: A cell may decide handovers based on other metrics, such as priorities, traffic load, and access control. REM keeps them without changes, and retains flexibility for operators.

REM's simple conflict-freedom guarantees: Compared to today's policies in §4.1.2, REM eliminates most events except A3. This leads to less conflicts between events, and simpler conflict resolutions than [LDL16, YLL18]. We start with the policy with delay-Doppler SNR only. We obtain the following result (proved in Appendix A.4):

Theorem 4.3.2 (Conflict-freedom with delay-Doppler SNR only)**.** *When only delay-Doppler SNR is used in* REM*'s simplfied policy, no persistent loops will occur if among any cells c_i, c_j and c_k ($j \neq i, k$, but i can be equal to k) that cover the same area, $\Theta_{A3}^{i\to j} + \Theta_{A3}^{j\to k} \geq 0$.*

Theorem 4.3.2 shows at most 3-cell threshold coordination is sufficient for policy conflict freedom. Compared to the conflict freedom conditions today [LDL16, YLL18], Theorem 4.3.2 is much simpler with less events and threshold coordination between cells. Violation of Theorem 4.3.2 happens in extreme mobility when operator tries proactive handovers to mitigate failures (§4.1.2). With REM, operators do not need this since REM has mitigated most failures.

We next show that, even with other criteria (preferences, load balancing, access control, etc.), Theorem 4.3.2 is still sufficient for conflict freedom.

Theorem 4.3.3 (Conflict-freedom in general)**.** *For any settings of non-SNR metrics in* REM*, satisfying Theorem 4.3.2 still guarantees loop-freedom.*

Theorem 4.3.3 is proved in Appendix A.5. Intuitively, with coordinated SNR events, Theorem 4.3.2 ensures handovers between cells will not be simultaneously satisfied. Regardless of other policies, this condition suffices for conflict freedom. This simplifies the policy configurations with provable conflict freedom.

4.4 Implementation

We implement REM on Ettus USRP software-defined radio running OpenAirInterface [ope19] software cellular stack, with one emulating a client and another as a base station. REM is realized as a signaling overlay between LTE physical layer and radio resource control (RRC) protocol [3GP15, 3GP19h] in the client and base station. Our implementation is backward compatible: If the client *or* base station does not support REM, both disable REM overlay and rollback to 5G/4G.

• **Delay-Doppler signaling overlay (§4.3.1):** We realize it on both the client and base station. In 5G/4G, the pending signaling messages are queued in the signaling radio bearer (SRB) at radio link control (RLC) layer [3GP17d, 3GP12b], We first estimate how many slots (thus subgrid size) they need by volume. Then we run OTFS modulation for them, and then forward them to medium access control (MAC) layer [3GP19g, 3GP14]. We further adapt MAC's traffic scheduler to always place all signaling messages in a subgrid in OFDM to meet the OTFS requirement. All data traffic will not be affected since they are handled by the data radio bearers (DRBs) in RLC and scheduled with lower priority in MAC layer.

• **Relaxed reliance on feedback (§4.3.2):** The base station reuses 5G/4G reference signals and modulate them with OTFS. For the client, it first groups cells by their physical base stations based on the global cell identifiers ECI in 4G LTE [3GP11] and NCGI in 5G NR [3GP20].. Then it chooses one cell per base station to measure (intra-frequency cell if any, otherwise inter-frequency cell), estimate its delay-Doppler channel with standard procedure [MAT18], runs Algorithm 1 to estimate other cells from the same base station, and reports them to the serving cell.

• **Simplified, conflict-free policy (§4.3.3):** The base station configures the client to measure all intra/inter-frequency cells' with A3 that meet Theorem 4.3.2 and 4.3.3, and disable other events (thus no multi-stage decision). The non-SNR policies (e.g., preferences and load balancing) remain unchanged.

4.5 Evaluation

We evaluate REM's reliability in extreme mobility (§4.5.1), and its efficiency and overhead of its key components (§4.5.2).

Experimental setup: To approximate real extreme mobility, we run trace-driven emulations over USRP-based testbed.

• *Extreme mobility dataset:* Table 7.2 summarizes our datasets, including **(1) Fine-grained HSR dataset:** We collected it over Chinese high-speed rails in 07/2019–08/2019. We have tested a 1,136 km rail route at 200–300km/h between Beijing and Taiyuan, China. We run a Skype video call in Xiaomi MI 8 phone using China Telecom, and collect the full-stack 4G LTE signaling messages (PHY, MAC, RLC, RRC) using MobileInsight [LPY16]. **(2) Coarse-grained HSR dataset:** We used an open dataset from [WZN19] for larger-scale evaluations. This dataset is collected when the mobile client runs continuous downlink data transfer via TCP-based iperf over Beijing-Shanghai HSR route at 200/300/350 km/h. It includes 357.9 GB data by traveling 51,367 km on the trains. Different from the fine-grained one, this dataset only has RRC messages, thus missing fine-grained OFDM channel information. Together with the LTE signaling messages, it also collects the tcpdump packet traces from the mobile client and server. **(3) Low mobility dataset:** It is our baseline. Since 02/2017, we have collected it with MobileInsight, by driving on highways in Los Angeles with AT&T, T-Mobile, Verizon, and Sprint.

• *Testbed:* Our testbed is based on §4.4. It consists of USRP B210/N210 as client and base stations, which connected to servers with Intel Xeon CPU E5-2420 v2 and 16GB memory. The servers run OAI [ope19] cellular protocol stack. To approximate operational settings, we configure the testbed's radio power, protocol configurations and mobility policies based on above datasets. We run USRP under the unlicensed 2412/2432MHz band instead of lincensed ones. To compare REM with legacy design, we replay our datasets and evaluate if REM can prevent failures in same settings.

Table 4.3: Overview of extreme mobility datasets

		Low mobility	High-speed rails (China)	
		Los Angeles (Fine-grained)	Beijing-Taiyuan (Fine-grained)	Beijing-Shanghai [WZN19] (Coarse-grained)
Movement speed		0–100km/h	200–300km/h	200–350km/h
Route distance		619 km	1,136 km	51,367 km
Mobile operators		AT&T, T-Mobile, Verizon, Sprint	China Telecom	China Mobile, China Telecom
# Signaling messages		46,814	49,781	601,720
Wireless	**Carrier frequency**	731.5–2648.6MHz	874.2–2120MHz	1835–2665MHz
	Bandwidth	5, 10, 20MHz	5, 10, 15, 20MHz	5, 10, 15, 20MHz
	Channel metrics (OFDM)	SNR/BLER/CQI/MCS/RSRP/RSRQ		RSRP/RSRQ
	RSRP range (dBm)	$[-136, -44]$	$[-134, -59]$	$[-140, -60]$
	SNR range (dB)	$[-20, 30]$	$[-20, 30]$	N/A (not collected)
Mobility	**# Cells (base stations)**	932 (503)	1,281 (878)	3,139 (1,735)
	# Feedback	4,023	3,588	81,575
	# Policy configuations	2,771	3,783	38,646
	# Handovers	1,157	2,030	23,779

Ethics: This work does not raise any ethical issues.

4.5.1 Overall Reliability in Extreme Mobility

We evaluate REM's reduction of network failures and policy conflicts in extreme mobility. To compare REM with legacy mobility management, we replay our datasets in Table 7.2, and evaluate how many failures/conflicts in Table 4.1 are reduced by REM. For each handover from our datasets, we extract its feedback and handover command , and infer its corresponding policies with the same approach in §4.1. Based on them, we configure our testbed with same policies, and adapt base stations' runtime transmission power of reference signals with same dynamics of signal strengths (RSRPs) and SNRs in datasets. We repeat this setup with/without REM overlay, and examine if this handover will succeed. To assess REM's benefits for end-to-end applications, we also replay the iperf's TCP data transfer

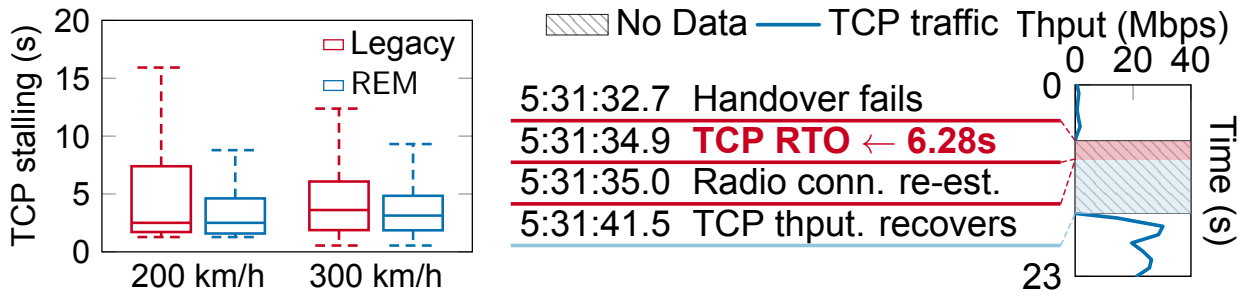
Table 4.4: Reduction of failures and policy conflicts in high-speed rails (LGC=Legacy)

		Low mobility			Beijing-Taiyuan			Beijing-Shanghai								
		0 − 100km/h			200 − 300km/h			100 − 200km/h			200 − 300km/h			300 − 350km/h		
		LGC	**REM**	ϵ	**LGC**	**REM**	ϵ	**LGC**	**REM**	ϵ	**LGC**	**REM**	ϵ	**LGC**	**REM**	ϵ
	Total failure ratio η	4.3%	3.0%	0.43×	8.1%	4.2%	**0.9**×	5.2%	2.4%	**1.2**×	10.6%	2.63%	**3.0**×	12.5%	3.5%	**2.6**×
	Failure w/o coverage hole	3.2%	1.9%	0.68×	4.6%	0.7%	**5.6**×	3.4%	0.7%	**3.9**×	8.6%	0.63%	**12.7**×	10.1%	1.1%	**8.2**×
Failure	Feedback delay/loss	0.78%	0.05%	14.6×	2.4%	0.1%	23×	1.7%	0.1%	16×	4.9%	0.2%	23.5×	6.9%	0.23%	29.0×
	Missed cell	1.8%	-	-	0.8%	0.2%	3×	0.6%	-	-	0.4%	-	-	0.8%	-	-
	Handover cmd. loss	0.61%	0.04%	14.2×	1.4%	0.4%	2.5×	1.1%	0	∞	3.3%	0.03%	109×	2.4%	0.03%	79.0×
	Coverage holes	1.1%	1.1%	0	3.5%	3.5%	0	1.7%	1.7%	0	2.0%	2.0%	0	2.4%	2.4%	0
	Total HO in conflicts	0.95%	0	∞	33.2%	0	∞	19.3%	0	∞	5.5%	0	∞	19.1%	0	∞
Conflict	Intra-frequency conflicts	0	0	0	31.2%	0	∞	18.2%	0	∞	5.5%	0	∞	12.7%	0	∞
	Inter-frequency conflicts	0.95%	0	∞	2.0%	0	∞	1.1%	0	∞	0	0	∞	6.4%	0	∞

in the tcpdump traces if the coarse-grained HSR dataset is used, and quantify their TCP performance with/without REM.

We compare REM and legacy LTE on failure ratios $\eta = \frac{K_{LTE}}{K}$ and reduction $\epsilon = \frac{K_{LTE}-K_{\text{REM}}}{K_{\text{REM}}}$, where K is total handover counts, and K_{LTE} (K_{REM}) is the total handover failure counts in LTE (REM). Since the failures occur randomly with wireless dynamics, we assess REM's *worst-case* failure reduction as a lower bound. For failures from signaling loss/corruption in §4.3.1–§4.3.2, we assume REM can prevent them only if it reduces the error rate to 0. This under-estimates REM's failure reduction since signaling may be delivered with non-zero block error rate. For failures from missing cells in multi-stage policy in §4.3.3, the client will eventually reconnect to a missed candidate cell if its SNR is better than old cell (before which the client has no service). We use this to detect if a cell is available but missed. Since SNR is not collected in Beijing-Shanghai dataset, we do not assess REM's failure reduction for missing cell and thus under-estimates its effectiveness.. Table 4.4 shows REM's reduction of network failures and policy conflicts, and Figure 4.9 shows REM's benefits for TCP and applications.

Overall reliability improvement: Table 4.4 shows REM reduces the overall fail-



(a) TCP stalling time

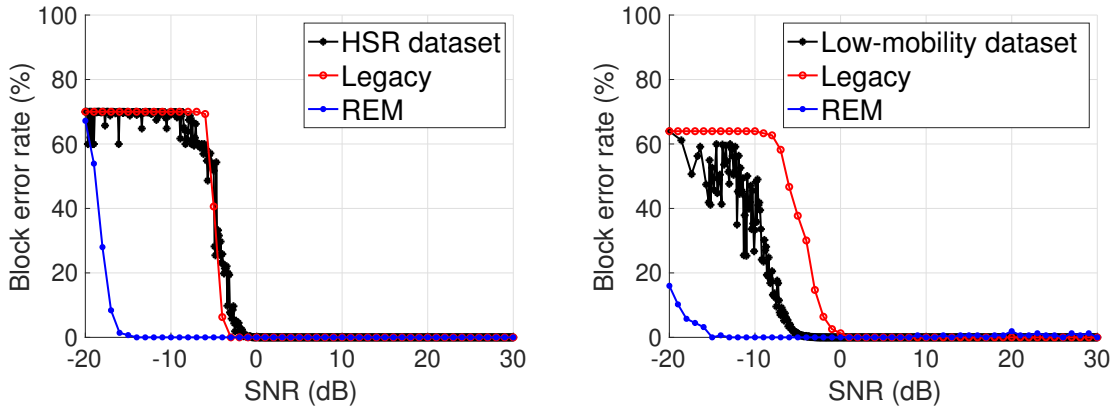
(b) TCP data transfer in failures (from dataset)

Figure 4.9: REM's benefit for TCP. The result at 350km/h is not shown since its LTE signaling messages and TCP traces were not simultaneously collected and evaluated.

ures and conflicts in both HSR datasets at all train speeds. In Beijing-Shanghai route, REM reduces existing LTE's failure ratio by 1.2× (5.2%→2.4%) at 100-200km/h, 3.0× (10.6%→2.6%) at 200–300km/h, and 2.6× (12.5%→3.5%) at 300-350km/h. In Beijing-Taiyuan route at 200–300km/h, REM the failure ratio by 0.9× (8.1% →4.2%). In all cases, REM achieves comparable failure ratios to static and low-speed mobility (e.g., driving in Table 4.1). Note all these failure ratios include the *unavoidable* failures from coverage holes, which can only be avoided with better coverage. Without coverage holes, REM achieves negligible failures (0.6%–1.1%) and failure reductions (3.9×–12.7×) by up to one order of magnitude.

Failure reduction in triggering: With the stabilized signaling (§4.3.1), REM reduces the feedback-induced failures to be negligible (0.1%–0.2%). Note failure reductions in decision and execution can also be indirectly related to faster feedback with cross-band estimation (§4.3.2). We currently classify them to later phases and are working on more accurate breakdown.

Failure/conflict reduction in decision: By eliminating the multi-stage policy, REM mitigates the failures from missed inter-frequency cells (3× reduction in Beijing-Taiyuan dataset). With coarse-grained dataset, we cannot evaluate this benefit in Beijing-Shanghai



(a) High-speed rails (350km/h)

(b) Low mobility (EVA)

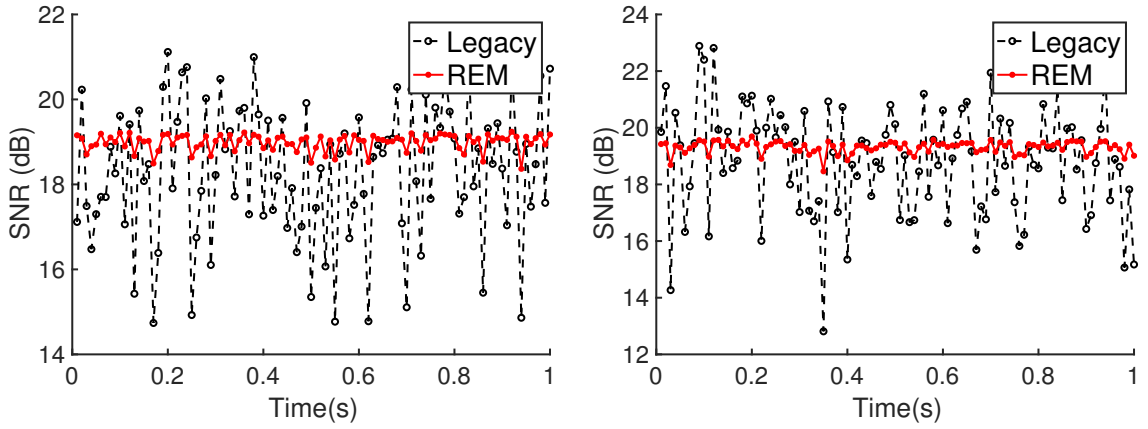
Figure 4.10: REM's error reduction for signaling

route since no SNRs were collected by that dataset. So REM's failure reduction is underestimated in this dataset. Moreover, with the simplified policy in §4.3.3, REM eliminates policy conflicts in all scenarios. While this also eliminates operators' proactive policies that try to prevent failures, such elimination will not negatively affect the failure mitigation with REM's failure reduction (§4.5.2).

Failure reduction in execution: REM reduces its failures to 0–0.4%. Our dataset shows many handover commands in OFDM-based LTE are corrupted/lost with acceptable SNR ([−5dB, 0dB]). Instead, REM explores the full frequency-time diversity in delay-Doppler domain to mitigates the signaling errors/corruptions.

On coverage holes: REM cannot reduce failures from coverage holes. After years of operation, HSRs have been mostly covered with more cells (thus <3.5% failures). Without coverage holes, REM achieves negligible failures (0.7%–1.1% depending on train speed) and more failure reductions (3.9×–12.7×).

Benefits for applications. We last assess how REM benefits TCP and application data transfer. We define the TCP stalling time as the duration that a TCP connection cannot transfer data. With the network failures, the radio connectivity is down and TCP data transfer is blocked. We replay the LTE signaling messages and packet traces in this dataset,



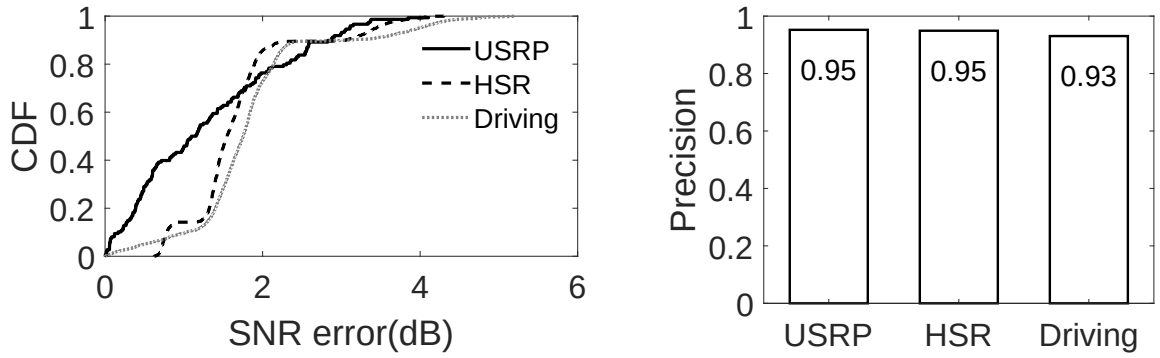
(a) High-speed rails (350km/h) (b) Low mobility (EVA)

Figure 4.11: Stabilized delay-Doppler domain.

and assess the TCP stalling time in legacy LTE and REM. Note in the coarse-grained HSR dataset, the iperf application at the client and server continuously generate data. So the TCP stalling will not be caused by the idle application or connection. Figure 4.9a shows REM's TCP stalling time reduction. With less failures, REM reduces the average TCP stalling from 7.9s to 4.2s at 200km/h, and from 6.6s to 4.5s at 300km/h. Note TCP stalling time is usually longer than the network failures because of its retransmission timeout (RTO). This is exemplified in Figure 4.9b: When network failure occurs, the TCP congestion control aggressively increases RTO for backoff, thus significantly delaying the data transfer. By reducing the failures in extreme mobility, REM mitigates such scenarios and benefits the applications' data transfer.

4.5.2 Efficiency and Overhead

Stabilized signaling in delay-Doppler domain (§4.3.1): We first examine how delay-Doppler domain helps reduces signaling errors/loss. We replay our datasets in Table 7.2 with same signaling message length and SNR, and evaluate their block error rate in a 5G/4G subframe ($M = 12, N = 14$ for 1ms [3GP19e, 3GP17c]) in standard reference multispath models for high-speed train and driving [3GP17a, 3GP19a]. Figure 4.10 confirms REM



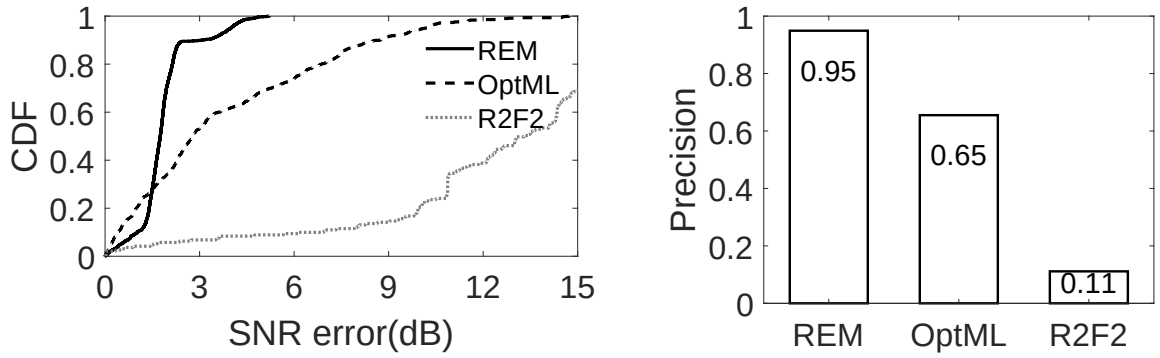
(a) Estimated SNR error (b) Handover decision precision

Figure 4.12: Viability of REM's cross-band estimation.

reduces errors by exploiting time-frequency diversity. This mitigates failures from signaling loss/corruption.

Besides less errors, delay-Doppler domain also facilitates more stable channels and SNRs. Figure 4.11 compares REM and legacy LTE's SNR in the same setting above. In OFDM, slots in different carrier frequency and time experience different channel gains $H(f,t)$ and thus diverse SNRs. Instead, REM adopts OTFS to spread signaling traffic across the entire time-frequency grid, explores the full frequency/time diversity and results in stable channel gains $h_w(\tau,\nu)$ for all slots in the grid (Equation 4.4). This results in more stable SNRs, facilitates SNR-based policy in REM and less transient loops.

Relaxed feedback (§4.3.2): We first explore whether REM retains accurate handover decisions by replacing directly measurements with cross-band estimation. With our dataset, we extract all handovers' measurements and triggering events/thresholds, run REM's cross-band estimation to estimate the target cell if it's co-located with another one, compare the estimated cell quality with the direct measurement, and evaluate whether REM's cell estimation can trigger the same events for handover. Figure 4.12 shows that, REM can achieve $\leq$2dB estimation errors for $\geq$90% measurements, and correctly triggers $\geq$90% handovers. To improve the correct triggering of handovers with cross-band estimation, the operator can further fine-tune its event thresholds (Table 2.1) to tolerate estimation errors.



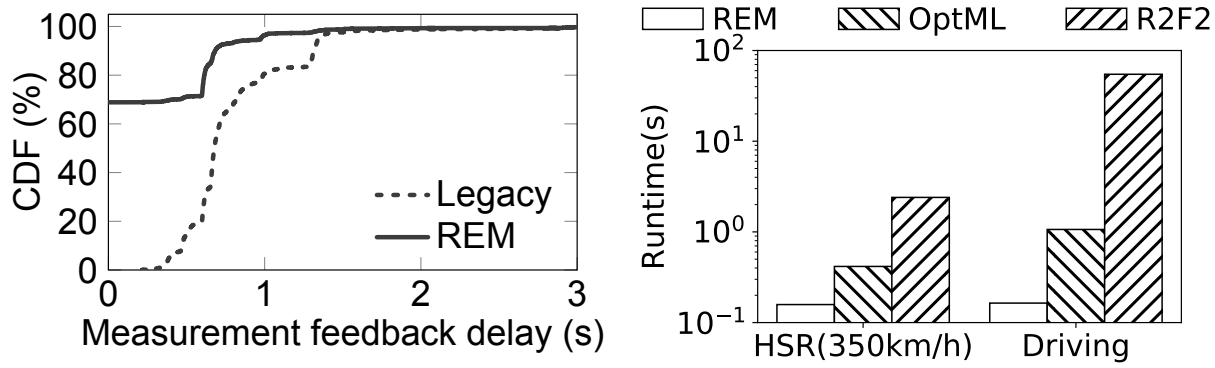
(a) Estimated SNR error

(b) Handover decision precision

Figure 4.13: Cross-band estimation with the HSR dataset.

We further compare REM's accuracy with R2F2 [Vas16] and OptML [Bak19], the state-of-the-art cross-band estimations. Note that R2F2 and OptML require to configure the maximum number of paths to be explored, which will affect their estimation accuracy. For fair comparison, we empirically find their optimal configuration (6 paths for both R2F2 and OptML), and show the results under this setting. Moreover, to train the OptML model, we randomly choose 80% data from the HSR dataset, and use the remaining 20% data to test OptML. Figure 4.13 REM achieves 86.8% lower mean SNR error than R2F2, and 51.9% lower mean SNR error than OptML in the high-speed rail scenario. As explained in §4.3.2, this is because REM explicitly tackles the Doppler effect in extreme mobility.

We last quantify REM's acceleration for the feedback. For each saved measurement in above experiment, REM reduces its measurement durations (including the triggering interval in §4.1.1) and round-trips of sending this feedback (totally T_1). Meanwhile, REM incurs extra delay due to its runtime of cross-band estimation T_2, so the feedback latency savings is $T_1 - T_2$. Figure 4.14a shows REM reduces the average feedback latency from 802.5 ms to 242.4 ms. We also compare REM's runtime T_2 with state-of-the-arts under 5G/4G reference multi-path channels without Doppler (unsupported by R2F2/OptML). Figure 4.14b shows REM outperforms both, without optimization or machine learning. In the HSR, REM saves the runtime from 2.4s (416.3ms) in R2F2 (OptML) to 158.1ms, thus 14× (1.6×) reduction.



(a) Feedback delay reduction.

(b) Cross-band estimation

Figure 4.14: Delays in REM.

While it is possible to accelerate R2F2 and OptML with advanced hardware (e.g., FPGA and GPU), such solution is too expensive for the resource and energy-constrained mobile devices.

Simplified, conflict-free policy (§4.3.3): As shown in Table 4.4, REM's simplified policy provably prevent conflicts. Since operators adopt these conflict-prone policies for proactive failure mitigation (§4.1.2), one may wonder if eliminating the conflicts will cause more failures. We show REM prevents this situation. For all the conflict-prone handover events in our dataset, we follow Theorem 4.3.2 and 4.3.3 to update thresholds, and repeat the evaluation in §4.5.1 to evaluate if more failures will happen in REM. Figure 4.15 compares the failures (without coverage holes) after REM fixes conflicts. It shows that REM still retains negligible failures, since it prevents late handovers with faster feedback and signaling loss/corruption with delay-Doppler OTFS modulation. Both ensure operators do not need to rush the handovers when channel quality is still satisfactory.

4.6 Discussion

Coverage holes and implementation issues: REM currently only mitigates failures with cell radio coverage. Otherwise, no network services exist and no solutions can pre-

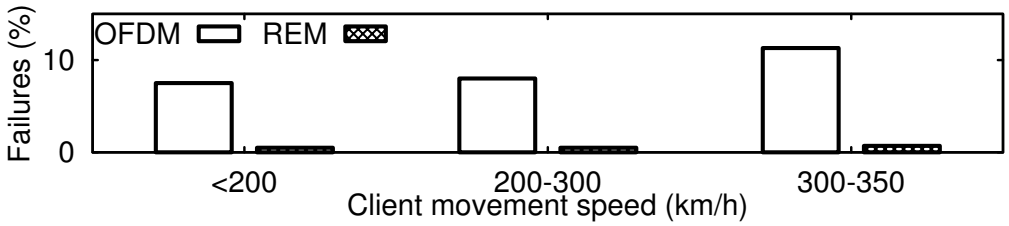


Figure 4.15: Failures without aggressive policies.

vent failures unless the coverage hole is fixed. Besides, the failures from the client/network implementation bugs is also beyond REM's scope.

On data speed: While primarily for reliability, REM also benefits data performance in general for three reasons. First, REM reduces failures and policy conflicts, thus avoiding serve performance downgrade. Second, REM's cross-band estimation saves MeasurementGap for inter-frequency cells, thus offering more spectrum for data transfer. Last but not the least, if data also uses OTFS, REM's SNR-based policy also selects the cell with high capacity $C = B\log(\text{SNR}+1)$. Theorem 4.3.2 and 4.3.3 still hold by replacing SNR with capacity.

Implications on IoT and edge: REM helps them simplify their application-layer operations. With REM, the IoT/edge will have a more stable network condition. This facilitates predictive solutions for IoT/edge to improve the quality-of-experiences (e.g., in virtual reality [TLL18]) and saves signaling overhead (e.g., in massive IoT).

CHAPTER 5

Resolving Policy Conflicts in Multi-Carrier Cellular Access

5.1 The Case for Policy-Based Inter-Carrier Switch

We next make a case for policy-based selection as the fundamental component for the inter-carrier switch. In policy-based switch, each carrier is assigned certain policy attributes, in the form of a preference value or certain threshold-based forms for specific measures. These attributes reflect the multi-carrier service provider (MCSP, such as Google)'s policy demands (e.g., faster network, better coverage, and roaming agreements with carriers). At a given location, the MCSP uses these carrier attributes to select the most preferred carrier.

Figure 5.1 illustrates an example. There are two carriers (C1 and C2), and two cells (4G/3G) for each carrier at the given location. The MCSP uses a preference-based policy by specifying the preference value for each RAT in each carrier, thus resulting in the preferred switch order $(4G, C1) > (4G, C2) > (3G, C1) > (3G, C2)$. Given this policy, the MCSP first switches the device to carrier $C1$, since 4G in $C1$ is the most preferred choice. Within $C1$, cell 1 with higher priority ($p = 4$) is selected based on the intra-carrier handoff policy. Note that cell 1 is a 4G cell in $C1$; this is consistent with the inter-carrier policy. In the example, the MCSP checks carrier-level preference only and leaves cell selection decision to carriers. This preserves the operation autonomy of each carrier.

Policy-based inter-carrier switch is needed by the MCSP for three reasons. First, the policy naturally arise at the MCSP level. MCSP builds its service on top of individual cellular

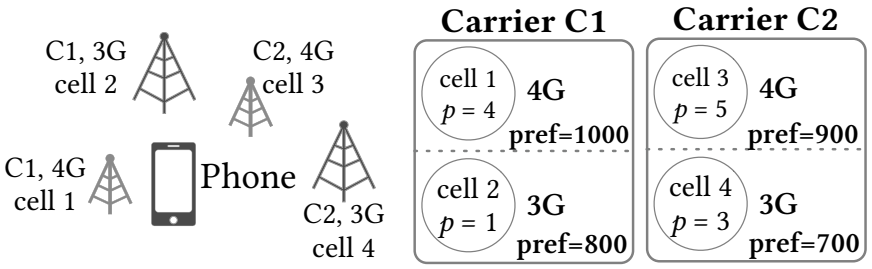


Figure 5.1: Policy-based inter-carrier switch example

carriers, and has to balance among carriers for both technical concerns (e.g., select the best-performing carrier) and nontechnical interests (e.g., which one is a more favorable partner). Second, the policy issues further exhibits in operational practices, such as dealing with geographical diversities of carriers, or even traffic engineering when distributing cellular data traffic across carriers. Third, the policy allows the MCSP to make *configurable* decisions to accommodate diverse demands (e.g., faster network, better coverage, and partner preference).

Policy-based switch further offers several nice properties. First, it decouples choosing a carrier network from the cell selection within the carrier. Consequently, MCSP only needs coarse-grained information on carriers rather than the fine-grained cell information within each carrier network[1]. Second, it leverages the largely standardized intra-carrier mechanism, and keeps the policy design simple at the carrier level. It thus does not require the standardization process again. Last but not the least, it preserves the autonomy and privacy of each individual carrier network. An MCSP works with the carriers without mandating the disclosure of the operational practices of these cellular carriers.

Examples of policy-based switch. We identified three common forms of policy:

• *Preference-based switch.* At each location, a carrier is assigned a local preference value. A carrier with the highest preference is selected assuming the same other conditions.

[1]It may be infeasible for an MCSP to access the fine-grained, cell-level information at runtime across all carriers. This has been the practice by Google Fi. The current hardware will not allow for the device to obtain all cell-level measurements and metrics unless registered to the carrier. The device has to constantly scan and switch to all available carriers to collect such detailed information; this incurs service outage.

• *Threshold-based switch.* For a carrier in a given location, it is assigned a threshold criterion for certain measures (e.g., latency, throughput or a mix) as its attribute. When the threshold conditions are satisfied, a new carrier would be selected. The goal is to find a carrier, which is better than the serving carrier and meets the threshold requirements.

• *Hybrid switch.* The carrier attributes are specified in the form of local preference *and* threshold forms.

The above forms of policy attributes are simple enough to realize, but still generic enough to cover many practical usage cases. Similar forms have also been used in other operational networking systems. The most notable example is that the Internet BGP routing has used the preference attribute in its inter-domain route selection [GW99, GR01]. The preference-based policies are also used in intra-carrier handoff management [3GP19c, 3GP12a, LXP16, LDL16] and WiFi AP selection [Sta16] (in latest Android/Linux). The threshold-based forms are also the common practice for intra-carrier handoff management [3GP19c, 3GP12a, LXP16, LDL16] and WiFi AP selection [CNR15, BMV10]. As we will show later, the major difference between our form and these efforts is the conflicts with the intra-carrier policy.

Operational system in reality. We have observed that Google as an MCSP has largely adopted policy-based switch when making the inter-carrier selection in its Google Fi. In fact, we have seen from the logs (via `logcat` on Pixel/Nexus phone models) that both preferences and thresholds forms are used when selecting a preferred carrier network by Google. Moreover, its recent machine learning-based switch module can also be viewed as a variant of threshold-based policy.

Specifically, Google Fi uses a monitor-controller architecture. Each monitor tracks some metrics in parallel and makes a switch decision proposal. The controller receives these proposals and decides the final carrier. Notable monitors include a *PoorNetwork* monitor (labeled as `PNP` in `logcat`), which assigns *preferences* on carriers and RATs to facilitate the carrier selection. A *GeoLocation* monitor (labeled as `Flock`) uses the crowdsourced carrier quality data to perform *pairwise comparisons* on target carriers. The newest version also

includes a machine learning-based monitor (labeled as `K2so`) that predicts carriers' quality and uses *thresholds* for decisions. Therefore, both preference and threshold-based policies are adopted in Google Fi's design. Given certain conditions, it may use only one, or both. For example, when location service is disabled, only *PoorNetwork* monitor remains active so the policy is preference only. If *GeoLocation* monitor is active, *PoorNetwork* monitor's decision is usually overshadowed, so effectively only the threshold policy is used.

5.2 Improper Inter-Carrier Policy

Policy-based inter-carrier switch is necessary for an MCSP and possesses appealing features. However, improper policy practice may also yield unexpected behaviors such as loops. In this section, we show an example to illustrate the incurred issues as well as their impacts.

5.2.1 An Illustrative Example

We now show an example to illustrate the policy conflicts and potential negative effects. Consider the scenario in Figure 5.2. It is an office building environment with two carriers C_1 and C_2, with two deployed cells belonging to each. The phone remains static with constant wireless channel conditions. It uses multi-carrier cellular access to the two carriers. The inter-carrier policy takes the preference-based form. Given the preference values for each RAT in $C1$ and $C2$, the preferred order is given by: $(4G, C1) > (4G, C2) > (3G, C1) > (3G, C2)$. This is a sensible policy by MCSP. It is well grounded at the inter-carrier policy level: $4G$ is favored over $3G$, while carrier $C1$ is favored over $C2$ since $C1$ has generally better performance (e.g., higher access speed). On the other hand, the cell-level policy at the intra-carrier level uses the priority-based policy. In carrier $C1$, the 4G cell 1 is set with $p = 1$, whereas its 3G cell 2 has priority 2. This is because cell 2 is a deployed urban/enterprise small-cell in the office building that seeks to offload the traffic for local users from the macro-cells. Note that small cells are indeed quite common. Recent data [For17] predicts that, the

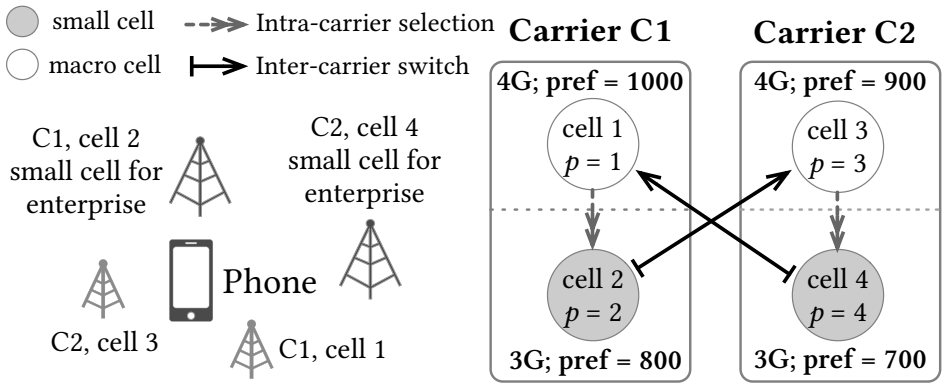


Figure 5.2: Example of policy conflicts and bad impact.

deployed small cells will reach 11.4M by 2025, and the annual growth rate is 14%. The rate increases to 36% in nonresidential areas. Similarly, in carrier $C2$, its 3G cell 4 (also an urban small cell) is assigned a higher priority value 4 than its 4G cell 3. Within each carrier, the intra-carrier cell policy is also well justified.

Policy conflicts then arise between inter-carrier and intra-carrier levels. The inter-carrier level prefers 4G RAT over 3G RAT for better technology and higher access speed, whereas the intra-carrier level favors 3G over 4G for better traffic offloading. Both are well justified from their own interests based on their knowledge. The MCSP uses the carrier-level information only and sets its preference based on what RAT is superior and which carrier offers better performance. Within each carrier, the carrier sets its policy to consider the unique small-cell deployment in the example setting.

The above policy conflicts also result in unexpected behaviors. If the MCSP intends to strictly enforce its inter-carrier policy $(4G, C1) > (4G, C2) > (3G, C1) > (3G, C2)$, it will suffer from two categories of bad effects. First, if MCSP does not resolve the conflicts, it is stuck into persistent loops. The device first switches to carrier $C1$ based on the inter-carrier preference $(4G, C1)$. However, cell 2 is selected since this cell has higher intra-carrier priority, once the device is in carrier $C1$. Unfortunately, cell 2 is a 3G cell, but not a 4G cell. Since this is not what the inter-carrier policy dictates, the device goes back to the carrier level. It then selects carrier $C2$ after the selection failure in $C1$. Once in carrier $C2$, the

3G small cell 4 is also chosen for higher priority among the two cells. This is also not what the inter-carrier policy wants. It then repeats the above steps and gets into the persistent loop $(4G, C1) \mapsto (3G, Cell\ 2) \mapsto (4G, C2) \mapsto (3G, Cell\ 4) \mapsto (4G, C1) \mapsto \ldots$. Note that, despite the existence of 4G RATs in both carriers, neither is selected. The inter-carrier policy mandates the continuous search to hopefully settle down at one 4G cell. Second, the MCSP may decide to stop the switch after several rounds (e.g., via recording historical switches, or maximum attempt counters). The device may switch to carrier $C1$ and select the 4G cell 1 without getting into the loop. However, this requires the device to disable its 3G access at the phone. This seems to honor the inter-carrier policy. However, it is against the intra-carrier policy for small cells within carrier $C1$. It is also not a good choice for the device and the user, since it unnecessarily disables the 3G option and compromises selection flexibility.

For the above example scenario, the best option is to switch to carrier $C1$ (preferred over C2 based on inter-carrier policy) but select the 3G cell 2 (that is favored based on the intra-carrier policy for small cells). This sheds lights on the simple rule that helps to resolve the policy conflicts: *Upon policy conflicts, intra-carrier policy should be prioritized over the inter-carrier policy in the resolution process.* This intuitive rule is also consistent with the two-tier switch scheme. At the carrier level, the MCSP uses policies to specify the general preference, but may not have the accurate information (e.g., small-cell deployment), which is only accessible within the carrier. Therefore, whenever conflict arises, intra-carrier policy, which is well defined and practiced by individual carriers, should be prioritized first.

Real-world instance. The above example is conceptual; however, we do observe a real trace in Google Fi that can be mapped to this example. Trace 1 demonstrates a loop when the user is static[2]. Google sets a preference list for four RATs: preference for both T-mobile and Sprint LTE is $P_{T,LTE} = P_{S,LTE} = 1000$; the preference for T-mobile HSPA (a 3G RAT) is $P_{T,3G} = 700$ and that for Sprint EHRPD (a 3.5G RAT) is $P_{S,3G} = 800$. In the experiment,

[2]It was observed in the latest version of Google Fi v5.1.11.

```
14:32:07 Connected to Sprint EHRPD.
         Already waited for 02:58, will have to wait for 27:02 more.
14:59:17 Evaluation.  Switch request Sprint -> T-Mobile is approved.
15:00:09 Switch done.  Current network: T-Mobile HSPA.
15:00:09 Reset monitor.  Elapsed time: 43:29, locked until 6:43:30.
21:17:20 Unlock switch.  Current network: T-Mobile HSPA.
         Already waited for 00:00, will have to wait for 2:00:00 more.
23:18:08 Evaluation.  Switch request T-Mobile -> Sprint is approved.
23:18:25 Switch done.  Current network: Sprint-EHRPD.
23:18:25 Reset monitor.  Elapsed Time: 9:05:30, locked until 15:01:47.
```

Trace 1: A loop in Google Fi

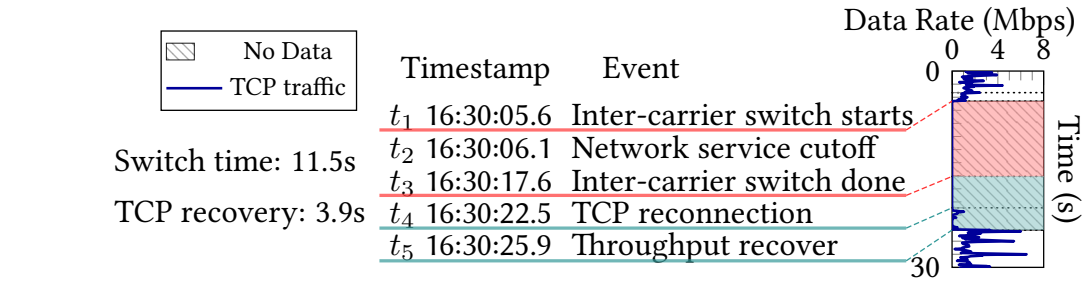
LTE signals in both carriers are weak, and the phone camped on HSPA in T-mobile and EHRPD in Sprint. As shown in the trace, the loop Sprint $\mapsto$ T-Mobile $\mapsto$ Sprint occurs, because inter-carrier policy attempts switch to the carrier with highest-preference RAT but could not stay. Note that, such a loop is not happening very frequently, due to an engineering fix (lock timer at Lines $2, 5, 7, 10$) implemented by Google Fi to limit the interval between switches. However, such fix can only prolong the period of a loop, but not eliminating the loop. We next quantify the loop's practical impacts.

5.2.2 Real Impact of Inter-Carrier Loop

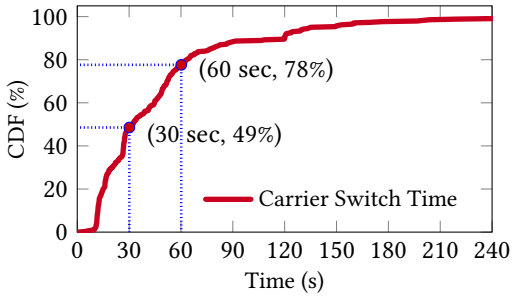
Inter-carrier loop disrupts user's cellular service, incurs significant battery drains at the device, and triggers signaling overhead on carriers. Its impact accumulates as the loop continues.

The phone loses its cellular data and voice service during the switch[3]. Figure 5.3b shows

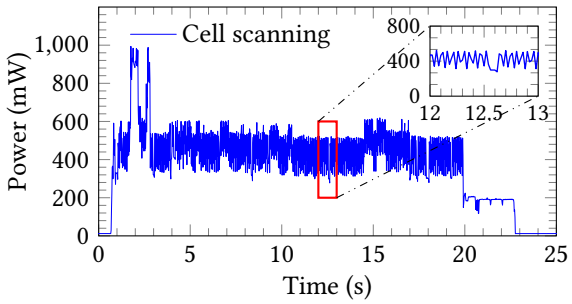
[3]The device may still be able to access network via WiFi, but this could be still an issue when WiFi is unavailable (e.g., outdoor environment).



(a) TCP throughput drop



(b) Service interruption time



(c) Modem power consumption

Figure 5.3: Impact of inter-carrier switch.

the time taken by a single switch in Google Fi, from our small-scale user study[4]. About 51% of the records took 30 seconds or more, while 22% of the switches took more than one minute. TCP throughput tanked during switch as shown in Figure 5.3a.

The battery consumption hikes (could be 3× higher than idle mode [HQG12, DWC13]). This is rooted in intra-carrier design; phone exhaustively scans cells and keeps the radio on during switch. See Figure 5.3c for the power consumption[5]. The average power draw is around 400 mW during cell scanning (a major phase of carrier switch), significantly higher than the idle state. Furthermore, the phone exchanges signaling messages with *every* carrier's *every* RAT it can reach, incur excessive signaling overhead [LDL16].

[4]It shows 350 records, spanning from 2017/02 to 2018/03 on four phone models that supports Google Fi: Nexus 6/6P and Pixel/Pixel 2. All data are collected anonymously and comply the IRB regulations.

[5]The measurement is done on Samsung S5 with minimal background service running in comparison with airplane mode; the energy contrast is similar on all phone models.

Table 5.1: Notations

C_i	Carrier i
RAT_j	Radio access technology j (e.g., 3G, 4G)
c^k/c_i^k	Cell k (in carrier C_i)
$P_{i,j}/P_i$	Inter-carrier preference on carrier C_i's RAT_j / C_i
$p(c^i)$	Intra-carrier priority of cell c^i
$M, M(C_i)$	Measure M (on C_i) for inter-carrier policy
$Q, q(c^j)$	Measure Q (on c^j) for intra-carrier policy
δ, θ, ϕ	Different inter-carrier thresholds (on carrier)
$\Delta^i, Thresh^{i,j}$	Different intra-carrier thresholds (on c^i/c^j)

5.3 Methodology and Overview

We next study the policy conflicts in the multi-carrier access. We present our methodology, and overview our main results.

5.3.1 Methodology

We take a three-step approach to study the policy conflicts and loops. First, we model the inter-carrier policy and derive the theoretical stability conditions. Then we show validations from Google Fi for these results. We last propose practical guidelines for provable stability, and assess them via emulations.

A model for inter-carrier switch. Consider N carriers $C_1, C_2, \ldots, C_N$ that provide services at the user's current location. Each carrier has K radio technologies, denoted as $RAT_1, RAT_2, \ldots, RAT_K$. There are $n(i)$ cells in carrier C_i: $c_i^1, c_i^2, ..., c_i^{n(i)}, i \in [1, N]$. Within each carrier, intra-carrier policy selects the serving cell for the device. An *inter-carrier switch* is the transition from one carrier C_i to another carrier C_j specified by the inter-carrier policy. Therefore, We model such a switch as a discrete transition $C_i \mapsto C_j$. Here, we only consider deterministic policy. That is, under static network conditions and staying on the serving

cell, a single phone would always makes the same carrier-switch decision.

Intra-carrier policy. There are two types of intra-carrier policy in LTE: *Idle-state policy* that is used when there is no active radio connectivity (standardized in [3GP15, 3GP19c]); and *active-state policy* that is used otherwise. In multi-carrier access, only the idle-state policy should be considered, because inter-carrier switch occurs in idle-state only (by deregistering from the old carrier and registering to the new carrier)[6]. The idle-state policy is based on the per-cell priority and threshold of measures. It moves the device from cell c^u to c^v *iff.* (1) *Absolute value:* $q(c^v) > Thresh1^{u,v}$ if $p(c^v) > p(c^u)$; (2) *Direct comparison:* $q(c^v) > q(c^u) + \Delta^u$ if $p(c^v) = p(c^u)$; (3) *Indirect comparison:* $q(c^u) < Thresh2^u, q(c^v) > Thresh3^{u,v}$ if $p(c^v) < p(c^u)$.

Assumptions. In static case, we assume the user does not move, and all cells' measures (e.g., radio signals, latency, throughput, ...) remain stable. Our results can still be generalized when this assumption does not hold (discussed in §5.9). We further assume that intra-carrier policy does not change, and it does not incur loops within carrier (e.g., via regulations in previous work [LDL16]). The device initially is connected to a carrier C_0's RAT_0[7]. It performs inter-carrier switch only when the intra-carrier reselection stabilizes, and use specific inter-carrier policies to be elaborated next.

Loops and stability. The inter-carrier policy can incur consecutive switches even under the assumed static condition (§5.2.1). Formally, an N-carrier loop is an inter-carrier switch sequence, starting from one initial carrier, traverse each carrier *exactly once*, and then switch back to the same initial carrier. For example, the switch sequence $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$ is one instance of N-carrier loop. The *order* of the sequence matters, for example, sequence $C_1 \mapsto C_3 \mapsto C_2 \mapsto C_1$ is a different loop to $C_1 \mapsto C_2 \mapsto C_3 \mapsto C_1$. An N-carrier loop is *persistent* when single instances of N-carrier loop happen repetitively under the *same*

[6]As a real example, Google Fi will suspend the inter-carrier switch until the device completes the data transfer or calls and moves back to idle state.

[7]C_0 could be any of the $C_1, C_2, \ldots, C_N$ and RAT_0 could be any of the $RAT_1, RAT_2, \ldots, RAT_K$

static condition. 8We have the following result (proof in §B.1):

Proposition 5.3.1. *An N-carrier loop is persistent loop under the static condition and deterministic switch policy.*

An inter-carrier policy is *stable* iff. it will not incur persistent loops. In the following sections, we will derive the theories and guidelines for the inter-carrier policy stability.

Real-world validation. We use Google Fi to validate our results due to its wide deployment in reality. We collect Android system logs that records Google Fi's decision and activity, and reconstructs its main operational logic and policies. We also validate our modeling via limited reverse engineering effort and online user forum reports. As shown in §5.1, Google Fi uses preference-only, threshold-only and hybrid policies in different scenarios. Therefore, we set the specific condition to make Google Fi's policy consistent with each subcategory. Then, we observe/construct the loop scenario and validate our analytical reasonings.

5.3.2 Roadmap and Overview

This work explores the theoretical conditions and practical guidelines for the policy conflicts (loops) in multi-carrier access. Figure 5.4 and Table 5.2 classify the conflicts based on their causes. Such conflicts can arise from the preference-based, threshold-based, and hybrid inter-carrier policies. We overview each category, examine how it conflicts with the intra-carrier policy, and summarize our results.

Preference-based policy (§5.4). In this category, the MCSP's inter-carrier preference settings contradict with carriers' priorities for the same carrier or RAT (exemplified in §5.2.1). Based on the granularity of the preferences that MCSP uses, there are two sub-categories:

• *RAT-aware preference* (§5.4.1): The inter-carrier policy assigns a preference to each (carrier, RAT) pair (exemplified in Figure 5.4a and Figure 5.2). We show that, the stability can be violated when the MCSP's inter-carrier preferences contradict with the carriers' internal priorities.

Table 5.2: Classification of main results (NC: necessary condition; SC: sufficient condition).

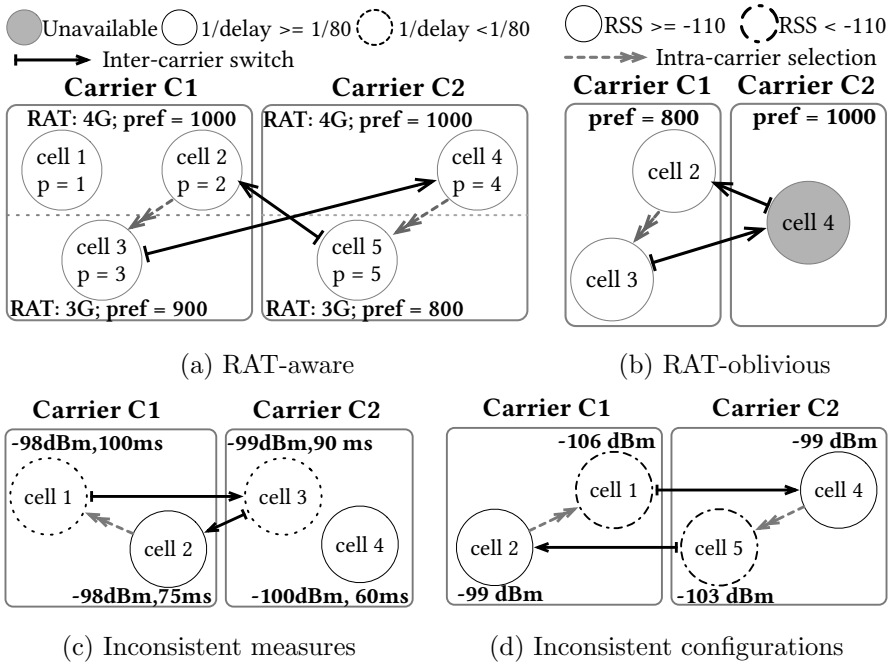
Inter-Carrier Policy		Theorem Results				Guide-	Vali-
Form	Subcategory	Reference	Insight	SC?	NC?	line	dation
Preference	RAT-aware	Thm. 5.4.1	Inconsistent preference on RAT	✓	✓	§5.7.1	§5.4.1
	RAT-oblivious	Thm. 5.4.2	Preference conflict w/ unavailability	✓	✓	§5.7.1	§5.4.2
Threshold	Incons. measures	Thm. 5.5.1 & 5.5.2	Loop-prone criteria; Min-measure rule	✓	✓	§5.7.2	§5.5.1
	Incons. config	Thm. 5.5.3 & 5.5.4	Some thresholds are unstable	No	✓	§5.7.2	N/A
Hybrid	Preference first	Thm. 5.6.1	Some threshold criteria are ruled out	No	✓	§5.7.3	§5.6
	Threshold first	Thm. 5.5.1–5.5.4	Same as threshold theorem	No	✓	§5.7.3	§5.6

• *RAT-oblivious preference* (§5.4.2): The inter-carrier policy assigns a preference to each carrier only. The stability is violated if the inter-carrier preferences conflict with intra-carrier policies to unavailable cells (exemplified in Figure 5.4b).

Threshold-based policy (§5.5). When the MCSP uses the threshold-based policy, it may conflict with the intra-carrier policies and incurs loops in two scenarios:

• *Inconsistencies of measures* (§5.5.1): The inter-carrier and intra-carrier policies evaluate the same carrier using different types of measures. This could happen since the MCSP and carriers may target different goals (e.g., latency v.s. radio quality, as exemplified in Figure 5.4c). We show that, some threshold-based evaluation criteria are loop-prone. Moreover, if measures are independent, the necessary *and* sufficient condition for the stability is that the MCSP applies the *minimum measure rule*. If they are correlated, our theorems are still sufficient, but not necessary.

• *Inconsistencies of configurations* (§5.5.2): Even if the inter-carrier and intra-carrier policies evaluate the same measures, they can conflict with each other due to uncoordinated threshold values. Figure 5.4d illustrates an example: Under constant and static measures, the inter-carrier switch and intra-carrier handoffs are triggered simultaneously, thus incurring loops. To ensure stability, we derive a set of necessary conditions for different criteria for threshold coordination. The key result is that, such coordination can be performed using *aggregated*



(a) RAT-aware (b) RAT-oblivious

(c) Inconsistent measures (d) Inconsistent configurations

Figure 5.4: Classification of policy conflicts and loops.

threshold values rather than fine-grained thresholds. This simplifies the coordination and prevents carriers from exposing its internal policies to MCSP.

Hybrid policy (§5.6). When the MCSP uses both preferences and thresholds, we show how above results can be generalized. Broadly speaking, there are two approaches to combine the preferences and thresholds: (1) *Preference-first policy:* It evaluates each carrier's preference first, then applies different threshold-based criteria based on the preference relations. We prove that the use of preferences poses more constraints on choosing the threshold-based criteria; (2) *Threshold-first policy:* It applies the same threshold-based criteria to all carriers, and then select the one with the highest preferences. We show that the results in threshold-based policy still hold here.

5.4 Stability for Preference Policy

We first study the preference-based inter-carrier policies.

5.4.1 RAT-Aware Preferences

5.4.1.1 Policy Form

The MCSP assigns a preference $P_{i,j}$ to carrier C_i's RAT_j (exemplified in Figure 5.2). The objective is to *select a most preferred carrier* according to such preference list[8]. Let P^i_{max} be the maximum RAT preference in C_i. Under this goal, a simple RAT-aware policy is as follows.

Policy 1 (RAT-aware inter-carrier switch)**.** *Let C_i be the serving carrier. Perform inter-carrier switch $C_i \mapsto C_j$ and mark C_j as selected, if (a) $P_{j,k} = P^j_{max} > P_{i,m}, j \neq i$; and (b) C_j has not been selected. When all highest preferred carriers have been selected, clear the marks to allow flexibility.*

Policy 1 can cover and emulate a broad range of RAT-aware inter-carrier policies. For instance, one may choose to select a carrier with *higher* preference than the current carrier's (instead of the highest preference). In doing so, a minimum acceptable preference is also needed. It is equivalent to setting all carriers that satisfy such "minimum preference" with *equal* highest preference value and Policy 1 still applies.

5.4.1.2 Stability Condition

The stability is violated by the conflicts between inter-carrier preferences and the intra-carrier priorities. To unveil the concrete conflict form, we first prove the following result (proof in Appendix B.1):

Lemma 5.4.1. *Assume preference satisfies $P^1_{max} \geq P^2_{max} \geq \cdots \geq P^N_{max}$, where $P^i_{max} = \max_j P_{i,j}$. An N-carrier loop occurs* iff. *the inter-carrier switch sequence (*) $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$ occurs.*

[8]We allow same preference value for different (carrier, RAT) pairs. The tie is broken using a given order, i.e., smaller index on carrier first and then RAT.

Lemma 5.4.1 shows that the ordering of switch sequence in an N-carrier loop follows the preference order. Proof is in Appendix B.1. The sufficiency is trivial to prove. For necessity, N-loop indicates the phone cannot connect to any of the most preferred RATs. Based on the RAT-aware policy, the phone would explore all carriers from the one with the highest preference to that with the lowest preference. This inspires the N-carrier loop condition below:

Theorem 5.4.1 (Inter/intra-carrier preference conflict)**.** *Assume the inter-carrier switch takes Policy 1. A persistent N-carrier loop happens* iff. *(a) Every carrier has one or more RATs (denote the set as $\mathbf{RAT}_H$) assigned with equal, highest preference; and (b) each carrier's intra-carrier priority and threshold result in reselection from a carrier $RAT_k \in \mathbf{RAT}_H$ to a different carrier $RAT_m \notin \mathbf{RAT}_H$.*

Theorem 5.4.1 explains how inter-carrier RAT preferences contradict with the intra-carrier priorities: The inter-carrier policy will seek C_i's $\mathbf{RAT}_H$, but since C_i's intra-carrier policy move to RATs out of $\mathbf{RAT}_H$, inter-carrier will switch to another carrier C_j.

5.4.1.3 Validation

We validated the loop between two carriers, T-mobile and Sprint, in Google Fi. The example is the same as Trace 1 in §5.2. The loop Sprint $\mapsto$ T-Mobile $\mapsto$ Sprint is triggered because inter-carrier policy prefers LTE equally but neither carrier can stay in LTE. The loop exists in the trace but is not persistent, because Google Fi implements an engineering fix (lock timer) to limit switch frequency. The side effect, however, is that the device may get stuck in a carrier even it leaves its radio coverage.

5.4.2 RAT-Oblivious Preference List

We next analyze the RAT-oblivious preference, and discuss its relation with the RAT-aware preferences.

5.4.2.1 Policy Form

With RAT-oblivious preference, the MCSP assigns a preference value P_i to carrier C_i, and still *selects a most preferred carrier*. Under this goal, The basic RAT-oblivious preference policy is specified as follows.

Policy 2 (RAT-oblivious inter-carrier switch)**.** *Perform inter-carrier switch to the highest preference carrier which has not been selected if (a) the serving carrier's preference is not the highest preference; or (b) the serving carrier is served by an unavailable cell. When the serving carrier is unavailable but all other carriers have been selected, clear the marks.*

Note that Policy 2 is similar to RAT-aware Policy 1. However, it is not a subset of that, because without further differentiate the cells within carrier, user may not get available service. It explicitly addresses this unavailability condition by performing an inter-carrier switch.

5.4.2.2 Stability Condition

Intuitively, we can draw a similar conclusion to Theorem 5.4.1. If every carrier may move the device to an unavailable cell, then the inter-carrier policy will keep trying and may form a loop. Theorem 5.4.2 confirms this intuition, and is proved in Appendix B.1.

Theorem 5.4.2 (Cell unavailable loop)**.** *An N-carrier loop occurs* iff. *the intra-carrier logic in all carriers moves the device to an unavailable cell.*

Remark. Theorem 5.4.2 gives the sufficient and necessary condition for N-carrier loop assuming Policy 2.

5.4.2.3 Validation

In Google Fi, Google distributes a RAT-oblivious preference list to phone. The logic of performing switch is highly similar to Policy 2. Unfortunately, we have not observed loop in

real-world. It is partially due to the strong condition Theorem 5.4.2 states (*all carriers are unavailable*). There is also an engineering fix by Google Fi: once an inter-carrier switch has been performed using such list in one location, it will lock switch for six hours. Essentially, such engineering fix limits the loop frequency, at the cost of getting stuck in one carrier for extended period.

5.5 Stability for Threshold Policy

We next analyze and validate the threshold-based inter-carrier policies by following the classifications in §5.3.2.

5.5.1 Inconsistency of Measures

5.5.1.1 Policy Form

We consider the following policies.

Inter-carrier policy. In finding a better carrier, the most basic and straightforward way is to *find a carrier whose measure is better than the serving carrier*. If we denote serving carrier's measure as $M(C_s)$, target carrier's measure as $M(C_t)$, and thresholds as δ, θ, and ϕ (all > 0), one can enumerate four possible comparison criteria that reflect this goal:

F1. $M(C_t) > \theta$ (candidate's measure is higher than threshold)

F2. $M(C_s) < \theta \wedge M(C_t) \geq \phi$ (serving carrier's measure is lower than a threshold, and candidate's measure is higher than another threshold)

F3. $M(C_t) > M(C_s) + \delta$ ($\delta \geq 0$; candidate's measure is offset higher than the serving carrier's)

F4. $M(C_s) < \theta \wedge M(C_t) > M(C_s) + \delta$ ($\delta \geq 0$; serving carrier's measure is lower than a threshold, and candidate's measure is offset higher than the serving carrier's)

Given these criteria, the inter-carrier policy performs the switch $C_i \mapsto C_j$ when C_i and C_j's measures satisfy criterion ($F*$) from *F1 – F4*.

Measures of carriers. Assume the inter-carrier policy uses the measure type M, while the intra-carrier policy uses the measure type Q ($Q \neq M$). Denote $M(C_j)$ as the measure M of carrier C_j, $M(c_j^u)$ as the measure of cell c_j^u in C_j, and $M^{min}(C_j) = \min M(c_j^u)$. The MCSP will compute the per-carrier measure C_j based on the per-cell measures $\{M(c_j^u)\}$ [9].

5.5.1.2 Stability Condition

We first show that, some criteria are inherently loop-prone and thus should not be used in *any* inter-carrier policies (proof in Appendix B.2).

Theorem 5.5.1 (Unstable comparison)**.** *If inter-carrier switch policy takes Criterion* F1*, then the inter-carrier policy cannot be loop-free no matter how the thresholds are configured.*

Intuitively, *F1* violates stability since it does not evaluate the serving carrier's measure. If both the serving and candidate carriers meet *F1*, the device will oscillate between them. For stability, the threshold evaluation criterion must be based on the measure of both carriers.

Besides, we restrict $\phi \geq \theta$ for *F2* to avoid trivial loops. All theorems regarding *F2* in this section assume $\phi \geq \theta$.

For *F2*, *F3*, and *F4*, stability is ensured iff. the following *minimum-measure* rule is applied (proofs in Appendix B.2):

Theorem 5.5.2 (Minimum-measure rule)**.** *Assume inter-carrier policy's measure M and intra-carrier measure policy's Q are independent. The stability is satisfied* if and only if $M(C_j) - M^{min}(C_j) \leq g(F*)$[10] *is guaranteed no matter how per-cell measures change, where*

[9] More practically, the MCSP could directly get the carrier's measure in the form of an aggregation of all internal cells' measure.

[10] $m^{min}(C_j) = \min m(c_j^u)$ is the minimum measure of all cells in carrier C_j.

$g(F*)$ is defined as the following:

$$g(F*) = \begin{cases} \phi - \theta & \textit{for } \text{F2}, \\ \delta & \textit{for } \text{F3 } \textit{or } \text{F4}. \end{cases}$$

The proof is shown in Appendix B.2. Intuitively, if the worst case after switching could even satisfy the expectation, no matter what intra-carrier policy will not conflict.

As a special case, the following sufficient condition offers a simpler rule regardless of the criteria form (F2 – F4):

Lemma 5.5.1 (Simple minimum-measure rule)**.** *Following the assumption in Theorem 5.5.2, the threshold policy is stable if the carrier's measure $M(C_j) = M^{min}(C_j)$.*

Fundamentally, both rules are caused by the *different control granularities* between the inter- and intra-carrier policies. Inter-carrier policy works at *RAT/carrier level only*. It cannot control the *cell-level* selection, which is done by the intra-carrier policy. With independent measures, the minimum rule is critical for the consistent decision between RAT/carrier-level switch (inter-carrier policy) and cell-level selection (intra-carrier policy).

Both results can also be generalized to the *different, yet correlated* measures (e.g., latency and signal strength): Lemma 5.5.1 still holds. Theorem 5.5.2 is sufficient, but not necessary.

In reality, there usually exists some cells which are never selected by intra-carrier policy. Therefore, we can relax the definition of carrier's measure to consider only reachable cells, to rule out some bad or even unavailable cells.

Corollary 5.5.1. *Consider the criterion* F2, F3, F4. *If the carrier's measure is defined as the minimum measure among all* reachable *cells in that carrier, we can still achieve loop-freedom.*

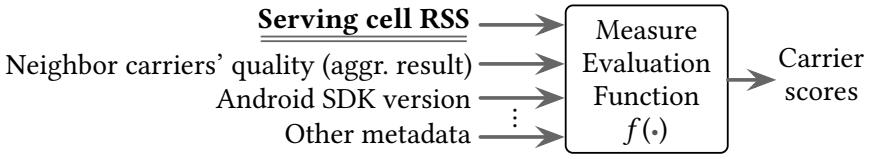


Figure 5.5: Google Fi's inter-carrier measure does not always satisfy.

```
17:54:04 K2so sorted carriers: T-Mobile, Sprint, U.S. Cellular.
17:54:04 Switch request U.S. Cellular -> T-Mobile is approved.
17:56:13 Switch done. Current network: T-Mobile
17:56:35 K2so sorted carriers: Sprint, T-Mobile, U.S. Cellular.
...
```

Trace 2: Switch decisions made by the monitor *K2so*.

5.5.1.3 Validation

We discover that Google Fi does not always follow Theorem 5.5.2 and incurs loops. Trace 2 illustrates one trace from Android's logcat. Initially, T-Mobile was evaluated as the best carrier and therefore a switch was triggered. After it was moved to T-Mobile, however, the score of T-Mobile fell behind Sprint. The reason is that, Google Fi uses a machine learning module ("*K2so*") to compute the metric to rank the carriers and select the candidate (i.e., F3 criterion). Figure 5.5 shows how the metric is computed. It is based on the *serving cell*'s signal strength, and neighboring carrier's aggregated radio qualities. However, they satisfy Theorem 5.5.2 and thus cause the loops. In practice, Google Fi eliminates switch loops by locking the device to T-Mobile. However, this is at the risk of losing network service when the device leaves the T-Mobile's coverage.

5.5.2 Inconsistency of Configurations

We next consider the scenario that inter-carrier policy and intra-carrier policy use the same measure. In this category, the stability can be violated if the threshold configurations of inter/intra-carrier policies are uncoordinated.

Table 5.3: Threshold incoordination in Theorem 5.5.3.

Criteria for $c_i^u \to c_i^v$	*F2*, with ϕ, θ	*F4*, with δ, θ
Absolute-value comparison	$\theta > Thresh1_i^{u,v} + \Delta^v \vee$	$\theta > Thresh1_i^{u,v} + \Delta^v \vee$
	$\theta > Thresh1_i^{u,v}$	$\theta > Thresh1_i^{u,v}$
Direct comparison	$\theta - \phi > \Delta^u$	$\delta + \Delta^u < 0$
Indirect comparison	$\theta > Thresh3_i^{u,v} + \Delta^v \vee$	$\theta > Thresh3_i^{u,v} + \Delta^v \vee$
	$\theta > Thresh3_i^{u,v}$	$\theta > Thresh3_i^{u,v}$

5.5.2.1 Policy Form

It is the same as §5.5.1, except that inter-carrier and intra-carrier policies use the same measure M. Moreover, as soon as a phone connects to a carrier, it will firstly camp on the cell with the highest M [11].

5.5.2.2 Stability Conditions

Given the same measures, Theorem 5.5.1 still holds, i.e., comparison criteria F1 is always loop-prone regardless of the threshold configurations. For F2 – F4, we have the following necessary conditions:

Theorem 5.5.3 (Unstable thresholds in *F2/F4*)**.** *Assume the inter-carrier policy uses* F2 *or* F4*. If the stability is violated, there must exist a carrier C_i with two cells c_i^u, c_i^v that satisfy the condition in Table 5.3.*

Theorem 5.5.4 (Unstable thresholds in *F3*)**.** *Assume the carrier's measure $M(C_j) = M^{max}(C_j)$, and inter-carrier policy uses* F3 *with offset δ. If the stability is violated, there would exist a carrier C_i satisfying: (1) There are two cells c_i^u, c_i^v in carrier C_i such that the*

[11] This is defined in 3GPP standard. It also holds for § 5.5.1, but could be omitted since the intra-measure is independent of inter-measure.

criterion used for handoff $c_i^u \rightarrow c_i^v$ *is in the form of* absolute-value*; or (2) There exists a cell sequence* $c_i^{u_1}, c_i^{u_2}, \cdots, c_i^{u_l}, l > 1$ *and each cell appears at most once in the sequence. It satisfies that* $\delta + \sum_{j=0}^{l-1} h(c_i^{u_j} \rightarrow c_i^{u_{j+1}}) < 0$ *where function* h() *is defined as:*

$$h(c_i^u \rightarrow c_i^v) = \begin{cases} Thresh3^{u,v} - Thresh2^u, & \textit{indirect comparison} \\ \Delta^u, & \textit{direct comparison} \end{cases}$$

The proofs are in Appendix B.2. We derive necessary condition of violated stability, based on the fact that intra-carrier policy redirects the phone from a "good" cell to a "bad" cell. Here, "good" (or "bad") indicates if the measure of cell is high enough to prevent switch from the current carrier (or not). Notably, both theorems imply that *aggregated intra-carrier thresholds* suffice for coordination with inter-carrier policies (elaborated in §5.7.2 and Table 5.6b). The carriers do not necessarily expose all of their per-cell thresholds to the MCSP for coordination.

5.5.2.3 Validation

We have not found real instances in this category today. Existing solutions (Google Fi) always use different measures from the intra-carrier policies, and thus will not incur such conflicts. The theorems in this section are thus serving as early guidelines for this category in the future.

5.6 Stability for Hybrid Policy

The hybrid inter-carrier policies decide the target carrier based on both pre-defined preferences, and runtime measures (and their thresholds). This section generalizes our results in §5.4–5.5 to this scenario. In combining the preferences and thresholds, there are two approaches in general:

Preference-first policy. In this approach, the MCSP will first check each candidate

carrier's preference, and evaluate its measure (via *F1 – F4*) based on the relations between their preferences and the serving carrier's (higher, lower, or equal). The idle-state intra-carrier policy (§5.3.1) belongs to this form. For each preference relation, the inter-carrier policy has the flexibility of choosing the threshold-based criterion (*F1 – F4*). But the following result shows that some criteria are unstable regardless of the threshold settings (proof in Appendix B.3):

Theorem 5.6.1 (Unstable comparison with preference)**.** *In hybrid mechanisms with preference-first, loop will happen under the following combinations of threshold-based criteria: (1) Criterion* A1 *is applied to neighbor carriers with equal preference; (2) Criterion* A1 *is applied to both neighbor carriers with higher preference and neighbor carriers with lower preference; (3) Criterion* A1 *is applied to neighbor carriers with higher preference and criterion* A3 *is applied to neighbor carriers with lower preference, or vice versa.*

Intuitively, if switch $C_i \mapsto C_j$ is completely based on the measure of C_j (as *F1*), then we should set absolute upper bound on C_j for switch-back $C_j \mapsto C_i$ to happen. Compared with Theorem 5.5.1, the use of preferences poses more constraints on selecting the threshold-based criteria.

Threshold-first policy In this approach, the MCSP uses one threshold-based criterion for all candidate carriers. For candidates that meet this criterion, the MCSP will select the one with the highest preference. In this category, coordinating the threshold suffices for stability; the preference values do not pose extra constraints. If such hybrid policy is unstable, then the corresponding threshold-only mechanism applying the same criterion and thresholds will also be unstable. The results in §5.5 still hold and can be readily applied here.

Google Fi validation. We have observed that Google Fi may apply preference-first and threshold-first policies in different scenarios. Although its preference-based policy (§5.4.1.1) and threshold-based policy (§5.5.1.1) are separate, they can be coupled by its internal per-

module priority. When the device has network access, the threshold-based policy is preferred whenever it makes a decision. If the threshold-based policy does not make a decision, the preference-based policy will be used. This corresponds to the threshold-first policy. When the device has no network access, the preference-based policy is elevated with higher priority, thus resulting in preference-first policy. Unfortunately, we have not observed real instances so far. We are in the process of collecting more traces to catch the actual occurrence.

5.7 Practical Stability Guidelines

Based on above results, we devise practical guidelines for multi-carrier access stability. We seek to achieve three goals (ordered by their importance):

- **G1: Guaranteed Stability.** We seek guidelines for any-loop-freedom under any static settings.
- **G2: Retaining policy flexibility.** In guaranteeing the stability, our guidelines should still retain high flexibilities for the MCSP and carriers to customize their policies.
- **G3: Protecting internal policies.** Intuitively, enforcing stability implies that the MCSP and carriers should share their internal policies for coordination. This is nontrivial for both technical and non-technical reasons. In regulating the policies, it is desirable to reduce the policy exposures.

In achieving them, there are two practical constraints:

- **C1: Regulating inter-carrier policy only.** Carriers may be reluctant to change their internal policies for the MCSP: These policies not only serve the multi-carrier customers, but also single-carrier customers.
- **C2: Limited visibility to intra-carrier policy.** In regulating its inter-carrier policy, the MCSP may not have full access to the carriers' internal policies.

To derive the guidelines, we start from the theoretical results in §5.4–5.6 that ensures stability (G1). We use them to regulate the inter-carrier policy only (C1), using the aggregated intra-carrier policies from carriers (G3 and C2). By considering the practical demands, we adopt these guidelines to leave sufficient flexibility for carriers and the MCSP (G2). We next elaborate these guidelines.

5.7.1 Guidelines for Preferences

We devise guidelines for different forms of preferences in §5.4.

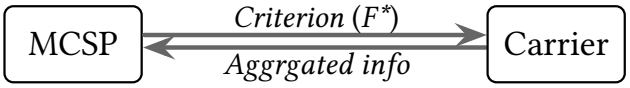
• *RAT-aware preference* (§5.4.1): For stability, Theorem 5.4.1 requires the MCSP to regulate its RAT-aware preferences based on carriers' internal thresholds and priorities. However, this requires carriers' fine-grained internal policies and thus violates G3 and C2. We thus present a practical assumption and derive Corollary 5.7.1 of Theorem 5.4.1 (proof in Appendix B.1).

Assumption 5.7.1. *Intra-carrier policy makes consistent decision for its per-cell priority and threshold. Specifically, intra-carrier will not move device to a low priority RAT from a high priority RAT, following idle-state policy in §5.3.1.*

Corollary 5.7.1 (RAT PREFERENCE CONFLICT)**.** *Under Assumption 5.7.1, also assume that the MCSP uses Policy 1. A persistent N-carrier loop happens* iff. *both the following conditions hold: (a) every carrier has one or more RATs (denoted RAT_H) assigned with equal, highest preference; and (b) in all carriers, RAT_H does not have the highest intra-carrier priority.*

Assumption 5.7.1 is commonly satisfied, unless the device is at cell coverage boundaries or receives extreme weak radio signal from high-priority RAT. Following the above corollary, we can thus lift G3 and C2 and avoid loops in the most common settings with *aggregated* carrier priorities:

Guideline 1 (Coordination via priority aggregation)**.** *If there exists a carrier C_i has and only has the most preferred RAT_H deployed, assign the highest inter-carrier preference to*



(a) Information exchange between MCSP and a carrier

Necessary aggregation info. from carriers	*F2*	*F3*	*F4*
$\min\{\min_{c^u \neq c^v} Thresh1^{u,v}, \min_{c^{u'} \neq c^{v'}} Thresh3^{u',v'}\}$	✓		✓
$\min\{\min_{c^u \neq c^v} Thresh1^{u,v} + \Delta^v, \min_{c^{u'} \neq c^{v'}} Thresh3^{u',v'} + \Delta^{v'}\}$	✓		✓
$\min_{c^u} \Delta^u$	✓		✓
$\min\{\min_{c^u \neq c^v} Thresh3^{u,v} - Thresh2^u, \min_{c^{u'}} \Delta^{c^{u'}}\}$		✓	
$\lvert\{c^u \lvert \Delta^u < 0 \vee \exists c_v, Thresh3^{u,v} < Thresh2^u\}\rvert$		✓	

(b) Aggregation items needed for coordination

Figure 5.6: Threshold coordination (The first four rows are aggregation values about thresholds of intra-policy. The last row is the number of cells with $\Delta < 0$ or $Thresh3 - Thresh2 < 0$).

it: $P_{i,H} = P_{max}$. Otherwise, inter-carrier preference assignment should be monotonic on carriers: $\forall i \neq j, P^i_{min} > P^j_{max}$ or $P^j_{min} > P^i_{max}$. The order of monotonicity is flexible but should reflect the MCSP's preference on carriers.

Guideline 1 only requires carriers to expose its maximum intra-carrier priority (G3). It still retains high flexibility of the preference settings (G2) since multiple monotonic ordering could exist. For instance, if both C_1 and C_2 have 3G and 4G, the preference order can be either of the following: $P_{1,4G} > P_{1,3G} > P_{2,4G} > P_{2,3G}$, or $P_{2,4G} > P_{2,3G} > P_{1,4G} > P_{1,3G}$. The MCSP may prefer the first ordering if C_1's service quality is generally better than C_2's. In case if C_1 only has 3G, while C_2 has 3G and 4G, the preference order is still flexible to ensure loop-free: $P_{2,4G} > P_{2,3G} > P_{1,3G}$, or $P_{1,3G} > P_{2,4G} > P_{1,3G}$.

• *RAT-oblivious preference* (§5.4.2): If the MCSP uses the RAT-oblivious preferences, the following guideline (based on Theorem 5.4.2) ensures stability and meets G1–G3 and C1–C2:

Guideline 2 (Avoid preference-unavailability conflict)**.** *Disable carriers whose intra-carrier policy can move the device to an unavailable cell.*

Guideline 2 ensures stability (G1) since it satisfies Theorem 5.4.2. It also retains high flexibility (G2): Except the disabled carriers, it allows arbitrary preference settings by the MCSP. It does not require the exposure of the carriers' internal policies (G3): Carriers only report a *binary confirmation* about whether it can move device to an unavailable cell.

5.7.2 Guidelines for Thresholds

We offer guidelines for various threshold-based policies (§5.5).

• *Inconsistency of measures* (§5.5.1): If the MCSP uses different measures from carriers, the following guideline helps the MCSP rule out the loop-prone criteria (Theorem 5.5.1):

Guideline 3 (Avoid loop-prone criteria)**.** *If the inter-carrier policy uses different measures from the intra-carrier policies, it should not use Criterion* F1 *to evaluate carriers.*

Next, the MCSP should regulate how it determines the measure for each carrier (based on per-cell measure metric). In principle, Theorem 5.5.2 provides necessary and sufficient conditions for loop-freedom. In addition, Lemma 5.5.1 gives more practical conditions to ensure loop-freedom. However, they may not be desired due to their limited flexibility in reality (G2). Consider a carrier that deploys 2G, 3G, and 4G. Using its own measure Q, the carrier may never move the device to 2G. However, based on the minimum-measure rule, the MCSP has to use 2G's measurement on M (such as latency) in determining the measure, which may be unfavorable The guideline below relaxes this constraint while still satisfying Lemma 5.5.1:

Guideline 4 (Relaxed minimum measure)**.** *Consider the inter-carrier switch policy that uses different measures (M) from the intra-carrier policies (Q) If a carrier's internal policy would only move the device to a subset of its cells (under Q), the MCSP should apply Theorem Lemma 5.5.1 to this subset.*

Compared with Lemma 5.5.1, Guideline 4 mitigates the impact of minimum measure. In the above example, if 2G is not selected in intra-carrier policy, its measures (e.g., latency) would not need be considered in inter-carrier policy either. This guideline does not require exposure of intra-carrier policy either (G3): Each carrier only reports a list of cells that its internal policy will not select.

• *Inconsistency of configurations* (§5.5.2): If the MCSP uses the same measure as the carriers', it should coordinate its thresholds for stability. In principle, the MCSP requires access to all carriers' per-cell thresholds, which however violates G3 and C2. To prevent it, we use the *aggregated thresholds* based on Theorem 5.5.3 and 5.5.4, and devise the following guideline:

Guideline 5 (Coordination via aggregated thresholds)**.** *If the inter-carrier policy takes criterion* F2 *or* F4*, set the inter-carrier thresholds to satisfy conditions in Theorem 5.5.3. If the inter-carrier policy takes criterion* F3*, set the inter-carrier thresholds to satisfy conditions in Theorem 5.5.4.*

Note that, to coordinate thresholds, the MCSP will query each carrier with a criterion (*F2 – F4*). The carrier returns aggregated information about intra-policy threshold (Figure 5.6a). The forms of those aggregation are listed in Figure 5.6b.

5.7.3 Guidelines for Hybrid Policy

If the MCSP deploys the preference-first policy, Theorem 5.6.1 offers the following guideline. Note that it does not require access to intra-carrier policy (G3), nor regulating the preferences or thresholds (G2).

Guideline 6 (Loop-prone criteria given preferences)**.** *If the hybrid inter-carrier policy uses preference-first, it should not use Criterion* F1 *and* F3 *under the conditions in Theorem 5.6.1.*

If the MCSP deploys the threshold-first policy, §5.6 has shown that Theorem 5.5.1 and

Table 5.4: Google Fi coverage.

City	Total Grids[a]	Has 4G LTE	Only 3G	Only 2G/No service
Los Angeles	122 335 (1261 km^2)	120 480 (98.48%)	1850 (1.51%)	5 (<0.01%)
St. Louis	136 350 (1295 km^2)	101 773 (74.64%)	34 574 (25.36%)	3 (<0.01%)

[*] Each grid's resolution is 0.001°, resulting in equivalently 110 m × 110 m grid.

Table 5.5: Intra-carrier policy statistics.

Cell priority	1	2	3	4	5	6
Count #	35719	3116	17300	4	11851	2698
Percentage (%)	50.5	4.4	24.5	<0.1	16.7	3.8

necessary conditions of loop in Theorem 5.5.2, 5.5.3 and 5.5.4 still hold. This implies that, the MCSP does not need to regulate the preferences, as shown in the following guideline:

Guideline 7 (Threshold-first)**.** *If the hybrid inter-carrier policy uses threshold-first, it only needs to regulate its thresholds by following Guideline 4–5.*

5.8 Validations of Guidelines

We assess the occurrence of conflicts in reality, and the effectiveness of our guidelines. Since we are not authorized to change Google Fi's policy in devices, we use trace-driven emulation to complement our real-world validations.

Emulation with operational traces. To approximate the real-world multi-carrier access, we extract our emulation parameters from the operational traces. To obtain the cell coverage, we crawled Google Fi's real coverage data [Goo18] as of 03/07/2018 for Los Angeles, CA (large city) and St. Louis, MO (mid-sized city). The statistics of the coverage data are summarized in Table 5.4. For each cell, we assign its intra-carrier priorities based on the op-

Table 5.6: Emulation settings.

Scenario	C_1.LTE ($c^{1,2,3}$), C_1.3G (c^4), C_2.LTE ($c^{5,6,7}$), C_2.3G (c^8)
Cell RSS range	LTE: $[-124, -80]$ dBm; 3G: $[-120, -75]$ dBm
Intra-priority	Enumeration of $1, 3, 5$ ordering
Intra-threshold	$Thresh1, 2, 3, \Delta$ varies
Inter-preference	75 combinations
Inter-threshold	*F2*: $\theta, \phi \in [-115, -109]$ dBm; *F3*: $\delta \in [0, 4]$ dB

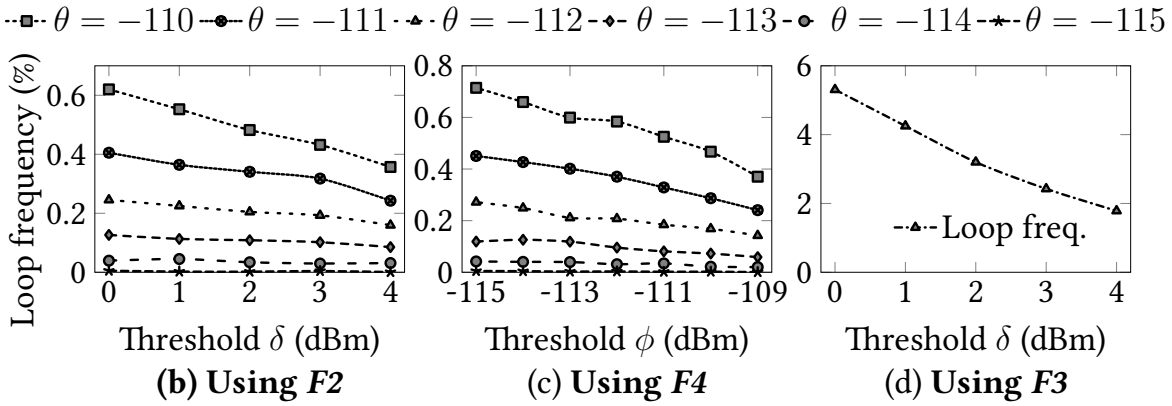
erational T-Mobile/Sprint/USCC configurations (extracted from the MobileInsight [LDX16] public dataset). Table 5.5 summarizes these priorities. We further select the most common and representative intra-carrier thresholds (the unit for θ and ϕ is dBm and that for δ is dB): $Thresh1 \in [-115, -117]$ for LTE and $Thresh1 = -108$ for 3G; $Thresh2 \in [-120, -116]$ for LTE and $Thresh2 = -108$ for 3G; $Thresh3 \in [-120, -116]$ for LTE and $Thresh3 = -114$ for 3G, $\Delta \in \{-2, 2, 3\}$. The cell signal strengths observed in the dataset range in $[-124, -80]$ for LTE and $[-120, -75]$ for 3G.

With these data, our emulation uses the settings in Table 5.6. We set two carriers, C_1 and C_2, both with LTE and 3G according to Google Fi's coverage. In the emulation, we vary each cell's signal strength according to observed range. We enumerate reasonable inter-carrier policies as follows: First, for preference policy, we enumerate all RAT-aware preference lists, which result in 75 different preference orderings. Out of the 75 different orderings, 40 are loop-prone. They fall into three categories in Table 5.7a. Second, for threshold policy, we use criteria F2, F3, and F4 (F1 is always loop-prone according to Theorem 5.5.1), and set $\theta, \phi \in [-115, -109]$ dBm for *F2*, and $\delta \in [0, 4]$ dB for *F3*. Setting is the same for *F4*. These range will not cause trivial loops. We repeat the emulation for 75 different settings of preference-policy and have 1.5 M rounds in total; For threshold-policy, we do emulation for 77 different settings involving criteria *F2, F3, F4*, and also have totally 1.5 M rounds.

Figure 5.7 shows the results.

Preference list	Loop freq. (%)
Only 3G assigned the highest pref.	6.160
One of C_1.LTE and C_2.LTE assigned the highest pref.	0.088
Both C_1.LTE and C_2.LTE assigned the highest pref.	0.003

(a) RAT-aware preference policy



(b) Threshold-based policy

Figure 5.7: Loop occurrence and evidence of guidelines.

Frequency of loops. For preference-based policy, Figure 5.7a summarizes the frequency of loops. The frequency ranges between 0.003% and 6.16%. Note that, the preference setting only with 3G assigned the highest preference is the most unstable. This setting is likely to happen if the 3G deployed as the enterprise's small cells.

For threshold-based policy, Figure 5.7b shows the frequency of loops versus configurations on θ, δ or ϕ . For *F2, F3, F4*, the frequency of loop drops as θ decreases, ϕ increases or δ increases. This is consistent with Guideline 5, thus indirectly validating the effectiveness of our guideline.

Effectiveness of guidelines. To evaluate the effectiveness of our guidelines, we rule out loop-prone inter-carrier policies following our guidelines in §5.7, and repeat the simulation under the same settings. We have validated that no loops will occur after this regulation. Moreover, as shown above, Figure 5.7b also validates the effectiveness of our guideline.

5.9 Applicability to Dynamic Policies

Dynamic policy updates. We so far assume invariant policies for MCSP and carriers (§5.3.1). It is possible that MCSP dynamically updates its policies. Our results can be generalized. Assume that stability is ensured before the update. A policy update is defined as *safe* iff. stability is still guaranteed after this update. The following proposition (proof in Appendix B.4) offers the conditions for safe preference and threshold updates, thus extending our results to the dynamic scenarios:

Proposition 5.9.1 (Safe policy update)**.** *The following inter-carrier policy updates are safe: (1) Increasing inter-carrier preferences for top-preferred carrier; (2) decreasing θ, ϕ or increasing δ in criteria* F2, F3, F4.

From the necessary condition of instability in Table 5.3, we can see the direction of adjusting parameters is towards eliminating the incoordination.

Dynamic measures. Our results are obtained by assuming the measures (e.g., signal strengths, latency) are fixed. When the measures are dynamic, our guidelines still ensure the persistent loops will not incur. Transient loops may occur (i.e., "ping-pong" loops), but can be mitigated using standard approaches (e.g., maximum attempt counters).

Mobility case. As the device moves, the inter-carrier policy also changes with locations. This can be viewed as a sequence of addition/deletion of carriers/RATs/cells (each associated with intra-carrier policies). The policy guidelines in §5.7 can thus be recursively applied in these sequences.

Our results also apply to the PLMN selection for roaming [3GP19c]. PLMN selection is a mandatory function for all commodity phones. As the device leaves the coverage of its home carrier, the PLMN selection searches the visiting carrier network based on pre-defined RAT-aware preferences. The results in §5.4 are thus applicable to regulate its stability.

CHAPTER 6

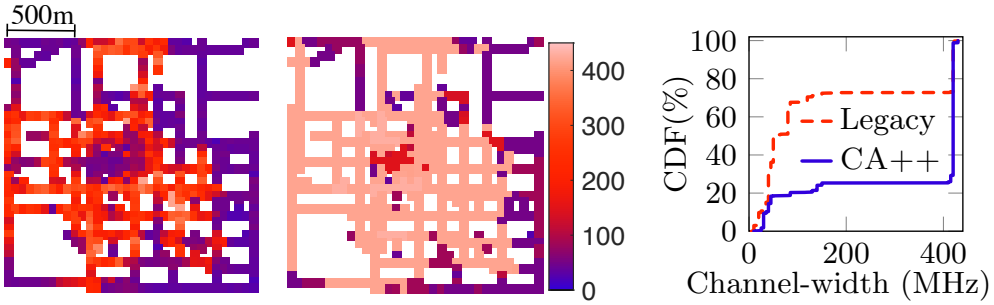
CA++: Enhancing Carrier Aggregation

6.1 Motivation and Overview

We use a measurement study with AT&T in two US cities to reveal current limitations to motivate our CA++ solution.

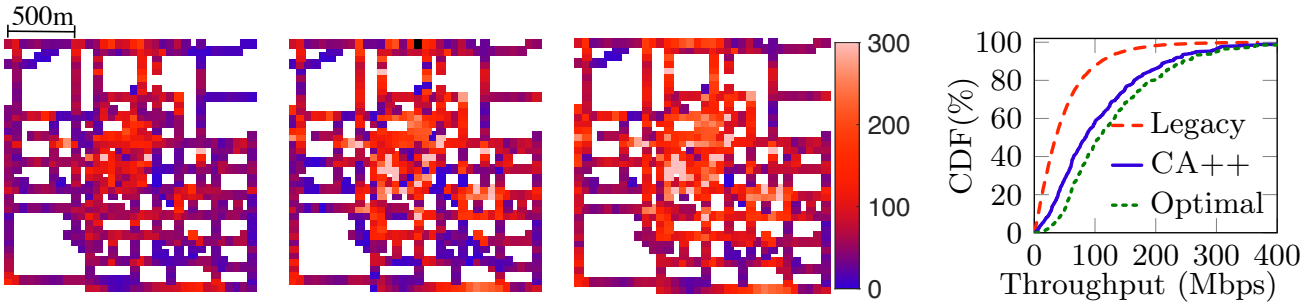
Current limitations. The fundamental issue is that current practice hardly keeps up with the increasing CA power. We have conducted extensive measurements in two cities C1 and C2 (> 4 km^2). Methodology and dataset information are detailed in §6.5 (Table 6.4).

Evidently, more spectrum resources have been added over time. Table 6.1 lists 5G/4G bands and channels observed. Since late 2019 (first 5G rollout), AT&T has acquired band n260 for 5G mmWave, repurposed 4G bands 66 and 5 partially for 5G, and added mid-band 46 for 4G. In a nutshell, AT&T increases its downlink frequencies from 258 MHz to 4033 MHz while the spectrum range expands to 700MHz – 40GHz, from up to 2.4 GHz



(a) Legacy (b) REM (c) CDF

Figure 6.1: 5G CA's frequency width (MHz) in C1.



(a) Legacy (b) REM (c) Optimal (d) CDF

Figure 6.2: Downlink data speed (Mbps) in C1.

(band 30). Figure 6.3 plots the cumulative distribution functions (CDF) of the number of candidates cells (with RSRP > a threshold) and the number of aggregated cells over all the test locations. Similar results are observed in C1 and C2. We present only those in C1. We observe more than 76, 55 and 23 cells with their RSRP (median) larger than -140dBm (all), -110dbm (good), and -90dbm (excellent) at more than half of locations in C1. We see that AT&T configures the acceptable RSRP threshold mostly in [-115dBm, -110dBm], so there are tens of good candidates available in most cases.

Given a huge number of good choices, it seems easy to fulfill the maximal CA power. However, the actual number of aggregated cells is much smaller than it can. At more than half of locations in C1, it is smaller than 4 (5G+4G, but mostly 4G), and even smaller than 1 for 5G only (5G is not used). In this study, we use Google Pixel 5 which supports CA up to 4 mmWave carriers (up to 400 = 4 × 100 MHz) and up to 8 carriers of 4G and 5G. The device's CA capability is below the network's one up to 8/16 carriers in Release-15. Therefore, it will be even harder to get good enough cells to aggregate when CA capability upgrades in the future.

Such under-utilization is more evident in the aggregated channel-width (Figure 6.1) and the resulted throughput (Figure 6.2). By comparing the current practice with our CA++ solution (elaborated later), we clearly see that the legacy solution used today largely fails to utilize CA power available. It does not use 400 MHz for 5G mmWave in more than 75% of

Table 6.1: Bands and channels used by AT&T.

Bands	n260	n66/66	n5/5	2	4	12	14	30	46*	**Total**
RAT	**5G**	**5G+4G**		**4G**						-
DL Freq	37000	2110	869	1930	2110	728	758	2350	5150	728
range (MHz)	40000	2200	894	1990	2155	746	768	2360	5925	40000
Width (MHz)	3000	90	25	60	45	18	10	10	775	**4033**
# of 5G Ch.	19	1	2	-	-	-	-	-	-	**22**
# of 4G Ch.	-	5	1	3	3	1	1	2	9	**25**

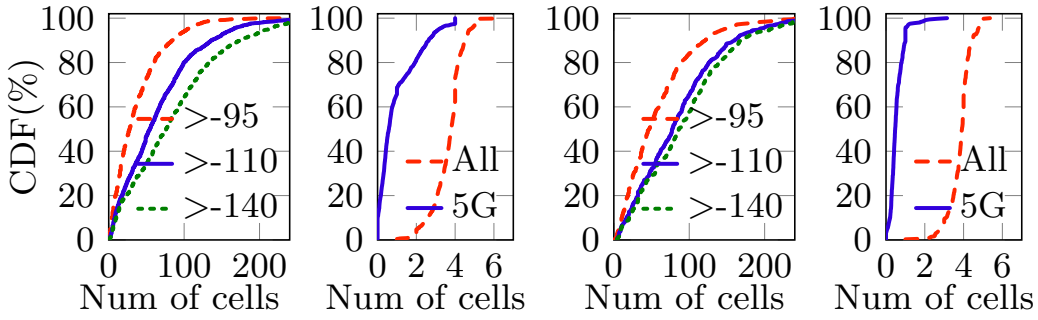
Band 46 is recently added in 2020 (after 5G rollout). 5G bands start with 'n*'.

Total spectrum width of 4G (prior to 5G): **258 MHz** (excluding bands n260 and 46).

cases, but it turns out that such spectrum resources are availableand attainable by CA++ in more than 50% cases. As a result, current practice only yields 35.4 Mbps (median) of downlink throughput (at speedtest via bulk file downloading), while the optimal CA can offer 107 Mbps.

Causes and challenges. Intuitively, it is easy to understand why from the optimization perspective. To maximize the resources to aggregate, the device needs a global view of all candidate cells' quality and assists the network to select the best N_s cells. However, the current, standardized CA is performed cell-by-cell *sequentially*, which undermines the CA potentials in two fold. First, in most cases, it limits the measurement scope without considering all the candidates; Many good candidates are never taken into account. Second, it is not to select a group of "best" cells together; Instead, cells are updated sequentially; The early-added cells restrict the scope of following cells to aggregate, thus missing better cell combinations. As a result, current practice is deemed to under-utilize the full CA power; It runs CA only on the best-effort basis, not even attempting to pursue global optimum (best performance). Even worse, the gap tends to rise as CA power constantly grows.

However, current operations are not entirely irrational. There are two inherent and important challenges.



(a) N_c (C1) (b) N_s (C1) (c) N_c (C2) (d) N_s (C2)

Figure 6.3: Number of candidate and aggregated serving cells (N_c and N_s) observed in C1 and C2.

First, measuring all cells over a wide frequency spectrum becomes less feasible with more cells and higher frequencies. The device measures one cell (channel) at a time, and hence the entire measurement time increases proportionally to the amount of candidates to measure (say, n_c), $T_{meas} \propto O(n_c)$. Constrained by OFDM symbol structure, measuring reference signals per frequency takes at least tens of milliseconds (40 ms or 80 ms, commonly configured by AT&T). In practice, it takes up to hundreds of milliseconds to get reliable measurement results, particularly on high-band cells due to Doppler. Channels are believed statistically constant (coherent) in a short window called coherence time [TV05]: $T_c \propto \frac{1}{f_m} = \frac{c}{v \cdot f}$. Here, f_m is the maximum Doppler shift, determined by the light speed c, the velocity v and the frequency f. From 2.4 GHz to 39 GHz, higher-frequency channels vary 16× faster at the same velocity, making measurement less reliable. In order not to miss good cells which would pass and change quickly, faster measurement is desired to switch the cell(s) promptly and correctly. It is unacceptable to wait for slow measurement on all (many) cells.

Second, Limiting the measurement to a very small number of cells can greatly reduce the total measurement time but raise the risk of missing good candidates. The make-up is to add more iterations at the outer loop. Even after some cells are updated, it does not stop seeking for more good cells to aggregate. However, this iterative approach suffers with one problem: cells are updated separately, without considering the quality of cells to join later.

We notice that cells cannot combine arbitrarily, constrained by co-location at the same cell tower or operator-specific policies. For example, we consistently see that certain 4G cells (at 739 MHz, band 12) never choose 5G mmWave cells regardless of their measurement results. Once such cells are used, 5G CA is unexpectedly disabled. Consequently, even if the early-added cells are good, they may restrict the scope of following cells to aggregate, and thus lead to a bad group. Moreover, the standardized mechanisms takes cell-by-cell measurement, feedback and decision, with no need of waiting for the measurement of other cells. As a result, aggregation is performed sequentially, which faces with the same risk of missing good aggregations limited by the previously-added cells.

Overview of CA++. Our ultimate goal for CA is to improve user-perceived performance (say, throughput) by aggregating frequency channels in an effective and efficient way. We devise CA++ to address aforementioned technical issues: (1) how to measure abundant frequency resources accurately and quickly? (2) How to get rid of negative impacts of sequential cell-based CA operations? As shown in Figure 6.4, we design two major components: (1) Fast and accurate inter-cell channel inference over wide spectrum (§6.2), and (2) Group-based CA operations (§6.3).

The inter-cell channel inference algorithm is to measure one (few) and infer all. We exploit a fact that aggregated cells typically reside on the same cell tower and share propagation paths. Our intuitive solution is to measure one cell, retrieve characteristics of the shared paths, and estimate quality of other cells on distinct frequencies. However, it is challenging to deliver highly accurate inference for the following reasons. First, the underlying channel model is unknown and complex, given the radio signal may propagate along multiple physical paths and combine at the receiver. Second, such inference is supposed to deliver high accuracy over a very wide frequency spectrum ranging from several hundred of MHz to 40 GHz. We transform the original measurement in time-frequency domain into the delay-Doppler domain, which enables us to separate frequency from the shared multi-path characteristics (§6.2.1). We further derive fine-grained representation for each path to enhance inference

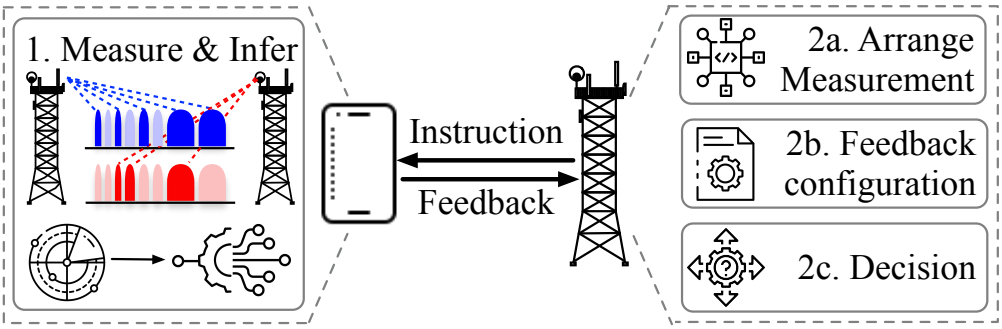


Figure 6.4: Architecture of CA++.

accuracy over wide spectrum (§6.2.2) and make inference light-weight (§6.2.3).

The fast and accurate channel inference makes it possible to replace the cell-based sequential operations with group-based manner. First, we divide measurement into groups to get rid of sequential scanning and speed up measurement (§6.3.1). Next, we enable group-based feedback and change the report condition from a single-cell form to a multiple-cell hierarchy. CA++ is thus efficient and incurs low signaling overhead (§6.3.2). Moreover, CA++ allows the network to make group-based decisions (§6.3.3).

6.2 Measure Few to Infer All

We devise inter-cell channel inference to measure one (few) and get all, thus achieving constant time measurement.

6.2.1 Inference on Delay-Doppler Domain

We perform channel inference on delay-Doppler domain because it makes it easier to infer channels at different frequencies than on conventional time-frequency domain.

As §4.2 introduces, the wireless channel is a linear combination of multiple propagation paths, represented as Equation 4.1 on delay-Doppler domain. Why delay-Doppler domain? Recall that the representation directly reveals the multi-path geometry shared by cells operating on the same base station. More importantly, the profile $\{(h_p, \tau_p, \nu_p)\}_{p=1}^{P}$ is frequency-

independent and thus makes cross-channel inference feasible (More background knowledge in §4.3.2).

How does channel inference work? A conventional inference process on the delay-Doppler domain involves three steps listed below. Assume we measure cell C_A on frequency f_A to infer cell C_B on frequency f_B.

• *Compute the channel response matrix of cell C_A.* The first step is to estimate the channel response based on the sent and received reference signals. The channel response is a 2-D matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ on the discretized delay-Doppler plane which is transformed from/to time-frequency grid with $\Delta\tau = \frac{1}{M\Delta f}$ and $\Delta\nu = \frac{1}{NT}$ [HRK18] (§4.3.1). Note that we can obtain $\mathbf{H}_A$ from measurement results using techniques in [SDA19, RPH19].

• *Retrieve the shared multi-path characteristics.* The most challenging step is to retrieve the underlying multi-path characteristics $\{(h_p, \tau_p, \nu_p)\}_{p=1}^{P}$ from $\mathbf{H}_A$. Based on [HRK18, RPH18], the relation between the channel response and the shared multi-path parameters is:

$$H[k,l] = h_w(k\Delta\tau, l\Delta\nu) = \sum_{p=1}^{P} h_p e^{-j2\pi\tau_p\nu_p} \mathcal{F}(k\Delta\tau, \tau_p)\mathcal{G}(l\Delta\nu, \nu_p) \tag{6.1}$$

$$\mathcal{F}(\tau, \tau_p) \triangleq \sum_{k'=0}^{M-1} e^{j2\pi(\tau-\tau_p)l'\Delta f}, \mathcal{G}(\nu, \nu_p) \triangleq \sum_{l'=0}^{N-1} e^{-j2\pi(\nu-\nu_p)l'T}.$$

We should exploit the relations and recover those parameters of the multi-path model.

• *Infer the channel response and quality for cell C_B.* We first recover the channel response $\mathbf{H}_B$ using the shared multi-path model and (6.1). Note that the Doppler shift is not the same, but linearly proportional to the frequency, i.e., $\nu_p^B = \nu_p^A \frac{f_B}{f_A}$. Then, it is straightforward to derive the radio quality metrics (i.e., SNR) from the channel response [San, Cab].

We focus on the retrieval of shared multi-path characteristics and present solutions next.

6.2.2 Retrieving Shared Paths Among Cells

Fine-grained representation of paths. Each path is located on the discrete DD domain with coordinates $k_p = \frac{\tau_p}{\Delta\tau}, \iota_p = \frac{\nu_p}{\Delta\nu}$ (Figure 4.6a). Note that an integer k_p is precise enough to represent the path delay, while we need a fraction ι_p for the Doppler. Specifically, the delay quantization step $\Delta\tau = \frac{1}{M\Delta f}$ is small enough to provide high resolution. So we could simplify the channel inference by approximating path delays to the nearest multiples in typical wide-band systems [RPH18, TV05]. However, one Doppler quantization step ($\Delta\upsilon = \frac{1}{NT}$) is large. In a typical 5G setting with carrier frequency $f_c = 38$ GHz, $NT = 5$ ms [1], we have the Doppler resolution $\Delta\nu = 200$ Hz, equivalent to a coarse resolution of 5.7 km/h if translated to the moving speed. Therefore, ι_p should be a fraction for accurate analysis. The problem is equivalent to retrieve $\{(h_p, k_p, \iota_p)\}_{p=1}^{P}$ for underlying paths where $h_p, \iota_p \in \mathbb{R}+, k_p \in \mathbb{Z}+$.

Decouple multiple paths by the delay. We decouple multiple paths based on a critical observation: Paths are separated along the delay axis. The channel model is complex as a combination of signal propagation effect of multiple paths. With deeper analysis, we realize that paths could be differentiated by their path delay. Specifically, each path has a distinct position l_i along the delay axis. Why is the observation true in practice? There are two reasons. (1) The paths are sparse compared to the range of delay. Given limited reflectors, the count of propagation paths is much smaller than the range of delay coordinates (typically over 200), i.e., $P \ll M$. (2) The delay quantization step is super fine-grained ($\Delta\tau$). As mentioned before, a mmWave cell of 100 MHz channel-width has the delay step of 1 ns. It is corresponding to a step of 0.3 m of path length. Such fine resolution makes path decoupling pragmatic in most cases. The quantization step (0.3 m) is much smaller than the length difference between outdoor paths. In particular, measurement studies [PLK20, WGA15] showed outdoor paths separated by over 100ns under high mobility. When paths get closer particularly at indoors, they can be combined and represented by one set of

[1] Based on the duration of reference signal in 5G (i.e., SSB burst set) [3GP19e].

aggregated parameters. Our indoor experiments validate the algorithm can still work well and outperform the start-of-the-art (§6.5.1).

Based on our insight, we derive the mathematical form of decoupling. Represent the channel response matrix using M *row* vectors, i.e., $\mathbf{H} = [\vec{\mathbf{h}}_1, \vec{\mathbf{h}}_2, \ldots, \vec{\mathbf{h}}_M]^T$. The p-th path with delay coordinate k_p is only associated with vector $\vec{\mathbf{h}}_{k_p}$. Therefore, we can decouple the multiple paths by taking the corresponding vectors. Theorem 6.2.1 describes this one-to-one mapping in a formal way, validated by our proof (details in Appendix C.1).

Theorem 6.2.1 (Path decoupling)**.** *If any two paths have different delay, i.e., $\forall i \neq j, k_i \neq k_j$, then for any vector $\vec{\mathbf{h}}_k$: (i) If there exists a path p with delay index k, then the vector $\vec{\mathbf{h}}_k$'s l-th element $\vec{\mathbf{h}}_{k,l} = Mh_p e^{-j2\pi\tau_p\nu_p}\mathcal{G}(l\Delta\nu, \nu_p), 0 \leq k \leq N-1$. (ii) Otherwise, $\vec{\mathbf{h}}_k = \mathbf{0}$.*

Extract parameters for each path. With paths separated, we extract each path's characteristics from the corresponding vector in the channel response matrix. Specifically, we invert the relation and represent the path parameters using the vector. According to Theorem 6.2.1, the p-th propagation path with delay coordinate k_p is associated with $\vec{\mathbf{h}}_{k_p}$. Therefore, we have N relations between each element in the vector and the path parameters (h_p, k_p, ι_p):

$$\vec{\mathbf{h}}_{k_p,l} = Mh_p e^{-j2\pi\tau_p\upsilon_p} \frac{l - e^{-j2\pi(l-\iota_p)}}{l - e^{-j\frac{2\pi}{N}(l-\iota_p)}}, l = 0, \ldots, N-1. \tag{6.2}$$

Then, we solve those equations to recover the p-th path's parameters from the vector $\vec{\mathbf{h}}_{k_p}$. Note that we already know the index of path delay, say k_p, by path decoupling. For the Doppler shift, we divide $\vec{\mathbf{h}}_{k_p,0}$ by $\vec{\mathbf{h}}_{k_p,\frac{N}{2}}$ and get:

$$\iota_p = \frac{N}{\pi}\left(x\pi \pm \arg\cot\left|\frac{\vec{\mathbf{h}}_{k_p,0}}{\vec{\mathbf{h}}_{k_p,\frac{N}{2}}}\right|\right). \tag{6.3}$$

where x is an integer. Given the range $0 < \iota_p < N$, (6.3) brings two possible values in that range. Then we filter out the wrong solution by validation with expressions of $\vec{\mathbf{h}}_{k_p,0}$ and

$\vec{\mathbf{h}}_{k_p,1}$. Specifically, the correct ι_p^* satisfies:

$$\left|\sin\frac{\pi}{N}\cot\frac{\pi\iota_p^*}{N}-\cos\frac{\pi}{N}\right|=\left|\frac{\vec{\mathbf{h}}_{k_p,0}}{\vec{\mathbf{h}}_{k_p,1}}\right|. \tag{6.4}$$

Finally, we derive path attenuation h_p.

$$h_p=\frac{1}{M^2N^2}\sqrt{\sum_{l=0}^{N-1}|\vec{\mathbf{h}}_{k_p,l}|^2}. \tag{6.5}$$

We prove all of the inverse relations in Appendix C.2, C.3.

6.2.3 The Algorithm

After solving the most critical issue, i.e., retrieval of the shared multiple paths, we devise an algorithm to conduct the inter-cell channel inference. Algorithm 2 elaborates on the step-by-step process. Initially, we get the channel response matrix $\mathbf{H}_A$ of cell C_A on frequency f_A via signal collection and processing. The goal is to recover the channel matrix $\mathbf{H}_B$ of co-located cell C_B on frequency f_B and then estimate its radio quality. Based on the theorem of path decoupling (Theorem 6.2.1), each vector with non-zero values in the original channel matrix reveals one propagation path (Line 2). We separate paths by taking the corresponding non-zero vectors and deriving the parameters (Line 3-7). Fractional Doppler is considered to improve inference accuracy. Next, we project the shared characteristics onto C_B on frequency f_B. Path attenuation and delay (h_p, τ_p) are invariant of frequency and thus remain the same. The Doppler parameter ν_p^B is linearly proportional to the frequency and thus equals to $\nu_p^A \cdot \frac{f_B}{f_A}$ (Line 6). With parameters of all paths extracted and mapped to frequency f_B, we reconstruct the channel matrix $\mathbf{H}_B$ to infer (Line 9). Last, we estimate C_B's radio quality (Line 10) which can replace the actual measurement.

Complexity. The algorithm is very efficient with polynomial computation complexity of $O(NMP)$. The overhead mainly comes from channel matrix reconstruction for the cell to infer, which dominates the cost of retrieving shared multi-path ($O(P)$). Our algorithm is

Algorithm 2 Inter-cell Channel Inference

Input: Cell C_A's channel matrix $\mathbf{H}_A$, frequency f_A, delay-Doppler grid $M_A, N_A, \Delta\tau_A, \Delta\nu_A$; Co-located Cell C_B's frequency f_B, grid setting $M_B, N_B, \Delta\tau_B, \Delta\nu_B$

Output: C_B's radio quality

$p = 1, \mathbf{P} = \emptyset$;
for each column vector with non-zero values in $\mathbf{H}_A$ **do**
    $k_p \leftarrow$ the index of the vector; $\tau_p \leftarrow k_p \Delta\tau_A$;
    Derive ι_p based on (6.3-6.4); $\nu_p^A \leftarrow \iota_p \Delta\nu_A$;
    Derive h_p based on (6.5);
    $\nu_p^B \leftarrow \nu_p^A \cdot \frac{f_B}{f_A}$;
    $\mathbf{P} \leftarrow \mathbf{P} \cup \{(h_p, \nu_p^B, \tau_p)\}, p \leftarrow p + 1$;
end for
Calculate $\mathbf{H_B}$ based on (6.1) and path parameters in $\mathbf{P}$, and the grid setting $M_B, N_B, \Delta\tau_B, \Delta\nu_B$;
Calculate SNR/RSRP/RSRQ of C_B;

much faster than solutions in the TF domain [Vas16] which are based on non-convex optimization. It also outperforms the DD-domain algorithm [LLZ20] by reducing the processing cost by a factor of $\frac{\max(N,M)}{P}$.

Resolved challenges. Our algorithm has addressed aforementioned challenges about accurate channel inference (§6.1). First, the algorithm can handle the complex channel models without any strong assumptions about the underlying paths. Because it decouples the multiple paths by exploiting the separation along the delay dimension. Second, the algorithm is robust to wide frequency spectrum given the accurate inference with fractional Doppler. Note that huge frequency gaps would exaggerate inference error. For example, Doppler shift is mapped from 700 MHz to 39 GHz by multiplying a factor >50. Therefore, the fine-grained analysis considering fractional Doppler would enhance accuracy.

Last but not least, our algorithm is also applicable to cells with diverse grid settings.

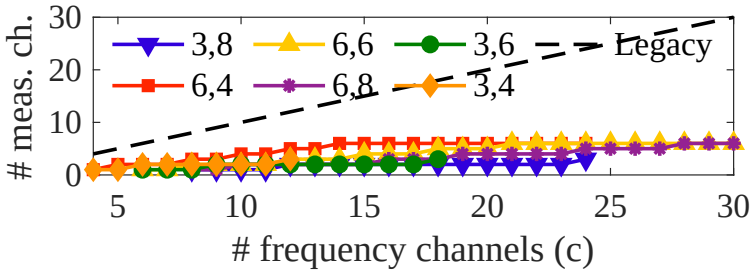


Figure 6.5: Frequency to measure (Legend for $Q, \overline{k}$).

Table 6.2: Group-based criteria.

Type	Group-based event
Update group	***(G-P)*** Neighbor eligible as PCell, $Q_n > \theta_{G1}$
	(G-S) Neighbor eligible as SCell, $Q_n > \theta_{G2}$
Update SCell	***(U)*** SCell worse than threshold, $Q_s < \theta_U$

Table 6.3: Cell setting.

Cell	P	S	Cell	P	S	Cell	P	S
A1	×	✓	B1	✓	✓	B5	×	✓
A2	×	✓	B2	×	×	B6	×	✓
A3	×	✓	B3	×	×	B8	×	×
A4	×	✓	B4	✓	✓	B8	×	×

Specifically, cells may have different symbol duration T, sub-carrier Δf, as well as grid size M, N. As a result, the transformation to delay-Doppler domain brings diverse settings. In this aspect, sub-6GHz and mmWave cells are usually different. Our algorithm inherently handles the diversity by maintaining the *absolute values* of multi-path parameters and transforming them to different representations for different grid settings.

6.3 Group-based CA Management

CA uses a group of cells to serve the devices. To search for best CA, groups of cells should be considered together. However, the existing CA operations take the conventional single-cell basis, thereby impeding the effectiveness. To solve this, we replace them with group-based management: to measure, report and decide on cells together if they can be aggregated (i.e., in the same CA group).

6.3.1 Group-based Measurement

The channel inference algorithm accelerates measurement of cells in the same CA group.

Considering multiple CA groups to examine nearby, we need to arrange which cell(s) to measure and which to infer jointly. The arrangement aims to minimize the number of frequencies to measure and thus maximize the acceleration.

We explain how the arrangement is expected to work. At one location in our experiment, we observed two 5G base stations, BS-A and BS-B. BS-A carries 4 cells on mmWave frequency channels denoted as F1 - F4. BS-B has 8 mmWave cells: 4 of them on the same 4 frequencies as BS-A; the other 4 are denoted as F5 - F8, respectively. By measuring only *one* frequency shared by both, like F1, the device could infer all other mmWave cells. And that arrangement leads to the optimal acceleration by group-based measurement.

Problem definition. CA++ is supposed to minimize the frequency channels to *physically* measure. The measured frequencies should cover all CA groups nearby. Our intuition is to deduce this practical issue from an equivalent set cover problem. We formalize the problem in Proposition 6.3.1.

Proposition 6.3.1. *To find the minimum number of frequency channels to measure is equivalent to the following set cover problem. A universal set* $\mathbf{S} = \{BS_1, BS_2, \ldots, BS_Q\}$ *represents* Q *neighbor base stations. Base station* i *carries* k_i *frequency channels. There are* c *unique frequency channels; each frequency* f_j *can be represented by a non-empty subset* $\mathbf{S}_j \subset \mathbf{S}$ *including base stations that carry the frequency. We have* $\max_{1 \le i \le Q} k_i \le c \le \sum_{i=1}^{Q} k_i$. *In addition,* BS_i *occurs* k_i *times among all subsets. Hence, minimizing the number of frequencies to measure is equivalent to a set cover problem, i.e., to find the minimal index sets* $\mathbf{I} \subset \{1, 2, \ldots, c\}$ *s.t.* $\bigcup_{i \in I} \mathbf{S}_i = \mathbf{S}$.

Algorithm and effectiveness. We seek for an approximation of the optimal solution, as the set cover problem is NP-hard. We apply a greedy algorithm, MinFreqNum (Algorithm 3), to the arrangement of group-based measurement. The algorithm incurs low computation overhead, with time complexity of $O(\overline{k}Q^2)$. For the effectiveness of minimization, Theorem 6.3.1 gives the upper bound of the number of frequency channels to measure

Algorithm 3 MinFreqNum: Minimizing Frequencies to Measure

Input: A universal set $\mathbf{S} = \{BS_1, BS_2, \ldots, BS_Q\}$, subsets $\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_c\}$

Output: Index set $\mathbf{I}$ representing selected subsets

1: $\mathbf{I} \leftarrow \emptyset, \mathbf{X} \leftarrow \mathbf{S}$

2: **while** $\mathbf{X} \neq \emptyset$ **do**

3: Let i be the index maximizing $\mid \mathbf{X} \cap \mathbf{S}_i \mid$

4: $\mathbf{I} \leftarrow \mathbf{I} \cup \{i\}, \mathbf{X} \leftarrow \mathbf{X} \setminus \mathbf{S}_i$

5: **end while**

(proof in Appendix C.4).

Theorem 6.3.1. *Assume Q base stations and the base station BS_i carries k_i channels. There are c unique frequencies ($\max_i k_i \leq c \leq \sum_i k_i$) in total. The number of channels* MinFreqNum *has to physically measure is no greater than* $\min\left\{\left\lfloor \frac{log\left(Q - \max_{1\leq j\leq c} |\mathbf{S}_j|\right)}{log\frac{c}{c-k_m}} \right\rfloor + 2, Q\right\}$, *where* $k_m = \min_{1\leq i\leq Q} k_i$.

To show the effectiveness in a more straightforward way, Figure 6.5 compares the upper bound of frequencies to measure by MinFreqNum and legacy mechanism. We use three parameters to characterize the deployment of base stations and frequencies: $Q, \overline{k}, c$ represent the number of base stations, the *average* number of frequency channels carried by one base station, and the total number of frequency channels nearby. The legacy approach has to measure c frequency channels to get the global view, invariant with $Q, \overline{k}$. For MinFreqNum, we consider practical settings of $Q = 3, 6$ and $\overline{k} = 4, 6, 8$ based on empirical observation. MinFreqNum could reduce the measurement time by 2.3 - 11.5 $\times$ *at least*. The efficiency depends on whether the base stations have more similarity (small c) or more heterogeneity (large c) in frequency channels. For example, measuring a frequency deployed by all base stations (more similarity) would cover all cells. In an extreme case of more heterogeneity, base stations do not share frequency channels. The device has to measure one frequency per base station. The measurement is still accelerated by a factor close to $\overline{k}$, as the number of base stations is less than the number of frequencies by $\overline{k}$ times.

6.3.2 Group-based Feedback

5G standards organize measurement reports by frequencies, which is designated for conventional single-cell operations. However, it would incur high signaling overhead for CA as cells to aggregate have different frequencies. To reconcile with the group-based cell usage, cells which can be aggregated should be reported together. It is a necessity to facilitate group-based selection and reduce the signaling overhead.

Trigger reports with group-based criteria. We devise group-based criterion to determine eligibility of reporting (listed in Table 6.2). To assist the switch of whole cell group during handover, we propose a hierarchical criterion with condition ***G-P***(rimary) and ***G-S***(econdary). They specify the minimum requirement to become PCell and SCell, respectively. A cell group is reported if and only if one or more cells meet the criterion ***G-P***; the report also includes all SCell candidates, i.e., any cell that meets the criterion ***G-S***.

Moreover, we should also support reports for SCell update w/o PCell change. Criterion ***U*** is created for this purpose. Generally, it is triggered when any SCell becomes weaker than a threshold. In addition, the report should also include co-located cells which are qualified to substitute it (i.e., above that threshold). Note that unlike the legacy mechanisms, CA++ could generate reports without physically measuring SCell candidates with Algorithm 2. All candidate cells that resides on the same base station can be inferred.

Benefits of group-based reporting. The reformation brings three advantages. First, it supports group-based cell selection to fulfill CA potential. The conventional mechanism adopts single-cell selection, and naturally incurs sequential aggregation. Second, it still holds on to connectivity as the baseline, as a qualified PCell candidate is required for a selected cell group. Finally, CA++ would greatly reduce the signaling overhead. The conventional report on frequency basis only carries information to acquire one serving cell [3GP19h]. Therefore, it incurs excessive signaling to track the real-time quality of prolific CA candidates. In addition, single-cell-based feedback might not meet a group's eligibility and cause

unnecessary reports.

We use examples to illustrate signaling overhead saved by group-based reporting. We reuse the setting of BS-A and BS-B in the real trace (§6.3.1). The device is selecting a new serving cell group from one base station. BS-A has 4 cells A1-A4 on frequency F1-F4. BS-B has 8 cells B1-B8 on frequency F1-F8. The network sets the criteria for PCell and SCell, respectively; Table 6.3 lists whether each cell satisfies the conditions. CA++ combines two conditions above for PCell and SCell to generate the group-based criteria, while the legacy mechanisms have to configure *both* criteria overall all frequencies. In this example, the legacy feedback mechanism incurs 9 reports in total, 2 for PCell candidates plus 7 for SCell. In contrast, the group-based form needs *one* report to include the only eligible group (belonging to BS-B).

The example in Table 6.3 also illustrates the efficiency of SCell update. We assume the group on BS-B is selected, with B1 as PCell and B4-B6 as SCells. Upon a time, the radio signal of B4 becomes worse than required, which triggers reports and scanning for substitute. The legacy mechanisms have to check each frequency until finding a qualified cell. Consequently, it incurs three rounds of negotiation before SCell replacement: to try B2, B3 and finally B7. The group-based form (event **U**) includes the status of all eligible co-located cells in the initial report. Therefore, the network can react immediately and thus avoid extra signaling.

6.3.3 Group-based Decision

CA++ have cleared roadblocks to make group-based decisions. With accurate radio quality of neighbor cells reported in groups, the network can assess each group as a whole and choose multiple cells concurrently. Typically, the network could adopt the strategy to maximize the aggregated channel-width. For example, CA++ prioritizes mmWave cells for decision making to use more spectrum resources as possible. In practice, operators have the flexibility to take any policy as needed. It is a complicated and independent topic out of the scope of this

work (more discussion in §10.3). We would highlight that, any group-based decision scheme must be empowered by CA++ with the prompt, precise, and adequate cell information organized in groups.

6.4 Implementation

We have implemented CA++ with two USRP X300 operating as the base station and the mobile client.

Delay-Doppler domain. CA++ is built on the delay-Doppler domain at the client and the base station. We adopt Orthogonal Time Frequency Space (OTFS), a delay-Doppler modulation scheme, onto PHY-layer measurement blocks. Specifically, 5G NR uses SS/PBCH Block (SSB) for measurement [3GP19e]. In general, multiple SSBs are transmitted continuously, which form an SSB burst (spanning 1 - 5 ms and 240 frequency subcarriers). CA++ places OTFS grids over the SSB burst for measurement. We implement OTFS on top of the existing OFDM operations, through the transformation of SFFT (OFDM $\rightarrow$ OTFS) or inverse SFFT (vice versa) [HRK18]. Note that OTFS can co-exist with OFDM as the former scheme is used for measurement signals only and the latter for other data types. In this way, the operational network can avoid prohibitive cost of altering modulation on the entire PHY layer.

Channel inference. This is the core component at the mobile client to enable CA++. We implement the inter-channel inference algorithm (§6.2) and finally use SNR to represent channel quality. This is because SNR reveals the equivalent channel condition independent of modulation scheme [TV05]. Considering that OFDM probably persists for data/control message transmission, the results based on OTFS modulation are supposed to hold for OFDM data blocks as well. Note that channel interference and noises may result in inaccurate $\mathbf{H}$, which causes false positive paths derived from non-zero vectors. We thus drop the weak paths whose attenuations are 30 dB lower than the strongest one.

Group-based CA management. At the base station, we implement group based measurement, feedback and decision (§6.3). Note that CA++ is extensible to dual connectivity (DC) [3GP21] which is a special case to aggregates cells on two separate base stations. In the current transition period, DC typically involves two connections of 4G and 5G, respectively. To seek for a good combination over DC, our implementation makes the following extensions: (1) It arranges group-based measurement for dual connections (4G and 5G) separately; (2) In the case of 4G-5G aggregation, we extend the group-based feedback with different criteria for 5G and 4G cells.

6.5 Evaluation

In this section, we first evaluate the key components of CA++: accuracy of inter-cell channel inference (§6.5.1) and effectiveness of group-based CA management (§6.5.2). Then we assess the overall improvement for CA (§6.5.3).

Methodology. Since we cannot deploy CA++ on operational networks, we conduct trace-driven emulation on our testbed with software-defined radio (USRP X300) based client and base station. We run experiments under unlicensed 5450/5550MHz band. As the equipment can not operate on mmWave channels, we replay the 5G traces recently collected in AT&T to approximate real-world environment including channel configurations, radio quality, frequency spectrum deployment and CA settings. On top of that, we perform the evaluation over wide frequency spectrum[2]. Moreover, we also emulate the performance of CA++ under high mobility with a public high-speed train dataset [WZN19].

Datasets. Table 6.4 presents three datasets used for evaluation: (1) **5G-C1 and 5G-C2**: We run walking and driving tests at two Midwest cities (C1 and C2) in the US, where AT&T 5G has deployed both sub-6GHz and mmWave channels in our test regions (downtowns). We run bulk downloading to examine CA performance, as well as ping traffic to collect cell radio

[2]We will acquire advanced equipment to run experiments with mmWave in the near future.

Table 6.4: Datasets.

Dataset	**5G-C1**	**5G-C2**	**4G-HST** [WZN19]
Date	Apr 2021 - Dec 2021 (288 hr, $\sim$ 3,802 km)		Oct 2018 - Nov 2018
Region	1.65×1.85 km^2	1.2×1.0km^2	1,300-km railway
Speed (km/h)	driving: 10-40 (mostly); walking: <5		300 - 350
Operator & RAT	AT&T, NSA 5G + 4G		China Unicom, 4G
# Aggr. carriers	1 - 6	1 - 7	1 - 3
# CA groups	3,310	2,416	534
CA channel-width	5 - 430 MHz	5 - 445 MHz	5 - 50 MHz
# 5G serving cells	sub-6: 45	32	N/A
	mmWave: 353	36	
# 5G channel	sub-6: 3 (826–2116M)	2 (same)	N/A
(freq. range, Hz)	mmWave:16 (38.6–39.5G)	9 (same)	
5G channel-width	sub-6: 5 MHz	5 MHz	N/A
	mmWave: 100 MHz	100 MHz	
# 4G serving cells	809	482	1,910
# 4G Ch (range)	20 (709 – 5824 MHz)	18 (same)	8 (1740 – 2155 MHz)
4G channel-width	5, 15, 10, 20 MHz		5, 15, 10, 20 MHz

quality measurements. Since we need sufficient data to know cell deployment and CA usage completely, we repeat extensive experiments to scan test regions (total over 3,800 km and 288 hr). (2) **4G-HST**: a public High-Speed-Train (HST) dataset [WZN19]. It was collected on the HST commuting between Shanghai and Beijing, China. By the time of collection, there was no 5G; 4G CA just started at its early stage.

6.5.1 Inter-cell Channel Inference

We run experiments with the testbed at 14 indoor locations (Figure 6.6) while walking/running. The blue icon, yellow circles and red circles show the placement of the base station, LOS and

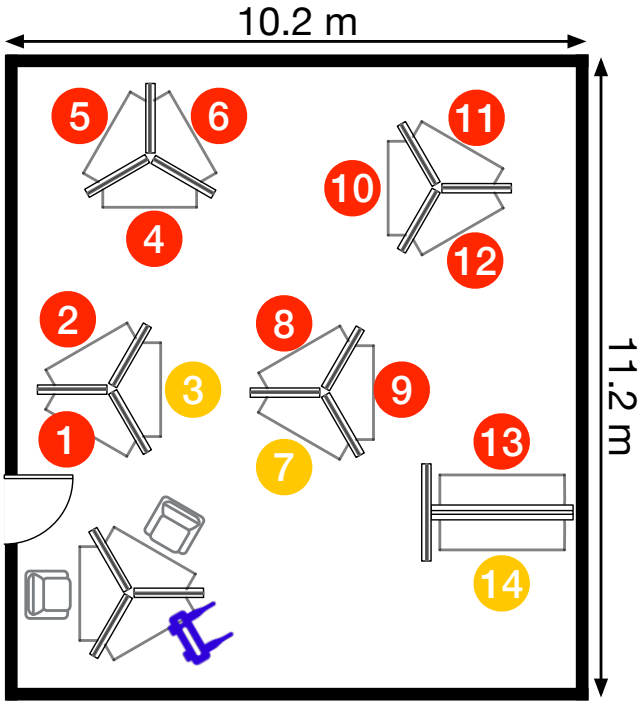
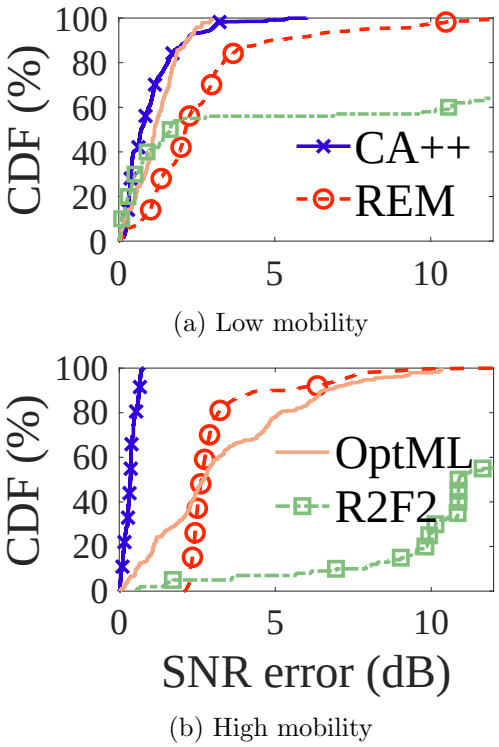


Figure 6.6: The floor plan.



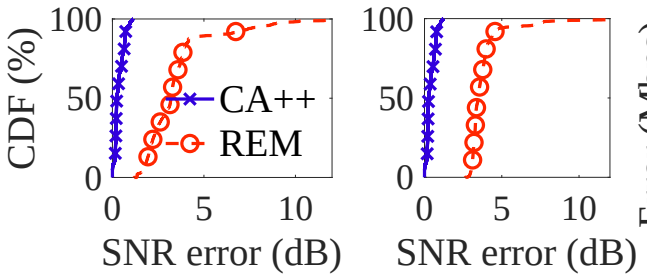
(a) Low mobility

(b) High mobility

Figure 6.7: Sub-6GHz.

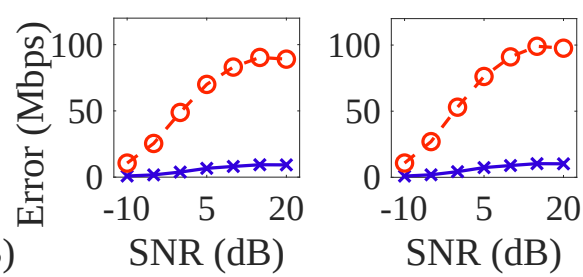
NLOS clients, respectively. We evaluate the channel inference algorithm under four different settings: narrow (within sub-6) or wide frequency range (between sub-6 and mmWave), low (10-100 km/h) or high mobility (> 200 km/h). Note that the testbed supports experiments within sub-6 bands at low moving speed. For mmWave or high mobility, we perform the evaluation based on the propagation model extracted from the testbed traces and replay the traces with the cell settings and mobility settings extracted from the 5G and HST dataset. We compare CA++ with the state-of-the-arts in both delay-Doppler domain, REM [LLZ20], and time-frequency domain OptML [Bak19] and R2F2 [Vas16]. Both OptML and R2F2 require setting the maximum number of paths for better accuracy. For fair comparison, we test and use their optimal configuration.

Channel inference accuracy. We first compare the channel quality inference by assessing the estimated SNR error. We use SNR as it is convertible to RSRP/RSRQ. For inference



(a) Low mobility (b) High mob.

Figure 6.8: Sub-6GHz→mmWave.



(a) Low mob. (b) High mob.

Figure 6.9: Error of data rates.

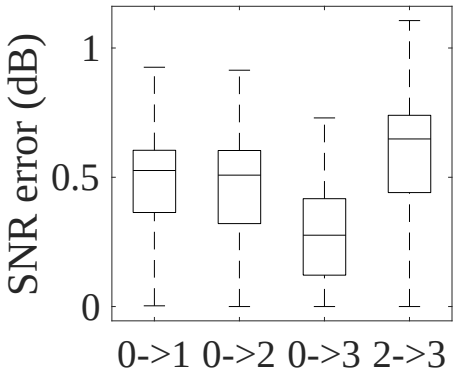


Figure 6.10: Settings.

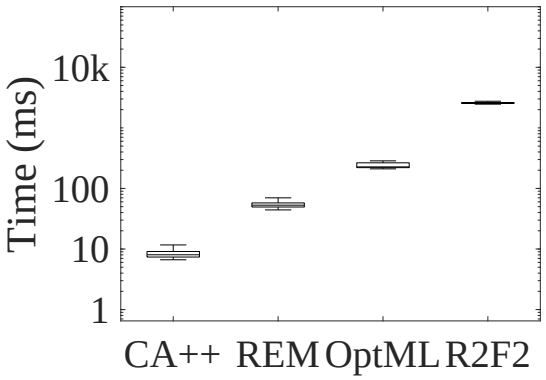


Figure 6.11: Runtime.

within sub-6GHz, CA++ incurs median error of 0.72 dB (0.36 dB) under low (high) mobility (Figure 6.7). It outperforms all state-of-the-arts. As comparison, REM, OptML, and R2F2 incur 2.14 dB (2.63 dB), 1.14 dB (2.71 dB), 1.64 dB (10.89 dB) median error under low (high) mobility. CA++ performs better under high mobility since the Doppler spreads for different paths become more significant. All three alternatives cannot provide accurate estimation as they fail to capture or precisely represent the time-varying Doppler.

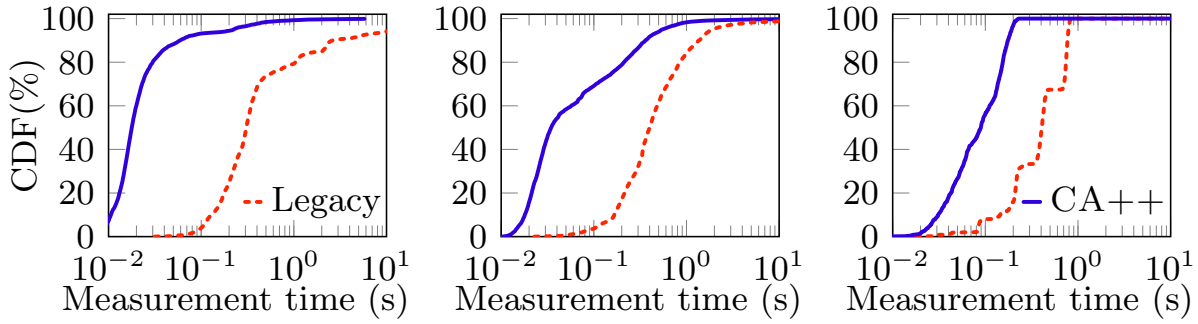
The supremacy of CA++ persists when performing channel inference across sub-6GHz and mmWave cells. Figure 6.8 presents the inference error. CA++ outperforms REM by reducing the error by a factor of 9.0, from 3.16 dB to 0.35 dB. REM fails since the frequency gap between frequency-to-measure and frequency-to-infer is too large such that

a small Doppler estimation error will be exaggerated. We do not compare with OptML or R2F2. Both algorithms cannot work normally over wide frequency spectrum, since they are designed for estimation between close frequencies.

To quantify how channel inference error affects achievable data rate, we gauge the estimation error of data rate based on the SNR inference accuracy. We use the standard SNR to spectral efficiency mapping to estimate achievable data rate in 5G [3GP19f]. The base stations decide the spectral efficiency by adapting modulation to received radio quality reports. Figure 6.9 compares CA++ with REM by data rate estimation error for mmWave scenario. If the estimation is wrong, the base station would aggregate cells with overestimated or underestimated quality, which causes under-utilization. We use 100MHz setting to assess how the data rate deviates from the ground truth with CA++ or REM under different mobility and SNR. Under different mobility, CA++ continuously outperforms REM and reduces the error by 88% - 92%.

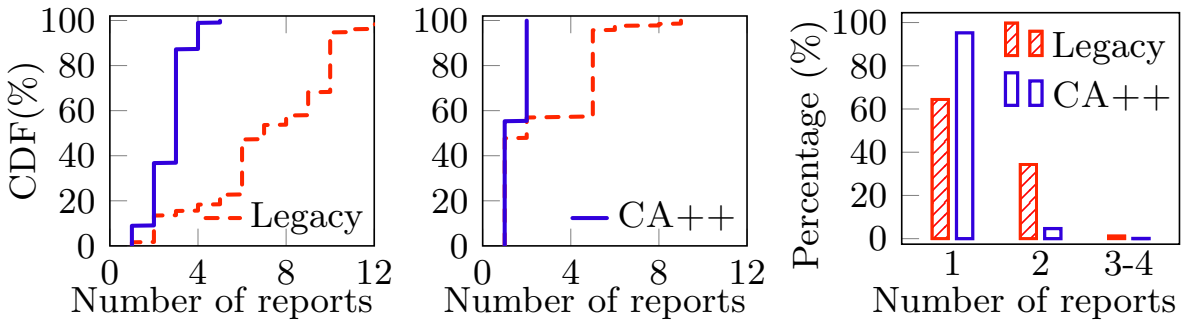
Robustness to different settings. CA++ is robust to various numerology settings. 5G supports 4 numerology, with 15kHz, 30kHz, 60kHz, 120kHz as subcarrier spacing (denoted with numerology index 0, 1, 2, 3). We test the inference with 15kHz→30kHz, 15kHz→60kHz, 15kHz→120kHz, and 60kHz→120kHz settings respectively for high mobility scenario as shown in Figure 6.10. CA++ is able to control the error median within 1dB for cross-numerology channel inference.

Efficiency. We compare the efficiency of all four algorithms by measuring the time needed to get channel inference using the same set of data. Figure 6.11 shows that CA++ takes 8 ms on average, compared with 52 ms by REM and 238 ms by OptML and 3.5 s by R2F2. CA++ and REM outperform OFDM-based algorithms by more than an order of magnitude as they do not rely on optimization with many iterations. Compared to REM, CA++ further reduces the execution time by decoupling the sparse propagation paths and performing inference separately. Such efficient inference makes CA++ promising to accelerate the measurement; The processing is faster than measuring one frequency in 5G,



(a) 5G-C1 (b) 5G-C2 (c) 4G-HST

Figure 6.12: CA measurement acceleration.



(a) 5G-C1 (b) 5G-C2 (c) 4G-HST

Figure 6.13: Reduction of signaling overhead.

i.e., 40 or 80 ms.

6.5.2 Group-based CA Management

We evaluate the effectiveness of group-based CA management (§6.3) with the following two micro-benchmarks.

Measurement acceleration. We first estimate how much the group-based operation could speed up measurement. We compare CA++ with the legacy mechanism, using the average time to detect one cell *eligible* as a serving cell. For the current mechanism, we can directly extract this metric for each CA instance from real traces. For CA++, we replay

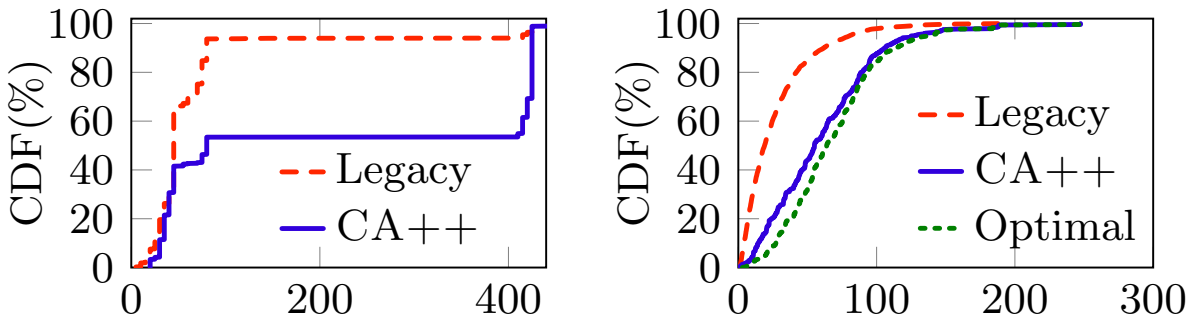
the dataset to collect eligible cells and emulate the measurement time using our algorithm. Here, we define cells' eligibility based on whether the cell is ever added as a serving cell.

Figure 6.12 shows the effectiveness of measurement speed-up by CA++. Under low or moderate mobility in 5G-C1 (C2), the median time to detect one eligible cell is 307ms v.s. 18ms (399ms v.s. 35ms) regarding the legacy mechanism and CA++. The acceleration is by 15.4× (7×) as for median, and at least 4.7× (1.9×) for more than 90% of cases. In the scenario of high mobility under 4G, the legacy operation and CA++ take 405 ms and 89 ms at median. CA++ speeds up by more than 4.4× at half cases. By comparison across databases, CA++ could achieve faster measurement as the network deploys more frequencies and cells. CA++ performs better in 5G (C1/C2) than 4G (HST). Given more cells deployed in C1 than C2, CA++ expedites the measurement more in the former region. The results confirm that our design is promising not only in 5G, but also in line with the future network.

Signaling efficiency. We examine the efficiency of group-based feedback in terms of signaling overhead. We compare the number of reports needed to include all eligible cells under the legacy mechanism and CA++. Figure 6.13 proves that CA++ greatly reduces the signaling overhead. Under low/moderate mobility in 5G-C1 (C2), the legacy mechanism incurs 7 (2) reports as the median while CA++ needs only 3 (1) reports. The signaling cost is reduced by a factor of 2.5 (2) as median, up to 8 (4.5). On the high-speed train under 4G (Figure 6.13c), CA++ only needs 1 report in 95.2% of cases while the legacy operations take at least 2 reports for 35.6% of time. Moreover, CA++ is more efficient when denser cells are deployed and more feedback are required for CA.

6.5.3 Overall Improvement by CA++

Since the device cannot know other CA groups at runtime besides the active one, we use historical data to learn available groups and their quality across the experimental region. Based on the profile, we perform a what-if study to compare the potential options enabled



(a) Channel-width (MHz)

(b) Throughput (Mbps)

Figure 6.14: Improvement for CA at city C2.

by CA++ and the actual serving CA group. Note that this evaluation is only feasible for our 5G datasets (C1 and C2); We cannot build the CA profile from 4G-HST dataset due to limited experiments.

Aggregated channel-width. In the what-if study, we compare the widest aggregation captured by CA++ to the actual serving cell group. Assume CA++ can only aggregate cells which are observed to work together in real traces. This rule complies with the capacity of mobile clients and restrictions on the network side. Figure 6.1 and 6.14a present the results in two cities. We see that CA++ would greatly enlarge the aggregated channel-width. In C1 (Figure 6.1c), 74.7% of cases have channel-width over 400MHz, compared to 27.3% before. The ratio grows from 6.0% to 46.5% in C2. City C1 witnesses higher growth than C2 because more 5G frequency channels and cells, especially mmWave, are deployed in the former region (Table 6.4). Figure 6.1a & 6.1b further compare the distribution of CA channel-width in C1, under the current mechanism and CA++. Each location is represented by the *average* channel-width of all instances. CA++ enforces aggregation over 400 MHz at 55.7% of locations, while the legacy one achieves it at only 7.5% of locations.

Throughput boost. Next, we examine how much CA++ can boost network performance in terms of throughput. In the what-if study, we retrieve the average throughput of each cell group at all locations. To emulate CA++'s decision on the serving cell group, we

assume it would prioritize mmWave cells to acquire high channel-width. Figure 6.2d and 6.14b compares the throughput of CA++ with the legacy mechanism in city C1 and C2, respectively. We also take the best observable performance as the optimal solution. By median values, CA++ would increase the throughput from 35.4 to 83.7 Mbps, with the optimal of 107.0 Mbps in 5G-C1. In 5G-C2, the throughput grows from 29.1 to 54.0 Mbps, and the best performance is 59.8 Mbps. By comparing the optimal results, 5G-C1 has higher upper bound of performance thanks to denser deployment of frequency channels of cells.

Meanwhile, the gap between CA++ and the optimal goes larger in city C1. We further examine the gap in the map of *average* performance per location (Figure 6.2a & 6.2b). CA++ benefits 73.9% of locations, and the median increase is 29.2 Mbps (1.6×). Meanwhile, CA++ downgrades the performance at 16.1% of locations, with median loss of 19.3 Mbps (1.8×). But the gain still outweighs the damage greatly. As comparison, the optimal would benefit 90.1% of locations with median increase of 32.4 Mbps (1.7×), without hurting others. To enlarge the gain and mitigate the loss, the the network needs to further differentiate cell groups of large channel-width based on runtime dynamic factors like cell load. It implies that more sophisticated decision-making scheme is needed as the next step on top of CA++ (discussed in §10.3).

CHAPTER 7

RPerf: Reconfiguring Cell Selection Towards Better Performance

7.1 Motivation

In this section, we first introduce the essential role of parameter configuration in cell selection. We then use one real-world instance to motivate the reconfiguration problem; Last, we present three drive forces behind reconfiguration for better performance.

7.1.1 Parameter configuration for cell selection

Parameter configuration plays a critical role in giving flexibility to network operators to customize their own policies while strictly following the standard mechanism. As §2.1 introduces, serving cell selection (including PCell and SCell) starts when the current PCell sends cell-switching-related configuration to the client. These parameters are pre-configured to define the criteria to trigger, decide and execute cell selection at runtime. They include whether to measure neighboring cells, what cells to measure (over the same/different frequency channels), whether to report the measurement results and what to report (i.e., events in Table 2.1), how to decide the target serving cell, and so on. At runtime, measurement and reporting are triggered at the device side when the pre-configured conditions are satisfied. Afterwards, the reported measurement results are used by the current PCell to assist its handover decision, and the serving PCell switches if a handover is decided and executed. SCell selection is similar and the difference is that the criteria to use are configured by the

Table 7.1: Main configurable parameters (R_s, R_n could be any form of RSRP/RSRQ for serving cell and neighbor cells, respectively).

Param.	Step (in Figure 2.1)	Criterion	PCell	SCell_{4G}	SCell_{5G}
Θ_{A1}	② Measure	$R_s > \Theta_{A1}$	✓	✓	
Θ_{A2}	② Measure	$R_s < \Theta_{A2}$	✓	✓	✓
Θ_{A3}	③ Feedback	$R_n > R_s + \Theta_{A3}$	✓		✓
$\Theta_{A5,1}, \Theta_{A5,2}$	③ Feedback	$R_s < \Theta_{A5,1}, R_n > \Theta_{A5,2}$	✓		

new serving PCell and the parameter values differ from those for PCell selection, e.g., cell constraints (what cells are allowed as SCells) and reporting event thresholds.

Main configurable parameters used in this work. We use AT&T and T-Mobile to study the reconfiguration problem in this work. Table 7.1 lists major parameters used for cell selection by both carriers, which are confirmed in our measurement study (§7.1.3). All the parameters and their associated criteria are regulated by 3GPP specifications [3GP15, 3GP19h], which define a complete list of configurable options in more complex forms for global operators. Generally, the criteria are based on radio signal strength measurements (in terms of RSRP or RSRQ) of the serving cells and available candidate cells; parameters in event A1 and A2 control which cells to measure (step ②); Parameters in event A3 and A5 decide which cells to report (step ③); They work together to affect the decision of cell selection and the resulting performance.

7.1.2 Example: An 8-fold Speed Increase via Reconfiguration

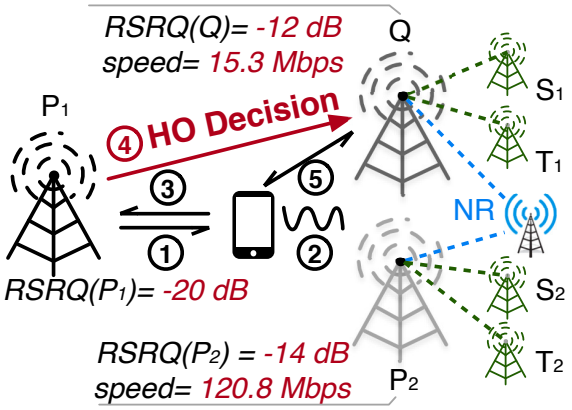
Figure 7.1 illustrates a handover instance which selects worse cells and results in much lower data speed (dropping from 120.8 Mbps to 15.3 Mbps on average). There are 8 cells involved: P_1, P_2, Q, S_1, S_2, T_1, T_2 and NR; Their cell information is given in Figure 7.1b. These cells run over the same (marked with the same letter) or different frequency channels. Each cell is uniquely identified by its short ID and frequency channel number which corresponds to

one specific frequency bandwidth regulated by 3GPP specifications [3GP17a] (for LTE) and [3GP19d] (for NR). Here, 5 channels over 4 bands are used (more observed, see §7.1.3); Band n5 is for 5G NR and exactly reuses band 5 for 4G (originally for 2G and 3G).

In this instance, the serving PCell hands over from P_1 to Q and then adds two SCells (S_1 and T_1); They together offer 15.3 Mbps (on average) to the device. However, this handover misses a much better choice with P_2 as PCell and S_2, T_2 and NR as SCells, which offers 120.8 Mbps (7.9x). It is repeatedly observed at one location ($\star$ of Figure 7.2) in our study.

We next explain why the handover selects Q as the new PCell and fails to offer higher data speed it could afford. This is due to handover configurations in place. Figure 7.1c lists main parameter values used by cell P_1 in the example. There are three criteria. First, there is one event A2. It specifies that inter-frequency cells are measured only when RSRQ of the PCell drops below A2 threshold Θ_{A2}; Otherwise, only intra-frequency cells are measured. Here, RSRQ of P_1 (-20 dB) is lower than Θ_{A2} (-17 dB), and thus both P_2 and Q are being measured by the device. Second, there are two A3 events that specify the criteria for intra-frequency and inter-frequency measurement reporting. It is reported if the measured RSRQ of the candidate cell is offset better than PCell by Θ_{A3} at least. As a result, only cell Q is reported because its RSRQ is greater than $RSRQ(P_1)$ by 8 dB, which satisfied the criterion. On the other side, P_2 is not reported because the difference in RSRQ (-20dB vs. -14 dB) is smaller than the offset $\Theta_{A3}^{s,inter}$ (7 dB). Finally, only cell Q is visible to the network and gets selected eventually. In practice, several A3 events may be configured, each associated with one or multiple frequency channels of candidate cells. Last, cell P_2 (over band 2) accepts NR cells as SCells, but cell Q (over band 12) does not. It is consistently observed in our measurement study and such cell constraints are likely set by AT&T to manage her spectrum resources. At hence, the handover misses not only P_2 as a PCell but NR as a SCell (the rest two SCells running over the same frequency channels in both handover choices).

The chance of selecting better cells is eliminated by current configurations. However, current configurations are not without rational. Signal strength-centric decision has been



(a) Handover to worse cells

Cell	ID	Freq #	Band
P_1	427	850	2 (LTE)
P_2	417	850	2 (LTE)
Q	455	5110	12 (LTE)
S_1	198	66461	66 (LTE)
S_2	52	66461	66 (LTE)
T_1	214	66986	66 (LTE)
T_2	449	66986	66 (LTE)
NR	371	174400	n5 (NR)

(b) Cell information

Parameter	Before (original)	After (reconfiguration)
Θ_{A2}	-17 dB	-17 dB (–)
$\Theta_{A3}^{s,inter}$	6 dB	9 dB (↑)
$\Theta_{A3}^{s,intra}$	7 dB	5 dB (↓)

(c) Parameter configuration and reconfiguration at cell P_1

Figure 7.1: An example of an 8-fold speed increase via reconfiguration (15.3 Mbps → 120.8 Mbps, AT&T, at ⋆ of Figure 7.2)

working well for decades to provide connectivity. But it is now insufficient to reach good service, especially at locations with dense cell deployment. On the one side, those configurations do not guarantee the strongest cell is selected. Because they only target connectivity by finding cells strong enough (i.e., above some threshold). Even though the selected cell has the highest signal strength, it does not indicate good performance compared to unselected cells. 5G NR and CA further enlarges the gap because it brings more options in terms of cell combinations which may have huge variance in capability.

We further illustrate how reconfiguration prevents such performance loss at the first place. One straightforward strategy is to include cell P_2 as a new candidate and get rid of cell Q.

Figure 7.1c gives one reconfiguration option, where Θ_{A3}^{intra} decreases from 7 dB to 5 dB, Θ_{A3}^{inter} increases from 6 dB to 9 dB and other parameter values remain the same. Therefore, cell P_2 becomes visible and finally wins when P_1 is preparing for a handover. Moreover, it is sufficient to get rid of cell Q as long as the reconfigured Θ_{A3}^{intra} is no smaller than 8 dB; A cell over the same frequency channel (an intra-freq handover) is preferred. There are more than one effective reconfiguration options. Reconfiguration at scale is not easy as illustrated in this instance; A number of factors must be considered and we will elaborate technical challenges in §7.2.1 and §7.2.2.

7.1.3 Three Drive Forces for Reconfiguration

We advocate reconfiguration not only for its potential performance gains, but also for its practicability and compatibility with network operations in place. It is driven by three forces.

First, reconfiguration is not new. Network operators do (often ask vendors to) configure tunable parameters to customize their operation policies while deploying and upgrading their network infrastructure. They reconfigure some or all of the parameters over time, particularly with major upgrades such as deploying a new technology (e.g., adding CA in 2016 and adding 5G NR in late 2019), acquiring new spectrum bands (e.g., adding band 12 for 4G in 2017 and band n260 for 5G in 2020), repurposing old bands (e.g., retiring band 5 for 4G and reusing it as band n5 for 5G since Nov 2019 [Ope20]), or performing regional or national updates. This indicates that new reconfiguration strategies and algorithms for enhanced performance does not require any major physical infrastructure upgrade; They are ready to launch by leveraging the off-the-shelf interfaces and tools for reconfiguration.

Second, reconfiguration to enhance data speed is largely missing despite feasibility. The above instance is not rare. It is commonly observed in our reality check in Los Angeles (C1), one of the largest cities in the US, where AT&T have full 4G coverage and early 5G rollout. Our results are consistent with recent measurement studies in a small college town [DLH20, DLG20].

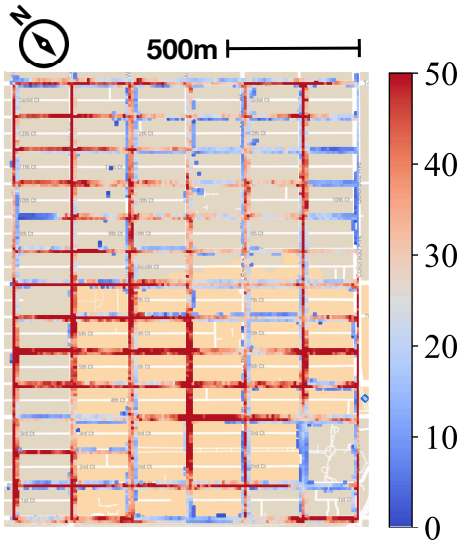


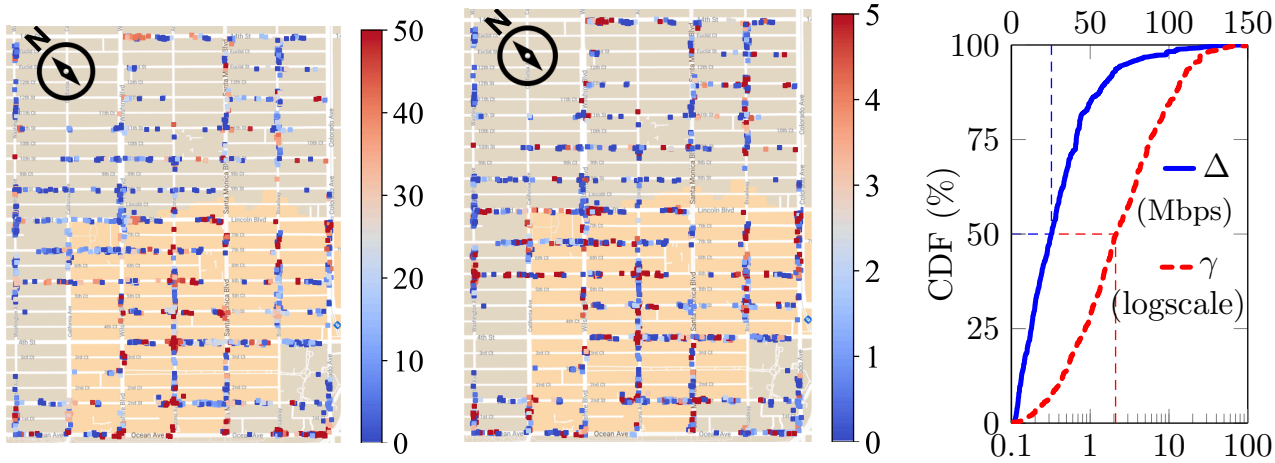
Figure 7.2: Map of cell density.

Area size	2.1 km^2 (1.3 km × 1.6 km)
Road length	28.0 km
Driving distance	551 km
Duration	32.9 h
RF bands (#: 9)	2,4,12,14,29,30,46,66,n5 (NR)
#. LTE RF ch.	14 (PCell: 5, PCell+SCell: 14)
#. LTE cells	1,504 (P: 113, P+S: 237)
#. NR RF ch.	1 (174400@band n5)
#. NR cells	21
#. location grids	809 (≥10 cells: 82.0%)
#. handovers	2,837
#. RSS meas.	5,232,516
Data speed (Mbps)	0.001 – 284.5 (med: 12.3)

Table 7.2: Dataset C1-A (P=PCell).

Methodology and dataset. We follow the methodology in [DLH20] but use two new phone models including Pixel 4a and Pixel 5 which support 5G in AT&T. This measurement study is mainly conducted in a 1.3 km × 1.6 km commercial region in April - May 2021. To reduce possible biases with selected locations and routes, we run driving experiments along all accessible roads to fully cover the test region. We sample out of the region because some driving routes across the surrounding areas impact the initial cell set. In each experiment, we run an elephant flow (speedtest via file downloading) or a mice flow (ping every second) at our test phones. The former is to collect cell selection instances and data speed samples while both are used to measure radio quality and cell deployment. Table 7.2 gives basic information of our dataset C1-A. Figure 7.2 shows the map, along with cell density observed. There are abundant candidate cells at any location (>40 at hotspots). This is thanks to continuous and heavy investment on network infrastructure upgrades (e.g., acquiring more bands, deploying denser cells and rolling out 5G).

Reality check. The purpose is to check whether default cell selection (configurations) could lead to high data speed as it can. We evaluate the performance of the selected serving



(a) Δ = best - default (b) $\gamma = \Delta$/default (c) CDF

Figure 7.3: Distribution of missed data speeds in our study.

cells, based on the comparison with the bound of affordable performance at the same location. The best serving cell is learned from its performance profile. We ran extensive experiments to collect sufficient performance data in the selected region and build profiles for each cell set. At a location, the best cell set has the highest median data speed among all available ones. We collect all cell selection instances and analyze whether they have chosen the best serving cells. We define a serving cell set is α-*optimal* if the ratio between its median speed and the median of the best cell set is no less than α ($0 \leq \alpha \leq 1$). The selection of a sub-optimal cell set implies that the existing configurations are improper for not preferring or even ignoring the better candidates. In the whole region, we observe that only 28.3% of handovers lead to 90%-optimal cells. Figure 7.3 shows the speed gaps of non-optimal selections. Note that cell selections do not happen everywhere and we only show the gaps at locations of cell selections. We use the absolute and relative gaps between the median data speed by the best and selected serving cell sets. At more than 50% of instances, the data speed gap is larger than 25.4 Mbps or 214%. In the worst case, the gap goes up to 148 Mbps; It is likely larger as 5G grows. This implies that current parameters are not well tuned towards higher data speed.

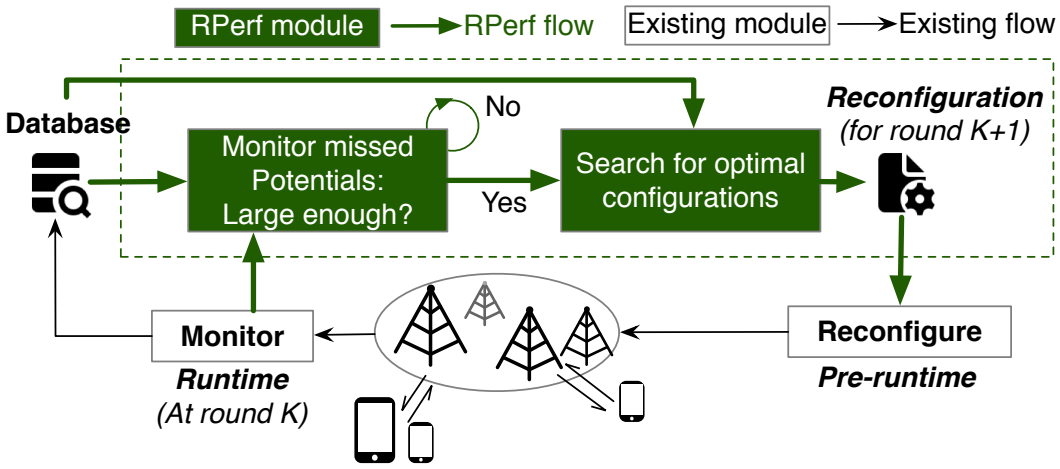


Figure 7.4: Overview of the RPerf design.

Last, reconfiguration is not a remedy, but a prevention. [DLH20] has devised a device-assisted solution to detect and correct improper cell selection at runtime to boost data speed. Despite effectiveness, the solution is a remedy. It seeks to reduce or recover from performance loss after such loss has already occurred. It requires heavy profiling and training in advance and raises runtime monitoring and learning which can not complete right away (takes at least several seconds). Naturally, it fails to help when traffic flows are short or traffic patterns vary over time. A more desirable solution is to prevent under-utilization instead of mitigation after it happens. Just by tuning some parameters, we can change the result of cell selection towards the target with better performance. More importantly, it is aligned with the needs of both users and operators. The trending technology is to make 5G more intelligent and maximize the efficiency of network resources. This calls for reconfiguration beyond connectivity by taking user performance into account. Reconfiguration should be proactive, not passive. It should timely and intelligently monitor performance of cell selection and tune parameters to reduce the likelihood of poor ones, rather than take actions upon bulk user complaints.

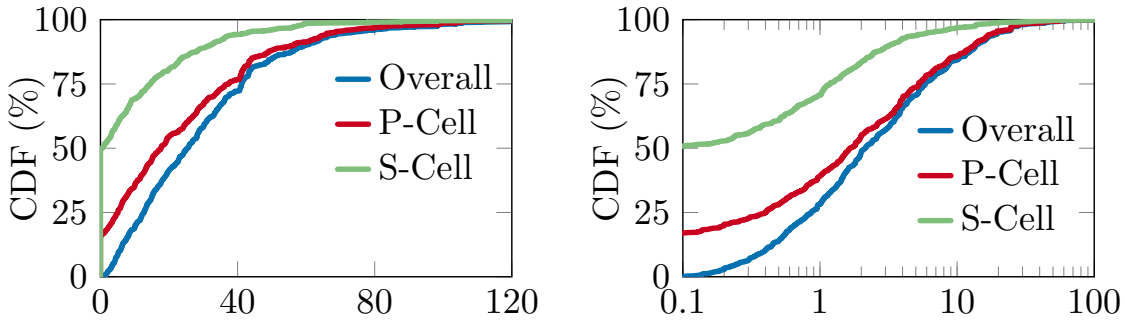
7.2 The RPerf Design

We propose RPerf to reconfigure parameters used for cell selection to enhance data performance afterwards. The overall design is depicted in Figure 7.4. RPerf is built on top of two existing network functions: monitoring at runtime and reconfiguration at pre-runtime; It adds two modules to trigger and execute performance-driven reconfiguration. Reconfiguration for the next round is triggered when the potentials of better performance missed by configuration at the current round has become large enough (see the triggering condition in §7.2.4); It is then executed by efficiently searching parameters that achieve better performance in all the impacted cell selection instances (§7.2.3). Before we dig into RPerf, we first present its technical challenges (§7.2.1) and design heuristics (§7.2.2).

7.2.1 Reconfiguration is not Easy

Intuitively, RPerf is to change the result of cell selection towards the target with better performance, just by tuning some parameters. However, it is not easy as illustrated before.

First, the mechanism of cell selection is based on radio signal strength, not designed for performance. To be compatible with minimal changes to the existing network infrastructure and operations, RPerf cannot directly change the outcome of cell selection but tune radio-centric criteria to *implicitly* impact the outcome. The alternative solution of explicitly changing the cells to select is discussed in §7.4. Specifically, RPerf must tune these threshold parameters to change the criteria so as to get rid of bad candidate and reach optimal (or close) cells. To be qualified for being considered as a target cell (step ④), a neighboring cell must be first measured by the device (step ②) and has its radio signal strength higher than the reporting threshold (step ③). To tell good from bad, we need to build up profiles of performance and signal strength based on historical measurements. Note that performance of any unselected cell at a specific time never exists and thus is unobservable. Facilitated with the knowledge of good/poor candidates, the serving cell is able to figure out



(a) Δ (Mbps) (b) γ

Figure 7.5: CDF of the speed gaps caused by PCell/SCell selection.

the reconfiguration towards enhanced performance and better resource utilization.

Second, reconfiguration is not optimized for individuals, but statistically for all the instances. It can not be tailored for every handover instance. Instead, it is applied to all the impacted handovers within an area. As a result, reconfiguration probably helps some cases while hurting some others. For example, reconfiguring A3 thresholds indeed increases the chance of selecting P_2, not Q, in the above instance. But it may also degrade performance in case Q (along with its SCells) performs better than P_2 at a different location but is not selected due to reconfiguration. Therefore, the goal of reconfiguration is all about improving the *overall* performance. Whether to trigger reconfiguration, depends on whether performance gains in all the cell selection instances outweigh losses, if the losses occur. We next show the "net" gain in a what-if study (§7.2.2), where gains in more cases outweigh losses in fewer cases.

Last, reconfiguration has a huge, high-dimensional space. In principle, it must tune all the parameters of all the cells *together* due to three coupling effects. (1) Parameters used by one single cell must work together to determine the steps of cell selection leading to the final target and thus cannot be tuned independently. Parameter values may be associated with different radio frequency channels (e.g., Θ_{A3}); the configuration space per cell expands with the growing frequency channels to use. For each cell, there are tens or even hundreds

parameters to tune. It also matches with a previous global-scale handover configuration measurement study [DPF18a]. (2) Parameters for all the cells involved in a cell selection must be tuned at the same time. Starting from the original serving cell, the device might go through one or multiple handovers until reaching the final target (without further switch). Note the target cell of a single handover may just be a transient state. Once the serving cell switches, the configured parameters and the associated criteria change accordingly. The outcome of cell selection depends on the parameters of all the involved cells. (3) All the parameters of all the cells impact each other with coverage locality. Parameters tuned to optimize performance at some locations may hurt data performance of cell selection at other locations. It is challenging to efficiently search for new parameter values across the high-dimensional space. Therefore, RPerf uses heuristics to reduce complexity of reconfiguration (§7.2.2).

7.2.2 Heuristics for RPerf

We find that reconfiguration can be simplified with several heuristics learned from real-world traces.

First, PCell selection takes the major blame of missed performance. We examine the speed gaps caused by the selection of PCell or SCell(s) in our reality check. Except 28.3% handover instances that achieve *90%-optimal*, 51.6% instances are caused by an improper PCell and 20.1% instances can be fixed by using different SCell(s). We further examine the distribution of speed gaps contributed by PCell/SCell selection in Figure 7.5. The gaps caused by PCell selection goes very close to the overall one. It is aligned with two more observations: The largest speed gaps exist between the serving cell sets with different PCells. With P-Cell fixed, the missing performance by SCell selection goes much smaller. Therefore, we should prioritize reconfiguring PCell selection.

Second, we find that not all the configuration parameters are equally important to cell selection. In fact, a subset of tunable parameters play a decisive role. Unfortunately, such

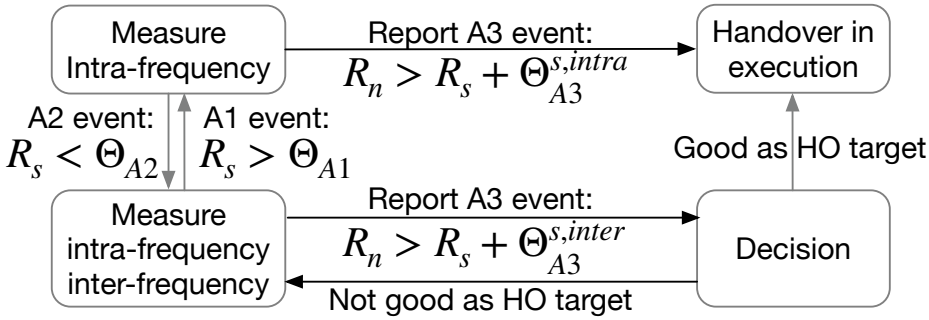


Figure 7.6: Model of the handover process learned from real traces.

information is decided by the operator and not released to public. We follow the approach in [LPY16] to infer the handover model. This is based on the mechanism defined by 3GPP standards [3GP15, 3GP19h] and inference from real traces. We use MobileInsight [LPY16] to collect 5G/4G signaling messages exchanged for each handover instance and learn a model of the handover process used by AT&T (Figure 7.6). The model is represented by a state machine, demonstrating which configuration parameters to use for cell selection and how. It is highly accurate and predicts the outcomes of cell selection at precision of 99.0%. Therefore, when trying different parameter values without actual deployment, this model is able to predict new targets with high confidence.

In general, we find that the following parameters are used by AT&T to influence cell selection in our test area (city): (1) Θ_{A2} and Θ_{A1} set the criterion to enable (disable) measurement on inter-frequency channels. There is no inter-freq measurement until the serving cell's signal strength runs below Θ_{A2}, while intra-freq measurement is always on. In our measurement study, we observe that Θ_{A1} always equals to Θ_{A2} as it functions in the opposite way. (2) Θ_{A3} specifies the condition to trigger reporting for the measured cells: only when the neighboring cell's signal strength is stronger than the serving one by Θ_{A3} at least. Θ_{A3} has two parameter values for the intra-freq and inter-freq cells. We notice that some handovers ($< 20\%$) are caused by unpredictable A5 reporting. We leave them because we have no access to all information available to the network operator. It can be resolved when reconfiguration is performed by the operator with all the information available on the net-

work side (discussed in §7.4). We focus on proof-of-concept reconfiguration in RPerf. As a result, our study focus on three parameters per cell: A2 threshold (Θ_{A2}), A3 offsets for intra-frequency (Θ_{A3}^{intra}) and inter-frequency neighbors (Θ_{A3}^{inter}). In our dataset, all parameters are based on RSRQ.

Last but not least, we find that it is promising to reduce the configuration dimensions as many factors can be decoupled without impacting the performance. We conduct a what-if study to examine the need and feasibility of performance-driven reconfiguration. We enumerate all possible values for those critical configurations to estimate the highest possible reward for all handover instances. We simplify the whole-space search with two tricks. First, parameters are tuned within a rational range (not too far away from actual values observed in our dataset), namely, $\theta_{A2} \in [-17, -8]$ dB, $\theta_{A3,intra}, \theta_{A3,inter} \in [0, 10]$ dB. This is reasonable to make reconfiguration practical at both the network and device sides. Second, we reconfigure each frequency channel separately, not per cell. Carriers would like to simplify configuration with area-specific policies instead of cell-specific ones. As our test area is small enough, we consider distinct configurations per frequency channel to use. Our study further shows that such simplifications are reasonable and the impacts are negligible.

Our what-if study is performed with three steps:

1. For each parameter setting, predict the new target of each handover instance based on the model (Figure 7.6) and profiles.

2. Estimate the overall reward considering gains and losses over all the cases. In this step, we define the reward as the possibility of gains minus the possibility of loss:

$$R = \frac{n_{\text{gain}}}{n_{\text{total}}} - \frac{n_{\text{loss}}}{n_{\text{total}}},$$

   where n_{gain} and n_{loss} refer to the number of cell selection instances with gains and losses after reconfiguration and n_{total} is the total number of all instances.

3. Repeat step 1 & 2 to enumerate all the possible settings and select the values associated

with the highest reward.

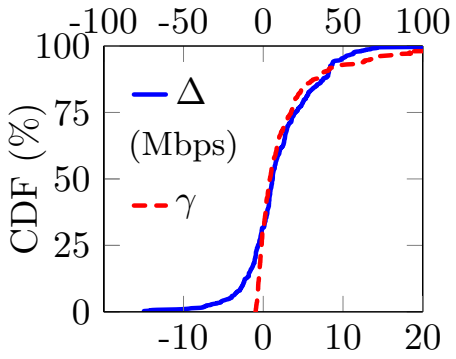
We note that the network operator could assess the benefit with different reward functions, considering the aggregated gains and losses together. For example, a conservative operator might first minimize the number of cases with performance losses, and then maximize the gains atop of that. In the what-if study, we have to use the profile extracted from our dataset, because the device cannot measure performance of multiple cells at the same time.

Reconfiguration helps more. Reconfiguration could bring promising benefits as expected. We notice that it is double-sworded but the gain outweighs the loss. Figure 7.7a shows the distribution of gains and losses in all the impacted handover instances. Performance are enhanced in 30.0% of all cell selection instances. Meanwhile, performance is degraded in 13.9 % cases. Considering all gains and losses, the median increase is by 83.9% or 5.9 Mbps as absolute change. There are no changes in the rest 54.3% of instances because reconfiguration would not impact the outcome of every cell selection.

On each individual channel, the percentage of cases w/ performance gain also dominates those w/ loss. In our study, there are five frequency channels used for PCells. Table 7.7b demonstrates the optimal reward for all channels, after filtering one channel with insufficient samples (here, 5330). On all the four observed channels, the percentage of improved cases exceeds the hurt cases by at least 12.3%. The median speed increase is at least 32.5% on all channels, and reached up to 327.4%. Note that we could obtain all such benefits by simply tuning some parameters within reasonable ranges. These results indicate great improvement to be achieved with proper reconfiguration per frequency channel.

The searching space reduces. Most importantly, we find that it is feasible to decouple those inter-dependent parameters by applying restrictions to reconfiguration. Instead of tuning all parameters together, we can reduce the search space with three key observations.

1) The possibility of immediate switches after one handover can be largely reduced as long as the signal strength is stable. Without impacting the final handover decision (impacting



(a) Dist. of gain and loss.

Channel	$\frac{n_{\text{gain}}}{n_{\text{total}}} - \frac{n_{\text{loss}}}{n_{\text{total}}}$	Median γ
850	12.3 (28.1 − 15.8)	57.0%
5110	20.6 (30.3 − 9.7)	133.7%
9820	15.7 (36.8 − 21.1)	32.5%
66461	56.9 (60.8 − 3.9)	324.7%
All	16.1 (30.0− 13.9)	83.9%

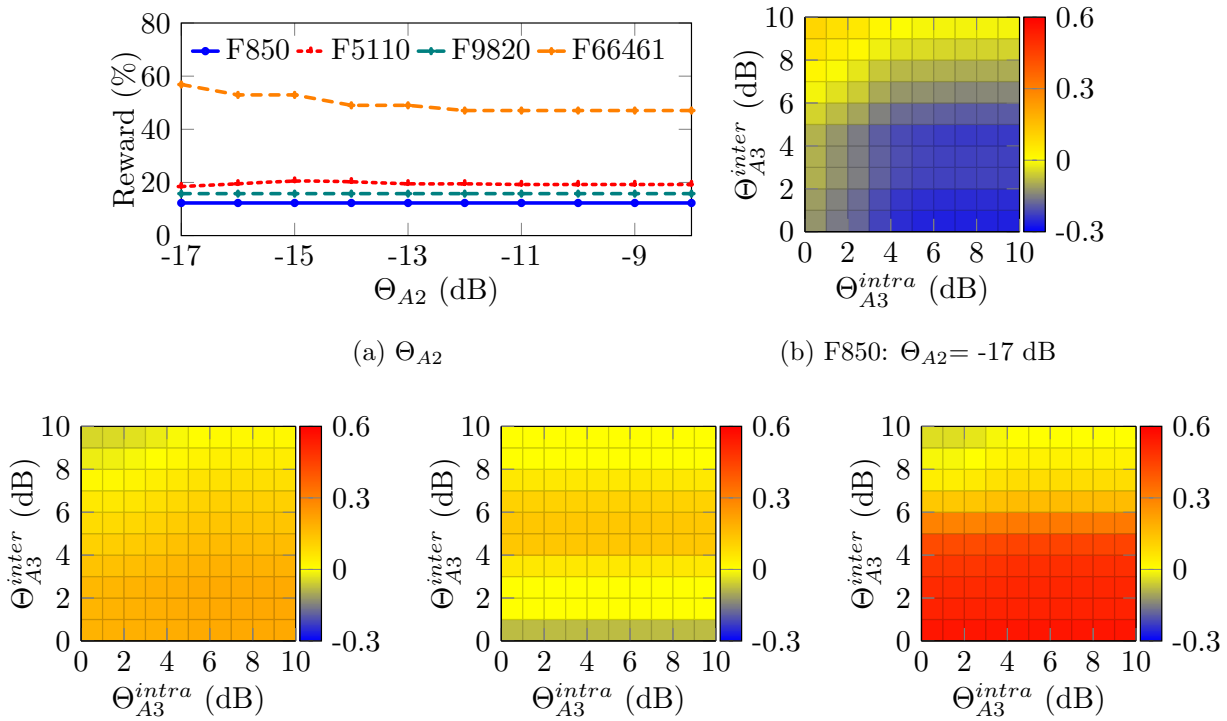
(b) Overall reward of optimal reconfig.

Figure 7.7: Performance gain/loss of "optimal" reconfigurations.

the reward), we can adopt the following tactics: (1) consider only non-negative values for A3 offsets, and (2) use Θ_{A2} as the lower bound of the candidate's signal strength when inter-freq handover is considered. Note that A2 and A3 are used to tune the measurement and reporting steps. All rules above will greatly reduce the possibility of continuous handovers at the same location, rather than eliminate it.

2) Changing A2 threshold would only cause marginal difference to the reward. Figure 7.8a demonstrates the trend of the maximal rewards by tuning Θ_{A2}. Across the value range, the absolute difference between the maximum and minimum rewards is no more than 10%. In addition, the maximum reward falls into [-17,-15] dB on all the tested channels. It justifies that the gain of reconfiguration is not compromised when A2 threshold is restricted to a small range for complexity reduction. In the following reconfiguration search, we should prioritize a small range of [-17, -15] dB.

3) A3 offsets play a more critical role in impacting the reconfiguration reward. These configurations directly determine the qualification as target cells. We examine the reward of all combinations of Θ_{A3}^{intra} and Θ_{A3}^{inter} from 0 to 10 dB. The range covers the dominant values used by AT&T: $\Theta_{A3}^{intra} = 3$ dB and $\Theta_{A3}^{inter} = 5$ dB for all channels. Figure 7.8 depicts how the reward changes with regards of A3 offsets, given a fixed A2 threshold. All channels have strong correlation with inter-freq configuration. Comparatively, intra-freq A3 offset has less



(a) Θ_{A2}

(b) F850: Θ_{A2}= -17 dB

(c) F5110: Θ_{A2}= -15 dB

(d) F9820: Θ_{A2}= -17 dB

(e) F66461: Θ_{A2}=-17 dB

Figure 7.8: The impact of Θ_{A2}, Θ_{A3}^{intra} and Θ_{A3}^{inter} in our what-if study.

impact on the reward. Take channel 66461 as an example (Figure 7.8e). With parameter Θ_{A3}^{inter} fixed, tuning Θ_{A3}^{intra} only changes reward slightly (absolute change less than 4%). In the opposite way, tuning Θ_{A3}^{inter} covers the reward from -3.9% to 56.9%. Another important observation is that, for all value of Θ_{A3}^{intra}, the best choice of Θ_{A3}^{inter} is nearly constant.

7.2.3 Fast search

We now incorporate these heuristics to RPerf. The core is to efficiently search for new values of configurations. The brute-force approach is unrealistic given high-dimensional configuration space and tremendous handover instances. In RPerf, we design fast search by prioritizing sub-space search and then using linear search for acceleration (Algorithm 1).

Subspace prioritization. To reduce the searching complexity, our first take is to de-

Algorithm 4 Fast search for optimal configurations

function FastSearch
  Set $\Theta_{A2}, \Theta_{A3}^{intra}, \Theta_{A3}^{inter}$ to initial values
  $r_{max} = -Inf$
  for Θ'_{A2} in [-17,-15] dB **do**
    $\Theta_{A3}^{intra'}, \Theta_{A3}^{inter'}, r' \leftarrow$ MaxReward(Θ'_{A2})
    if $r' > r_{max}$ **then**
      $r_{max} \leftarrow r', \Theta_{A2} \leftarrow \Theta'_{A2}$
      $\Theta_{A3}^{intra} \leftarrow \Theta_{A3}^{intra'}, \Theta_{A3}^{inter} \leftarrow \Theta_{A3}^{inter'}$
    end if
  end for
  return $\Theta_{A2}, \Theta_{A3}^{intra}, \Theta_{A3}^{inter}$
end function
function MaxReward(Θ_{A2})
  Set $\Theta_{A3}^{intra}, \Theta_{A3}^{inter}$ to initial values
  $\Theta_{A3}^{inter} \leftarrow \arg\max \mathrm{R}(\Theta_{A2}, \Theta_{A3}^{intra}, \Theta_{A3}^{inter})$
  $\Theta_{A3}^{intra} \leftarrow \arg\max \mathrm{R}(\Theta_{A2}, \Theta_{A3}^{intra}, \Theta_{A3}^{inter})$
  return $\Theta_{A3}^{intra}, \Theta_{A3}^{inter}, \mathrm{R}(\Theta_{A2}, \Theta_{A3}^{intra}, \Theta_{A3}^{inter})$
end function

couple those inter-dependent parameters by applying restrictions to reconfigurations. In particular, we focus on A2 thresholds in a range of $\Theta_{A2} \in$ [-17, -15] dB and only positive values for A3 offsets. Meanwhile, we also prioritize candidates with RSRQ $\geq \Theta_{A2}$ in the stage of final decision. Our study shows that 93.5% of original handovers would not precede another handover at the same location, given the above subspace. This is good enough to enable distributed reconfiguration on each channel, which makes the design scalable. Note that our design cannot make 100% handovers stable; Otherwise more stringent condition is expected to select the target, which may impede on-time handover and even hurt disconnectivity.

Linear search. We then iteratively search for the optimal parameters in the space to

explore. Given distinct impacts of configuration parameters (Figure 7.8), we take two strategies: (1) enumerate all values of A2 threshold, given a small range, and (2) tune Θ_{A3}^{inter} before Θ_{A3}^{intra} , as the reward is largely decided by the former. It helps us to benchmark the highest reward "level". Then, tuning Θ_{A3}^{intra} further optimizes the reward on that level. The time complexity is $O(NK(|\mathbf{\Theta}_{A3}^{inter}| + |\mathbf{\Theta}_{A3}^{intra}|))$, in which N is the total number of handover instances, K is the number of frequency channels and $|\Theta_{A3}|$ is the size of the range.

We argue that such heuristics-based fast search may sacrifice the reward optimality but it is acceptable and practical. First, the extra reward from the sub-space search to the whole space search is marginal. This is likely because network operators do not reconfigure parameters for performance and thus the reward is significant with such reconfiguration. At hence, there is no much need to push to the limit once the potential of reconfiguration is almost fulfilled. Second, current parameter values are not set randomly. They came from many-year experience and professional field trials. The engineers and technicians do radio planning and (re)configure these parameters for radio connectivity. Abundant cell deployment and increasing capabilities result in good radio $\neq$ good performance, which opens room for performance-driven reconfiguration. However, good values must comply with good radio coverage; It is often harmless to narrow down reconfiguration to a small subspace (validated by empirical studies).

7.2.4 Triggering Reconfiguration

Reconfiguration is triggered based on the possible performance reward of all handover instances. Generally, the network evaluates it using statistical measures periodically (e.g., every day or two) to avoid frequent changes. Which measure to monitor is up to the operator's decision. For example, our design uses the ratio of handover instances whose targets are *not* 90%-optimal. Reconfiguration is invoked when the ratio goes above 30%. The operator can define measures out of their needs, as long as they are consistently used for reconfiguration triggering and optimization. Given selected measures, a triggering condition is then

created to indicate when the gap goes beyond tolerance.

7.3 Evaluation

RPerf is evaluated by trace-driven emulation. Unfortunately, we cannot test RPerf on real systems or large-scale testbeds, since we do not have internal access to change configurations. In order to emulate results in practice, we use real data from collected traces to approximate the actual network conditions, including handover instances, cell signal strength and performance.

In addition to the previous C1-A dataset (Los Angeles, AT&T), we use two more datasets for evaluation: C1-T and C2-A. C1-T is collected at the same region as C1-A to test RPerf with T-Mobile. We ran similar driving experiments over T-Mobile for 17.8 hours and 305 km in Aug 2021. C2-A is a public dataset over AT&T in West Lafayette, IN (C2) [DLH20]; It contains performance and radio information in a region of 2.5 km^2 with 876 grids. Note that C2-A does not include 5G measurement as a result of no coverage. Then, we run RPerf on three datasets and evaluate the overall improvement (§7.3.1) and efficiency (§7.3.2). We also compare the results on different datasets and show insights on 5G (§7.3.3).

7.3.1 Overall Improvement

AT&T. We evaluate RPerf by analyzing the performance gain and loss. In C2-A, we see that AT&T use the same parameters as C1-A: $\Theta_{A2}, \Theta_{A3}^{intra}, \Theta_{A3}^{inter}$, which are the decisive factors to cell selection. We make three observations.

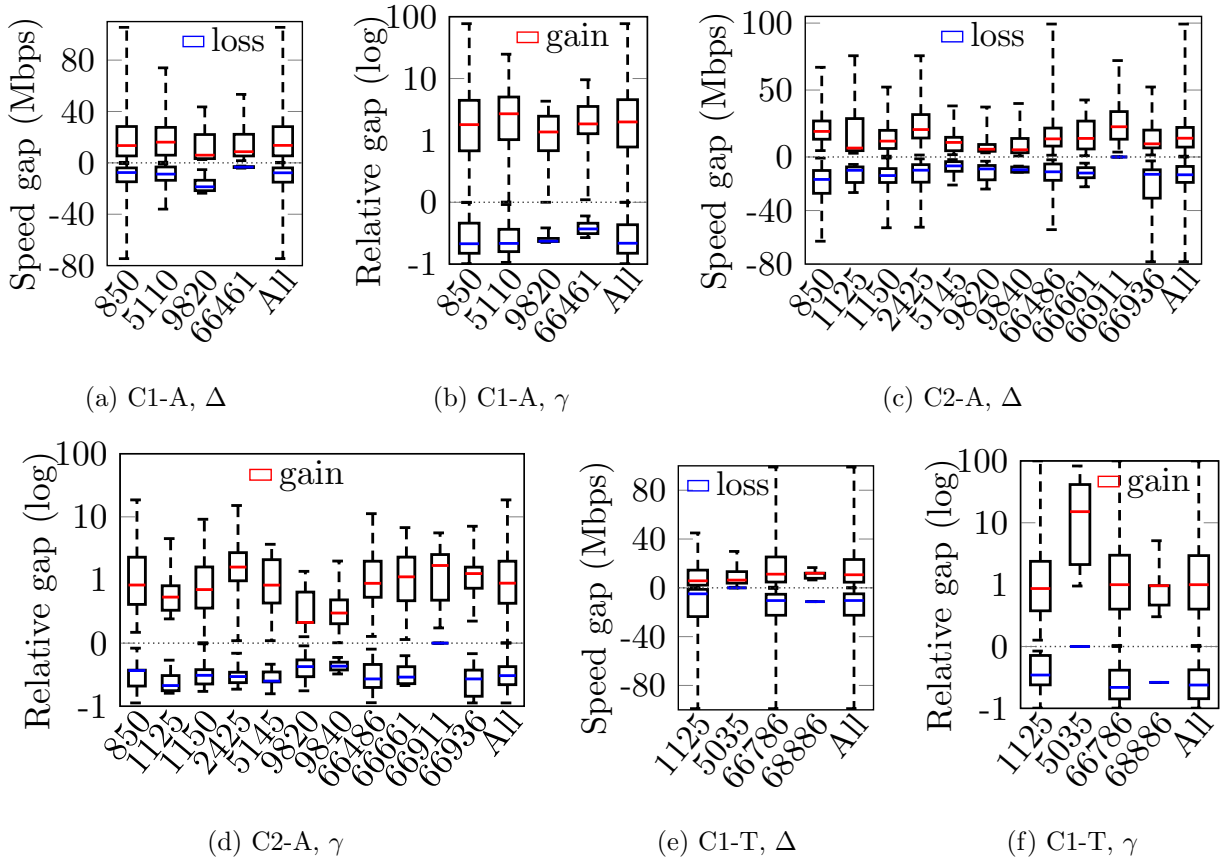
First, RPerf would greatly enhance the performance by improving a large proportion of users. We use the metric R defined in §7.2.2, the "net" gain which takes both improved cases and worsened cases into account. Larger values indicates that the number of improved cases is much more than the number of worsened ones. As shown in Table 7.3, the "net" gain is 16.0% in C1-A (30% of cases with gains v.s. 14.0% of cases with losses); It is higher

in C2-A, which reaches 27.8% (42.6% v.s. 14.8%). We examine the results per frequency channel and see the percentage of better instances is always higher than the worse one in both datasets.

Second, considering all cell selection instances impacted by reconfiguration, the majority still get a big surge in performance. We use the median of absolute speed increase (i.e., Δ in Mbps) and relative increase (i.e., γ in %) in performance over all instances with performance change (not just limited to gain). Larger numbers indicate higher increase overall. In C1-A, the median speed increase is 5.9 Mbps, or 81.3% as relative value (Table 7.3). In C2-A, the speed increases by 9.6 Mbps or 56.2%. Therefore, despite worsened cases, RPerf still benefits users with a decent overall increase.

Last, the gain outperforms the loss in terms of the increased data speed. Figure 7.9 shows the absolute difference (Δ) and the relative difference (γ) in three datasets. In AT&T, the median speed grows by 13.6 Mbps (200.0%) in those improved instances, while the median drop is 7.7 Mbps (45.9%) in those worsened cases in dataset C1-A; We see that the gain declines a little bit in C2-A: the median gain is 14.0 Mbps (89.1%) while the median drop is 13.3 Mbps (32.7%) in those worse instances. This is because the gains over some frequency channels (1125, 9820, 9840) are fewer than those at other channels and these channels are observed in C2-A; According to the median speeds, the absolute gain outperforms the loss on most frequency channels: 3 out of 4 in C1-A and 6 out of 11 in C2-A.

T-Mobile. RPerf is applicable to other carriers. We first find that the handover model used by T-Mobile is almost the same as the one by AT&T, except for A5 event used to select inter-frequency cells. This model imitates the network's operations with high confidence, given the prediction accuracy of 98.7%. Accordingly, there are three tunable parameters critical to cell selection: Θ_{A2}, Θ_{A3}^{intra} and $\Theta_{A5,2}^{inter}$. The first two are used to monitor serving cell and intra-freq candidates, which are the same as AT&T; $\Theta_{A5,2}^{inter}$ is used to search inter-freq candidates, and $\Theta_{A5,1}^{inter}$ is omitted because it always overlaps with Θ_{A2}. T-Mobile uses RSRP, rather than RSRQ by AT&T. Table 7.3 and Figure 7.9 show that RPerf works well



(a) C1-A, Δ (b) C1-A, γ (c) C2-A, Δ

(d) C2-A, γ (e) C1-T, Δ (f) C1-T, γ

Figure 7.9: Performance gain/loss with REM.

over T-Mobile. 41.6% of cell selection instances are improved with median gain of 10.7 Mbps (99.8%), while 24.6% of instances have median performance drop of 10.4 Mbps (by 41.9%). Compared to C1-A, RPerf benefits more instances while hurts more instances at the same time, which leads to the similar "net" gain of 17.0%.

7.3.2 Efficiency

Next, we evaluate the efficiency of RPerf. As mentioned in §7.2, RPerf utilizes several heuristics and adopts fast search to pursue considerate performance gain while minimizing the complexity of reconfiguration. To assess the efficiency of RPerf, we introduce reconfiguration via brute-force search as the baseline. Generally, brute-force search enumer-

ates all combinations of parameters values and ends up with the optimal reconfiguration on the searching space. We compare RPerf with the baseline, by checking the difference in performance improvement and the execution time needed to figure out corresponding reconfiguration.

RPerf turns out to approach the optimal performance while reducing the cost by a factor 3.6 to 4.8. We compare the "net" increase in Table 7.3, which is the goal for RPerf and the brute-force approach to optimize. The absolute value of gap between the reward achieved by RPerf and the optimal reconfiguration is negligible: 0.1% in C1-A, 0.6% in C2-A, and even no loss for C1-T at all. Moreover, RPerf could achieve the optimal reward on 3 out of 4 channels in C1-A and 7 out of 11 channels in C2-A. On other channels, the amount of unachieved reward is within 0.1% and 5.6%, respectively. In the meanwhile, RPerf would speed up the reconfiguration by a factor of 3.6 and 4.8. The overhead gets reduced greatly, almost without comprising the performance gain.

7.3.3 Comparing Results on Difference Datasets

While RPerf has achieved considerate performance improvement for both datasets, we still observe some differences and obtain insights from the comparison.

Reconfiguration towards 5G. As we know, 5G is only deployed in C1 but not C2. To be more specific, AT&T 5G cell are used as SCells only if the PCell is from frequency channel 850. Such restriction is probably enforced by radio resource planning. Therefore, in order to increase the usage of 5G service, the operator should tune parameters in favor of cells on channel 850. Parameter values chosen by RPerf have already implied a similar intention. On channel 850, RPerf sets $\Theta_{A3}^{intra} = 0$ dB and $\Theta_{A3}^{inter} = 10$ dB. Such new parameters indicate high preference of intra-frea handovers over inter-freq handovers. It would help secure the user on channel 850, instead of migrating to a different channel. As a result, RPerf has successfully avoided performance loss on 28.1% cases with only 15.8% cases getting worse.

Table 7.3: Gain and loss after applying reconfiguration. (Configurations: AT&T - RSRQ, T-Mobile - RSRP.)

Dataset	Freq.	Fast search							Optimal [1]
		$(\frac{n_{\text{gain}}}{n_{\text{total}}} - \frac{n_{\text{loss}}}{n_{\text{total}}})\%$	median γ (%)	median δ (Mbps)	cost↓	Θ_{A2}	Θ_{A3}^{intra}	Θ_{A3}^{inter} (A) / $\Theta_{A5,2}^{inter}$ (T)	$(\frac{n_{\text{gain}}}{n_{\text{total}}} - \frac{n_{\text{loss}}}{n_{\text{total}}})\%$
C1-A	850	12.3 (28.1 - 15.8)	57.0	4.5	3.6×	[-17,-15]	0	10	12.3 (28.1 - 15.8)
	5110	20.5 ↓ (30.5 - 10.0)	131.7	8.3	3.3×	-15	10	2	20.6 (30.3 - 9.7)
	9820	15.7 (36.8 - 21.1)	32.5	2.8	6.2×	[-17,-15]	10	5	15.7 (36.8 - 21.1)
	66461	56.9 (60.8 - 3.9)	184.7	8.3	3.7×	-17	3	0	56.9 (60.8 - 3.9)
	Overall	16.0 ↓ (30.0 - 14.0)	81.3	5.9	3.6×	N/A	N/A	N/A	16.1 (30.0 - 13.9)
C2-A	850	0.7 (12.2−11.5)	19.9	5.8	3.9×	[-17,-15]	2	10	0.7 (12.2−11.5)
	1125	8.3 (50.0−41.7)	24.2	2.8	5.6×	[-16,-15]	10	[2,3]	8.3 (50.0−41.7)
	1150	16.2 ↓ (49.4−33.2)	30.1	4.5	3.9×	-15	10	3	21.8 (49.4−27.6)
	2425	37.5 ↓ (51.4−13.9)	136.3	17.8	3.8×	-15	10	3	38.9 (51.4−12.5)
	5145	19.5 (34.1−14.6)	51.3	5.3	4.4×	-15	10	1	19.5 (34.1−14.6)
	9820	1.8 ↓ (37.5−35.7)	12.6	3.1	4.3×	[-16,-15]	10	6	3.6 (39.3−35.7)
	9840	19.1 (38.1−19.0)	17.2	3.0	4.9×	[-16,-15]	10	3	19.1 (38.1−19.0)
	66486	62.9 ↓ (72.6−9.7)	67.7	12.2	3.8×	-15	10	1	63.6 (73.1−9.5)
	66661	57.9 (75.4−17.5)	82.2	9.3	4.0×	-15	10	1	57.9 (75.4−17.5)
	66911	84.6 (84.6−0.0)	169.7	22.6	4.0×	[-16,-15]	10	5	84.6 (84.6−0.0)
	66936	24.4 (51.1−26.7)	69.4	5.9	4.4×	-15	10	1	24.4 (51.1−26.7)
	Overall	27.8 ↓ (42.6−14.8)	56.2	9.6	4.0×	N/A	N/A	N/A	28.4 (42.8−14.4)
C1-T	1125	24.2 (42.5−18.3)	40.8	2.7	5.4×	-101	6	[-117,-114]	24.2 (42.5 - 18.3)
	5035	26.7 (26.7−0)	1511	6.4	5.9×	[-106,-100]	[0,10]	[-104,-100]	26.7 (26.7−0)
	66786	15.9 (41.5 - 25.6)	31.2	3.7	4.7×	[-118,-117]	9	[-105,-100]	15.9 (41.5 - 25.6)
	68886	50.0 (60.0−10.0)	96.2	12.0	5.1×	[-107,-100]	10	[-115,-114]	50.0 (60.0−10.0)
	Overall	17.0 (41.6−24.6)	31.8	3.6	4.8×	N/A	N/A	N/A	17.0 (41.6−24.6)

On the contrary, we notice the gain on the same channel in C2-A is negligible. It proves that channel 850 has been enhanced by aggregation with 5G cells, which makes cells on channel 850 more likely outperform others. It also shows the urgency of reconfiguration: As the operators are rolling out mmwave cells, the potential gap between good and bad cells will be further enlarged. Therefore, cell selection is encouraged to consider performance and RPerf could be easily patched onto the infrastructure to prevent under-utilization.

Reconfiguration on the level of frequency channel. We also notice that the overall gain in C2 is larger than C1. This is mainly because 4 bad channels on band 66 (the last 4 channels in Table 7.3) are densely deployed and frequently selected as PCell. Cells on those channels do not accept any SCell, which results in much more narrow channel width compared to others. Therefore, tuning parameters on band 66 towards inter-freq handovers could greatly save loss.

This finding sheds light on reconfiguration on the level of frequency channel, instead of per cell. It reveals the tendency of operators to manage radio resource for each frequency channel, instead of individual cells. For example, in our dataset, cells on the same frequency share the same channel width. In addition, operators may just support limited combinations of frequencies for carrier aggregation. Such behavior of radio planning make cells on the same frequency share common capabilities. Therefore, reconfiguration on a channel is aligned with such behavior of radio planning. It will get promising gain for making good use of the commonality within one channel and discrepency among channels.

7.4 Discussion

RPerf aims to optimize data performance impacted by cell selection but itself is far away from "optimal" due to its inherent limitations and remaining issues.

Limited traces. We design and evaluate RPerf based on real-world traces. However, our traces are limited as they are collected from only a few mobile phones. Network operators

have a much larger sample set and complete ground truth regarding their cell selection configurations and operations. In addition, the handover model is extracted from their operation directly, and thus, the overall reward is much less biased. Our effort is to leverage what we can to demonstrate reconfiguration potential and call for attention and action from network operators. Suppose the standards could enforce consideration of throughput onto the cell selection. In that case, we will have a complete view of network performance and new perspectives to solve this problem in policies and mechanisms other than configurations.

Spatial granularity for reconfiguration. In this work, reconfiguration is performed on two selected regions of 2 – 2.5 km^2. As the regions are small, we do not split them for finer-grained reconfiguration. However, given broad coverage, an operator must split the entire area into smaller regions to reconfigure. A practical solution needs a proper spatial granularity to trade off the performance gain and practicality. It can be aligned with network deployment (the network is divided to serve different geographical areas). An alternative solution is to start with reconfiguration in small regions and merge adjacent regions with close parameters into larger ones. After the merge, reconfiguration is performed in vast regions. It is feasible as regions nearby are likely to share common features in radio resource management, data usage, etc. As the operator updates cell deployment and radio planning, previously merged regions may gradually lose their commonality. Therefore, the operator should regularly adjust reconfiguration regions after a big system update. To validate this solution, we will enable measurements in much broader areas. We will also release our tools to conduct measurement and analysis of cell selections at places of user interest.

Run-time dynamics. Run-time dynamics like radio quality fluctuation, scheduling and cell load could be impact factors. We aim to reduce the impact of transient factors and focus on for the overall reward affected by *persistently* worse cell selections. This is first validated by [DLG20] and further confirmed by our latest experiments with 5G/4G. Accordingly, reconfiguration proposed in RPerf is to prevent such persistent performance loss by promoting cell selections towards better cells. We admit that reconfiguration learned from

the historical data may not work for cell selections in the next second. But it seeks for the overall reward of statistical significance which eliminates that impact of runtime dynamics. An alternative solution is to make decision based on run-time situations, which provides a different angle for cell selection. It could ensure quality cell selection individually. This will complement the proposed reconfiguration and warrants future work.

5G-related issues. In our study, AT&T and T-Mobile deploys their 5G networks by adopting dynamic spectrum sharing (DSS) technology which runs two generations of cellular networks (4G and 5G) over the same frequency channel. As a result, the achieved speed in 5G is quite comparable to the legacy 4G, despite of small speed growth. This is why we observe similar gains in both cities while no 5G is deployed in C2. However, with more advanced 5G technologies including mmWave and Standalone 5G, we believe that 5G can be much faster, which will raise a pressing need for reconfiguration to reconfiguration to reduce poorly-performed cell selection and promote good ones.

Miscellaneous. There are unexplored design options in triggering and executing reconfiguration. Instead of heuristic-based search, advanced ML techniques like neural networks can be exploited with a much larger dataset. Reinforcement learning seems to fit by iteratively tune parameters. When to trigger can be performed with periodic checking, runtime monitor over down-sampled traces, or hybrid. RPerf is far away from a perfect solution. Instead, it is more like a proof-of-concept demo which demonstrates that reconfiguration is simple, ready-to-launch with immediate benefits. More practical solutions will follow once network operators take actions.

CHAPTER 8

Mobility Support on Open-Source Platform

We set up a mobility testbed on FLORA (**F**lexible **M**obile **N**etwork **P**latform), an open-source software-defined 4G/5G system [Flo]. It provides a quick solution to trying new designs in real cellular networks. Figure 8.1 shows the architecture with a server to run the core network and a software-defined radio as the base station. To facilitate mobility experiments in the real world, we support handover and CA for commodity phones based on srsRAN [srs].

8.1 Handovers

Handover is the essential function of a testbed for mobility support. Unfortunately, the existing software [srs] cannot perform handovers on commodity phones because the *execution* (aka step ⑤ in Figure 2.1) is over-simplified compared to the standard in practice. FLORA produces two extensions. First, it enables *intra-base-station* handover for real smartphones. In other words, we can migrate the client among cells on the same base station. Second, FLORA builds *virtual* cells based on limited hardware resources (e.g., two cells per X300) to support handovers among multiple cells with various policies.

Primer: Process of handover execution We first show more details about execution, the last step of handover. It contains four actions at the base station [3GP22] (Figure 8.2):

1. Send handover command, including configurations of the target cell, to the client.

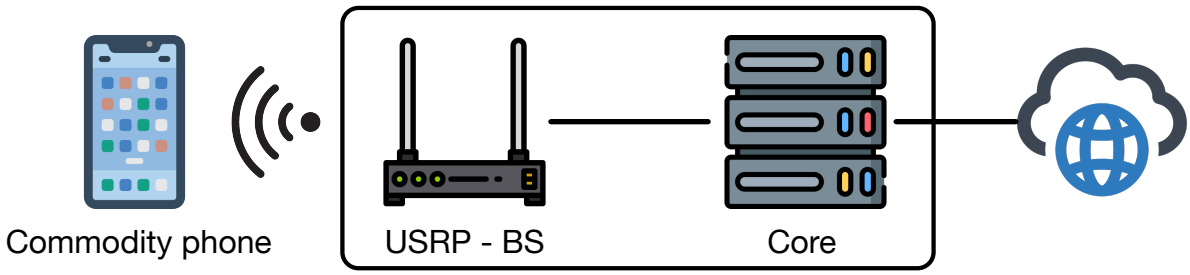


Figure 8.1: Flora: Flexible Mobile Network Platform

2. Disconnect the client by closing the data radio bearer[1] and uplink signaling radio bearer.

3. Schedule uplink transmission for the client after receiving random access signal.

4. Set up data radio bearer to resume the data transfer.

Handover for commodity phones How does FLORA support handovers on commodity phones? It checks in necessary functions which are omitted by the current design:

- FLORA constructs the handover command with all *mandatory* configurations of the target cell. For example, add the RACH-ConfigCommon element to instruct the client on random access to the new cell.

- FLORA updates security keys for communication with the new cell. Initially, the handover complete message was received by the eNodeB but did not pass the integrity check. According to 5G/4G standards, the client and network will calculate new keys after handover. The calculation takes old keys, cell ID, and frequency ID as input based on a specific algorithm negotiated during service attach. FLORA implements the security feature and thus performs successful handovers.

We test handovers with USRP X300 and Google Pixel as the base station and client. The demo [han22] shows more details.

[1]The "bearer" here refers to a logical channel between the client and the base station.

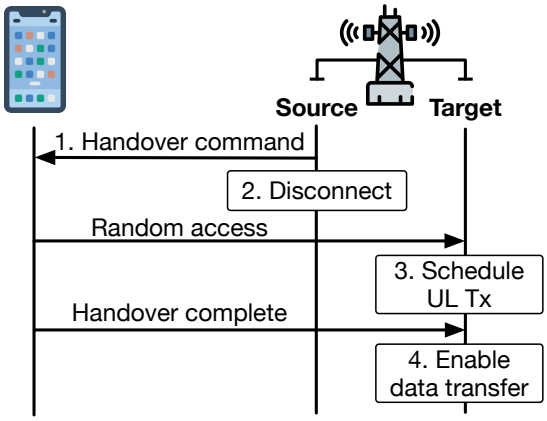


Figure 8.2: Handover execution.

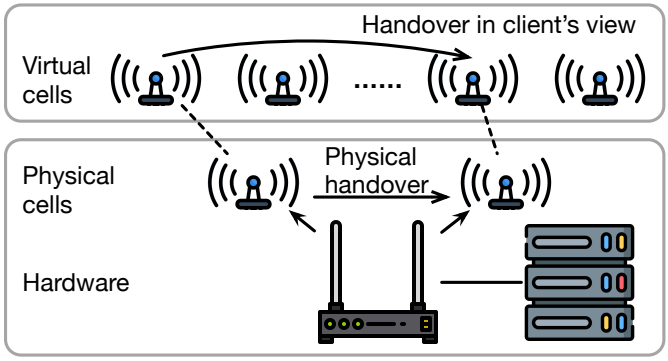


Figure 8.3: Handover among virtual cells.

Handover among virtual cells How can we test handover involving more cells? The straightforward idea is to add more base stations, given a limited number of cells on each one. However, some functional modules are not supported in the current architecture, like inter-base-station communication and connecting multiple base stations to one core network. Therefore, FLORA provides a quick solution with virtual cells to enable the mobility setting of multiple cells with different handover policies.

Figure 8.3 shows how virtual cells work. The system contains two physical cells for wireless communication with the client. On top of that, there are many ($\gg 2$) virtual cells with different upper-layer settings (specifically for handover and data transmissions). At any time, the serving cell adopts the upper-layer configuration of one virtual cell and the lower-layer setting of one physical cell. During handovers, the base station selects a virtual cell. The client is migrated from one cell to the other on the physical layer. After the transition, the new cell copies the selected virtual cell configurations for the upper layer. With this platform, we can emulate handovers among multiple cells and test upper-layer features (especially mobility support).

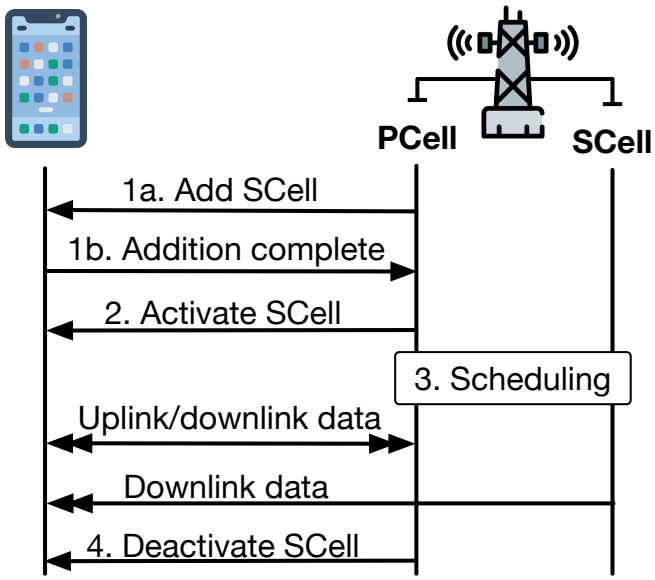


Figure 8.4: Bring SCell in use.

8.2 Carrier Aggregation

5G/4G can serve one user with multiple cells together with CA. The function is feasible on open-source platforms as the latest USRP models (e.g., X300/X310/N310) can open 2 to 4 channels, each serving as one cell. However, existing open-source CA implementation [srs] does not work with commodity phones. After inspection, we find the primary cause: the base station configures uplink data transfer on SCell(s) which is not yet supported by real smartphones. FLORA has fixed the issue and is compatible with both 5G/4G standards and commodity phone's capability. Next, we introduce the standardized procedures of using an SCell, based on which we further explain FLORA's implementation.

Primer: Transmit data on SCell(s) As Figure 8.4 shows, the base station takes four steps to use an SCell given client connected to the same PCell.

1. PCell sends a command to tell the client which SCell to add and expects confirmation from the client. In this step, if the client cannot support the SCell's configuration, it will reply with refuse. Specifically, the existing CA implementation fails here since the configuration includes uplink data transfer on SCell, but the actual phone cannot do it.

2. PCell sends an activation flag to the client, which synchronizes both ends to get ready for data transmission on SCell. After this point, the network starts scheduling resources on SCell; the client receives data and sends back channel quality indicators (CQI).

3. The network schedules data transmission over PCell and SCell, and grants radio resources accordingly. Now, SCell starts working for downlink.

4. SCell can be deactivated at runtime. While deactivated, it stops data transmission.

Flora's implementation. FLORA makes CA work with commodity smartphones by disabling uplink data transfer on SCell. We udpate steps 1 and 3 (Figure 8.4) and reuse the existing implementation of steps 2 and 4.

- In step 1 , FLORA's base station configures SCell without uplink data transmission channel (i.e., PUSCH). We set relevant parameters to the proper value to disable the uplink data. For example, FLORA disables one major configuration block, pusch-ConfigDedicatedSCell-r10 [3GP15], for SCell uplink data by setting its present flag to false. Please note that FLORA still keeps the uplink control channel on SCell because downlink data transfer relies on feedback from the client to work correctly, like MAC-layer ACK/NACK and fine-grained channel quality indicators.

- In step 3, FLORA skips resource allocation for uplink data if the cell is not PCell. Meanwhile, FLORA keeps the scheduling for uplink control signals as long as the SCell is active.

We test FLORA with USRP X300 (opening up to 2 cells) and Google Pixel. FLORA has successfully enabled CA on the smartphone with two cells, and please refer to our demo [ca 22] for more details. We believe our implementation is generic and supports more cells' aggregation if testing on advanced USRP with more channels (e.g., USRP N310).

CHAPTER 9

Related Work

We first present the state-of-the-art, which aims at enhancing mobility support. Unlike our study, most prior works are confined to the conventional mobility support design; or focus on the performance of emerging technologies but are not aware of the inherent correlation with cell selection. Next, we introduce the state-of-the-art in other fields, which shed light on our design to combat the new challenges of 5G/4G and beyond.

9.1 State-of-the-art on Mobility Support

Reliable and fast mobility management. Reliable and fast mobility management have been an active topic for years. Most efforts follow the wireless signal strength-based design today and explore how to refine its signaling procedures [QWP17, LYP17], handover decision [XNM19, TLL18], transport-layer data speed in mobility [WZN19, LXL18], policy conflicts [LDL16, YLL18], to name a few. As reliable mobility support becomes more challenging in extreme mobility, our study revisits the wireless-based design, unveils diverse network failures and policy conflicts below the IP layer, and proposes REM as the first movement-based reliable mobility management.

Exploration of multi-carrier access. Multi-carrier access offers a promising alternative to the dominant single-carrier cellular access technology. Early systems support multi-carrier access inside commodity phones using dual SIM cards [DNS11, Wik18] and a single universal SIM card [Goo, App, Eng16]. Recent research has focused on improving multi-carrier access,

such as better performance [KSR16, LDP16] and concurrent access to multiple carriers in 5G [KVB17]. We complement prior work by investigating policy management for multi-carrier access, a topic not been studied so far. While our study draws insights from the operational Google Fi, our paper is also forward-looking by extending policy-based switch to a more generic setting (more carriers, policy forms, etc..).

Well-utilized CA and radio resources. There are active efforts to achieve effective CA and improve the aggregated throughput using different techniques [DLH20, JEA21, GMC17, LCC14]. Some prior studies devised resource scheduling or joint carrier selection to maximize utilization [JEA21, GMC17, LCC14], which did not address mobility support problems. [DLH20] focuses on selecting better serving cells with device-centric solutions. Our solutions, RPerf and CA++, make huge differences: RPerf is to prevent under-utilization of radio resources rather than fix the issue afterward. To fit in the extreme mobility with mmWave cells, CA++ takes a revolutionary step: Instead of seeking the "local optimal" within the current architecture, it targets the inherent limitations that impede fulfilling CA potentials. Therefore, CA++ transforms the sequential, cell-by-cell operations into a group-based manner and fundamentally improves CA's effectiveness.

9.2 Inspiration from State-of-the-art in Other Fields

OTFS modulation. The delay-Doppler domain from the radar community and recent advances in OTFS modulation [HRT17, RPH18, RPH19] could help refine wireless robustness. Our solutions leverage the technology but move beyond wireless modulation and generalize to mobility scenarios. Specifically, REM adopts OTFS for movement-based design considering Doppler frequency shift under extreme mobility. Furthermore, both REM and CA++ exploit the compact, frequency-independent multi-path representation in the delay-Doppler domain to make cross-channel estimation feasible (see more details below).

Cross-channel estimation. [Vas16, Bak19] perform cross-channel estimation based on

common propagation paths. REM extends the design philosophy to mobility scenarios and simplifies the estimation in the delay-Doppler domain. CA++ adopts cross-channel estimation to speed up measurement over a wide spectrum. It further complements REM to achieve highly reliable and accurate inference over the wide frequency spectrum which is already in use for 5G CA and will continue to expand.

Policy inconsistency. Instability and policy inconsistencies have been examined in other networking systems, such as BGP routing [LMJ98, GW99, GR01] and SDN and data center networks [SMR14, JLG14, LWZ13]. Our work has a different setting (mobile networks) and addresses a different mechanism (mobility support). Stability results have been recently reported for configurable handovers within a carrier [LXP16, LDL16]. We find that instability happens more frequently under extreme mobility as one engineering remedy to mitigate connectivity failures. Thus, REM simplifies the handover configurations to enable easy-to-satisfy conditions for conflict-free policies. When it comes to multi-carrier access, our problem is different because we examine policy conflicts among inter-carrier policies and between inter-carrier and intra-carrier policies. We devise policy guidelines at the inter-carrier level by assuming policy autonomy within each carrier.

CHAPTER 10

Conclusion and Future Work

This chapter briefly summarizes our work, emphasizing on how to address new challenges for mobility support in 5G/4G. Next, we share insights and lessons learned from the study. Finally, we propose future work to enhance mobility support for 5G/4G and beyond.

10.1 Summary of Results

REM: Reliable extreme mobility management. Extreme mobility has become popular with emergent high-speed mobility scenarios (rails, vehicles, drones, etc.) and high-frequency radio (e.g., mmWave). However, 5G/4G is not well prepared to support them. The fundamental problem is that 5G/4G's *wireless signal strength-based* design is vulnerable to dramatic wireless dynamics in extreme mobility. We thus devise REM, *movement-based* mobility management in the delay-Doppler domain. REM relaxes the feedback with cross-band estimation, simplifies the policy for provable conflict-freedom, and stabilizes the critical signaling traffic scheduling-based OTFS modulation. As a result, REM gets rid of policy conflicts and achieves low failure ratios similar to static and low mobility scenarios.

Resolving policy conflicts given multi-carrier access. Multi-carrier cellular access provides mobile users better service compared to single-carrier access. The Google Fi solution already shows early signs of success and great benefits without requiring cellular infrastructure upgrades. Multi-carrier access has extended mobility support to a two-tier switch: It first selects the most preferred mobile carrier dynamically before proceeding to cell selec-

tion within the carrier (handover). The policy-based switching is adopted as the primary mechanism, a double-edged sword. On the one hand, the inter-carrier policy has excellent features and is needed by MCSPs. This is evident from operational practice in Google Fi and experiences from BGP and data center networks. On the other hand, conflicts arise between the customized inter-carrier switch policy and the standardized intra-carrier cell selection, akin to BGP loops. We identify several such cases, derive theoretical conditions when policy occurs, and provide practical guidelines to resolve the issues. While the detailed carrier selection algorithm may evolve, we believe that the framework in this study is fundamental, and our results will continue to help with stable multi-carrier access.

CA++: Enhancing carrier aggregation. CA is a promising technology to combine the chunks of expanding frequency spectrum and boost throughput. However, the current CA mechanisms fail to catch up with increasing resources and capabilities. The problem stems from the current cell-by-cell CA practice due to the tension between timely operations and quality selections. We devise CA++ to solve the problem fundamentally. The principle is to enable group-based selection with two major forces. On the one hand, parallel channel quality estimation can speed up the measurement of many cells. More importantly, it guarantees high inference accuracy over a wide spectrum in 5G and beyond. On the other hand, CA++ transforms CA operations from cell-by-cell to group-by-group, coherent to the nature of CA. Our experiments and trace-driven emulation confirm its effectiveness in enhancing throughput.

RPerf: Reconfiguring cell selection towards better performance. We present RPerf to prevent improper cell selection, which fails to select the cells with high throughput in today's 5G/4G networks. It attempts to re-configure parameters in cell selection policies to make the connectivity-centric design aware of throughput and thus fulfill the potential of network resources. We devise a simple reconfiguration algorithm based on profiling and heuristic searching. It is compatible with the 5G/4G standard and infrastructure and thus ready to launch. Our trace-driven emulation with 5G datasets shows good throughput gain,

with a data speed boost in 30% of test cases and a median increase of 89.1%.

Mobility testbed on Flora. We set up a 5G/4G mobility testbed by extending the current functions of FLORA, a flexible software-defined mobile network platform. The extension focuses on mobility functions, handovers and CA. More importantly, they run successfully on commodity phones, unlike other testbeds limited to customized clients on FPGA. Therefore, it provides an experiment setting closer to operational mobile networks.

10.2 Insights and Lessons

We present the insights and lessons for 5G/4G mobility support. In addition, we seek to be forward-looking by abstracting and generalizing the issues in a more generic setting beyond mobility support. The ultimate goal is to embrace, rather than suppress, the new challenges and to meet the high standard of 5G/4G networks in terms of reliability, throughput, etc.

Decouple wireless from mobility management for robustness. Under extreme mobility, the handover failure ratio goes up due to dramatic wireless dynamics. Instead of directly tackling wireless and combating the dynamics, we seek to *decouple* the wireless from mobility support. We shift to client movement, which inherently decides the wireless but evolves much slower. Therefore, the client movement is informative and robust to drive mobility management in extreme mobility. Aimed at highly reliable network, our work takes the initial step toward movement-based mobility management design. Beyond reliability, this idea can be generalized to broader scopes such as channel prediction, wireless performance optimization, geographical routing, and delay-Doppler-based localization. We hope our work could stimulate more innovation toward intelligent and robust mobile networks.

Use parallelization to tackle increased diversity in 5G/4G. Inherently, the dilemma between responsiveness and quality originated from the increased diversity of radio frequency. Therefore, 5G/4G should parallelize the operations over frequency channels. Specifically, we adopt this philosophy for cross-channel estimation. It is based on a critical observation that

one base station typically deploy cells on different frequency channels, but they share the underlying propagation paths to the client. Moreover, it makes more sense with CA, as only cells on the same base station can be aggregated in most cases. Therefore, the client can only measure one cell, retrieve path features that can be separated from frequency, and map them to other cells for parallel estimation. This approach speeds up the measurement and feedback without reducing the cells to be explored. Our work exploits it to relax the reliance on sequantial measurement and thus mitigates handover failures with earlier action under extreme mobility. Furthermore, we faciliate cross-channel estimation with higher precision over wide frequency spectrum (e.g., sub-1GHz through 39GHz) to select good cell groups quickly and thus fulfill the potential of 5G CA.

Parallel cross-channel estimation is not limited to mobility support. Mobile networks heavily rely on feedback from clients to operate on fast-varying wireless channels. Along with the surge in spectrum resources and diversity in 5G/4G, the cost of real-time feedback rises and turns unaffordable with increased payloads and delay. Therefore, the network should consider parallelization based on underlying invariant characteristics to tackle the diversity and make light-weighted and informative feedback.

Examine the interplay between existing architecture and new technology to integrate. It is not a new problem that adding a solid design to an established system produces a mal-functioned combination. Out of a different goal, the new technology may conflict with the existing architecture in some aspects and thus cause harm like service instability, performance downgrade, etc. However, the operational networks still tend to ignore the coordination when rolling out new technology. Our work fixes a similar issue with multi-carrier access, which incurs policy inconsistencies with the legacy intra-carrier cells switching (handover). To solve the problem, we prioritize the decisions of the standardized intra-carrier handover over the new inter-carrier selection for conflict-free interplay. It makes a practical case where coordination is indispensable to integrate a new technology smoothly. We also gain wisdom about regulating the new design only if we cannot interfere with the

existing system in reality.

Upgrade mobility support to keep up with network advances. The industry has been actively deploying more resources and enhancing network capabilities. This dissertation shows that making good use of what we currently have is as important as developing new advances. In particular, mobility support is critical as it decides on serving cells that are the foundation for obtaining any advanced network capabilities. Our work makes two cases following this insight: First, RPerf turns the connectivity-centric cell selection into throughput-aware via reconfiguration. It seeks to make better utilization of the increased network resources. Second, CA++ upgrades the mobility support mechanisms to unleash CA potential. It challenges the cell-by-cell design, which impedes CA potential, and adopts the group-based method for fundamental improvement.

We believe that this insight complies with the current 5G and future networks with rapidly increasing resources and capabilities. However, while wireless resources proliferate along various dimensions (e.g., multi-connectivity, antenna diversity), more efforts are needed to revisit the existing mechanisms beyond mobility support. In addition, we should ensure that the method of using wireless resources keeps up with growing capabilities.

10.3 Future Work

Enhance mobility management to support thrilling applications. In 5G and even 6G for the future, a lot of revolutionary and thrilling applications appear with a more stringent requirement on reliability (failure $< 10^{-7}$), data speed (reaching Tbps), and latency (sub-1ms) [DKP21] even on the move. They include high-speed drones (UAV), collaborative autonomous driving, smart industry, etc. To meet the ultra-high standard, we still desire breakthroughs in mobility support. For example, 5G/4G takes "hard" handovers as they have to disrupt data transfer during the switch for tens of milliseconds. However, it is not tolerable by any ultra low latency applications. Therefore, the ultimate goal is to perform

"soft" handovers and achieve completely seamless data service.

Intelligent and collaborative decision-making for optimal throughput. Selecting serving cell(s) remains a challenge to optimize the throughput. Specifically, we have two fundamental problems to address. First, how to optimize the selection from the perspective of one cell (or base station)? The cell should pick appropriate metrics and intelligent mechanisms to estimate candidates' performance and decide. Second and more importantly, how to coordinate the decisions from neighbor cells? 5G/4G needs a good balance of load among neighbor cells to achieve high data speed for users and high resource efficiency for the network. Cloud-based, virtualized radio access networks (RAN) provide a promising solution for neighbor base stations to collaborate efficiently.

Adaptation to multi-connectivity (DC/MC). Currently, 5G NR is rolling out dual-connectivity (DC) and even multi-connectivity (MC) [SG], which aggregates cells from two or multiple base stations, respectively. In this case, we need to adapt mobility support to the new dimension, say connectivity. Note that both DC and MC share a similar goal and requirement as CA: To expand the aggregated channels, and demand fast and good cell selection. Therefore, the design philosophy of cross-channel estimation will continue to assist select cells on two or more base stations. More significant benefits raised by DC/MC may drive further mobility support upgrades, like handover without disruption or simultaneous access to multiple mobile carriers, to name a few. In that case, we should revisit the mobility support to address new challenges.

Tackle new diversity accompanying beamforming. Considering beamforming in 5G [3GP17b], the cell and mobile client may adjust the antenna array at runtime. It will challenge the foundation of cross-channel estimation: Even though multiple cells reside on the same base station, they may use different dynamically-changed antennas and thus do not share propagation paths. Mobility support will also incorporate beam management [GPR18], i.e., adjusting antenna parameters. In that case, the network will take the antenna as a new aspect for measurement and selection. Intuitively, our design would still help with effective

adaptation. For example, to accelerate measurement, the device does not need to measure all carriers for each antenna; Instead, per antenna, we measure a specific cell and infer others operating on different frequencies. In future work, we will analyze various use scenarios of beamforming and extend our design accordingly.

Automatic verification/upgrade for mobility support. This dissertation shows the increased heterogeneity in mobile networks and accompanying challenges for mobility support. All the problems identified by our work plus future challenges above indicate higher and higher complexity of manual examination and upgrade of mobility management and other radio control mechanisms. The cost would be prohibitive to revisit and adapt the entire architecture to every new scenario. Instead, we should consider an automatic upgrade as the solution, e.g., AI/ML-based examination, suggestion, and correction. There are two typical cases where we need update the mobility support mechanisms. First, the network plans to integrate new technology. The automation should ensure that the integration will not hurt connecitivty, and the current mobility management will not impede the effectiveness of the new advance. The second case is to support a new application. The automation will customize the mobility management to meet the application's performance requirement.

APPENDIX A

Supporting Materials for Chapter 4

A.1 Stable Delay-Doppler Channel

The variance of delay-Doppler channel $h_w(\tau, \nu)$ over time $\frac{\partial h_w(\tau,\nu)}{\partial t} = \frac{\partial h_w(\tau,\nu)}{\partial \tau}\frac{\partial \tau}{\partial t} + \frac{\partial h_w(\tau,\nu)}{\partial \nu}\frac{\partial \nu}{\partial t}$ relates to the path delay and Doppler variance. The path delay $\tau = \frac{d}{c} \propto \frac{vt}{c}$ (d is path length, a is client acceleration), so its change $\frac{\partial \tau}{\partial t} \propto \frac{v+at}{c} \to 0$ since $v \ll c$ even under extreme client movement (e.g.,, 10^{-7} for v=500km/h). The Doppler change $\frac{\partial \nu}{\partial t} \propto \frac{\partial (fv/c)}{\partial t} = \frac{f}{c}a$ relates to the client's acceleration a and is negligible unless the client speeds up or down (infrequent in high-speed rails). Therefore, $h(\tau, \nu)$ remains constant in a much longer duration than $H(t, f)$ (whose coherence time $T_c \propto \frac{1}{\nu_{max}}$).

A.2 Proof of Theorem 4.3.1

Proof. We prove that when $P \leq \min(M, N)$ and $\tau_p - \tau_{p'} = k\Delta\tau$ and $\nu_p - \nu_{p'} = l\Delta\nu$ for any p, p', the delay-Doppler decomposition $\mathbf{H} = \mathbf{\Gamma P \Phi}$ results in unitary matrices $\mathbf{\Gamma}$ and $\mathbf{\Phi}$ and $M \times N$ diagonal matrix $\mathbf{P}$, thus being a SVD decomposition. Given $P \leq \min(M, N)$ paths, we can always insert "virtual paths" (with 0 attenuation) and expand $\mathbf{P}$ as a $M \times N$

diagonal, non-negative matrix as follows[1]:

$$\mathbf{P} = \begin{bmatrix} |h_1| & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & |h_2| & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & |h_P| & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}$$

This is equivalent to a $\min(M, N)$-path channel with $|h_p| = 0$ when $p > P$. Therefore, we only need to prove Theorem 4.3.1 always holds when $P = \min(M, N)$, and $P < \min(M, N)$ will also hold with this expansion. The following proof focuses on $M < N$ so that $P = M$; $M > N$ follows the similar proof.

First consider the delay spread matrix $\mathbf{\Gamma}$. Note that

$$\Gamma(k, p) = \frac{1}{M} \sum_{d=0}^{M-1} (e^{j2\pi\Delta\tau\Delta f})^{kd} \cdot (e^{-j2\pi\Delta f})^{d\tau_p} = \sum_{d=0}^{M-1} \Gamma_1(k, d)\Gamma_2(d, p)$$

where $\Gamma_1(k, d) = \frac{1}{\sqrt{M}}(e^{j2\pi\Delta\tau\Delta f})^{kd}$ and $\Gamma_2(d, p) = \frac{1}{\sqrt{M}}(e^{-j2\pi\Delta f})^{d\tau_p}$. So we can factorize $\mathbf{\Gamma} = \mathbf{\Gamma_1}\mathbf{\Gamma_2}$, where $\mathbf{\Gamma_1}, \mathbf{\Gamma_2} \in \mathbb{C}^{M\times M}$

$$\mathbf{\Gamma_1} = \frac{1}{\sqrt{M}} \begin{bmatrix} (e^{j2\pi\Delta\tau\Delta f})^{0\cdot 0} & \cdots & (e^{j2\pi\Delta\tau\Delta f})^{0\cdot(M-1)} \\ \cdots & (e^{j2\pi\Delta\tau\Delta f})^{kd} & \cdots \\ (e^{j2\pi\Delta\tau\Delta f})^{(M-1)\cdot 0} & \cdots & (e^{j2\pi\Delta\tau\Delta f})^{(M-1)\cdot(M-1)} \end{bmatrix}$$

$$\mathbf{\Gamma_2} = \frac{1}{\sqrt{M}} \begin{bmatrix} (e^{-j2\pi\Delta f})^{0\cdot\tau_1} & \cdots & (e^{-j2\pi\Delta f})^{0\cdot\tau_M} \\ \cdots & (e^{-j2\pi\Delta f})^{d\tau_p} & \cdots \\ (e^{-j2\pi\Delta f})^{(M-1)\cdot\tau_1} & \cdots & (e^{-j2\pi\Delta f})^{(M-1)\cdot\tau_M} \end{bmatrix}$$

since $P = \min(M, N) = M$. We show that both $\mathbf{\Gamma_1}$ and $\mathbf{\Gamma_2}$ are unitary, thus $\mathbf{\Gamma} = \mathbf{\Gamma_1}\mathbf{\Gamma_2}$

[1]This is how SVD is widely used for matrix dimensionality reduction.

being unitary. For $\mathbf{\Gamma_1}$, we have

$$\mathbf{\Gamma_1}\mathbf{\Gamma_1^*}(k,k') = \sum_{d=0}^{M-1} \Gamma_1(k,d)\Gamma_1^*(d,k') = \frac{1}{M}\sum_{d=0}^{M-1} e^{j2\pi\Delta\tau\Delta f(k-k')d}$$

If $k = k'$, we have $\mathbf{\Gamma_1}\mathbf{\Gamma_1^*}(k,k) = 1$, Otherwise

$$\mathbf{\Gamma_1}\mathbf{\Gamma_1^*}(k,k') = \frac{1-e^{j2\pi\Delta\tau\Delta f(k-k')M}}{1-e^{j2\pi\Delta\tau\Delta f(k-k')}} = \frac{1-e^{j2\pi(k-k')}}{1-e^{j2\pi\Delta\tau\Delta f(k-k')}} = 0$$

since $\Delta\tau = \frac{1}{M\Delta f}$, so $\mathbf{\Gamma_1}\mathbf{\Gamma_1^*} = I_M$ and $\mathbf{\Gamma_1^*}\mathbf{\Gamma_1} = (\mathbf{\Gamma_1}\mathbf{\Gamma_1^*})^* = I_M$ is unitary. For $\mathbf{\Gamma_2}$, we have

$$\mathbf{\Gamma_2^*}\mathbf{\Gamma_2}(p,p') = \sum_{d=0}^{M-1} \Gamma_2^*(p,d)\Gamma_2(d,p') = \frac{1}{M}\sum_{d=0}^{M-1} e^{j2\pi\Delta f(\tau_p-\tau_{p'})d}$$

If $p = p'$, we have $\mathbf{\Gamma_2^*}\mathbf{\Gamma_2}(p,p) = 1$. Otherwise

$$\mathbf{\Gamma_2^*}\mathbf{\Gamma_2}(p,p') = \frac{1-e^{j2\pi\Delta fM(\tau_p-\tau_{p'})}}{1-e^{j2\pi\Delta f(\tau_p-\tau_{p'})}} = \frac{1-e^{j2\pi k}}{1-e^{j2\pi\Delta f(\tau_p-\tau_{p'})}} = 0$$

since $\tau_p - \tau_{p'} = k\Delta\tau$ for some integer k and $\Delta\tau = \frac{1}{M\Delta f}$. Therefore, $\mathbf{\Gamma_2^*}\mathbf{\Gamma_2} = I_M$ and $\mathbf{\Gamma_2}\mathbf{\Gamma_2^*} = (\mathbf{\Gamma_2^*}\mathbf{\Gamma_2})^* = I_M$ are also unitary, and $\mathbf{\Gamma^*}\mathbf{\Gamma} = \mathbf{\Gamma}\mathbf{\Gamma^*} = \mathbf{\Gamma_1}\mathbf{\Gamma_2}\mathbf{\Gamma_2^*}\mathbf{\Gamma_1^*} = I_M$ is unitary. Similarly we can prove $\mathbf{\Phi}$ is also unitary when $\nu_p - \nu_{p'} = l\Delta\nu$ for any p, p'. So $\mathbf{\Gamma}$, $\mathbf{P}$ and $\mathbf{\Phi}$ meets the definition in SVD, and $\mathbf{H} = \mathbf{\Gamma}\mathbf{P}\mathbf{\Phi}$ is a SVD decomposition. □

A.3 Derivation of Algorithm 1

We detail how Algorithm 1 leverages SVD to estimate per-path delay-Doppler for cross-band estimation. Given band 1's channel estimation matrix $\mathbf{H_1}$, we run SVD and use it as an approximation of $\mathbf{H_1} = \mathbf{\Gamma}\mathbf{P}\mathbf{\Phi_1}$. Note that band 1's $\mathbf{\Gamma}\mathbf{P}$ is frequency-independent and thus can be reused by another band. To estimate band 2's channel $\mathbf{H_2} = \mathbf{\Gamma}\mathbf{P}\mathbf{\Phi_2}$, we need to infer $\mathbf{\Phi_2}$ from $\mathbf{\Phi_1}$. To do so, note that

$$\Phi_1(l\Delta\nu_i, \nu_p^1) = \sum_{c=0}^{N-1} e^{j2\pi(l\Delta\nu_i-\nu_p^1)cT} = \frac{1-e^{-j2\pi\nu_p^1 NT}}{1-e^{j2\pi(l\Delta\nu_i-\nu_p^1)T}}, \forall l$$

$$\Gamma(k\Delta\tau, \tau_p) = \sum_{d=0}^{M-1} e^{-j2\pi(k\Delta\tau-\tau_p)d\Delta f} = \frac{1-e^{j2\pi\tau_p M\Delta f}}{1-e^{-j2\pi(k\Delta\tau-\tau_p)\Delta f}}, \forall k$$

So we have

$$\frac{\Phi_1(p,l)}{\Phi_1(p,l')} = \frac{1-e^{j2\pi(l'\Delta\nu-\nu_p^1)T}}{1-e^{j2\pi(l\Delta\nu-\nu_p^1)T}}, \frac{\Gamma(k,p)}{\Gamma(k',p)} = \frac{1-e^{-j2\pi(k'\Delta\tau-\tau_p)\Delta f}}{1-e^{-j2\pi(k\Delta\tau-\tau_p)\Delta f}}$$

for any (k,k') and (l,l'). Then we can extract

$$e^{-j2\pi\nu_p^1 T} = \frac{\Phi_1(p,l)-\Phi_1(p,l')}{\Phi_1(p,l)e^{j2\pi l\Delta\nu T}-\Phi_1(p,l')e^{j2\pi l'\Delta\nu T}}$$

$$e^{j2\pi\tau_p\Delta f} = \frac{\Gamma(k,p)-\Gamma(k',p)}{\Gamma(k,p)e^{-j2\pi k\Delta\tau\Delta f}-\Gamma(k',p)e^{-j2\pi k'\Delta\tau\Delta f}}$$

When the conditions in Theorem 4.3.1 was not strictly satisfied (mainly due to small (M,N) and thus imperfect sampling), SVD and above derivations are approximations of delay-Doppler estimation. For high accuracy, Algorithm 1 computes the average of above delays/Dopplers across all (k,k') and (l,l') (line 4–5). Then we can convert each path's Doppler $\nu_p^2 = \frac{f_2}{f_1}\nu_p^1$ for every path p (line 6). Now with $\{h_p,\tau_p,\nu_p^2\}_{p=1}^{P_{max}}$, Algorithm 1 follows the definitions in §4.3.2, construct $\mathbf{\Phi_2}$ and estimate cell 2 as $\mathbf{H_2} = \mathbf{\Gamma P \Phi_2}$ (line 9–10).

A.4 Proof of Theorem 4.3.2

Proof. We prove it by recursion. If $\Theta_{A3}^{i\to j}+\Theta_{A3}^{j\to i}\geq 0, \forall i\neq j$, the following two conditions will not happen simultaneously:

$$\begin{cases} \mathrm{SNR}_j > \mathrm{SNR}_i + \Theta_{A3}^{i\to j} & (c_i \to c_j) \\ \mathrm{SNR}_i > \mathrm{SNR}_j + \Theta_{A3}^{j\to i} & (c_j \to c_i) \end{cases}$$

This asserts that no 2-cell persistent loops will occur for any $(\mathrm{SNR}_i, \mathrm{SNR}_j)$. Assume $\Theta_{A3}^{i\to j}+\Theta_{A3}^{j\to i}\geq 0$ asserts for any 1, 2, ..., $(n-1)$-cell loop freedom among $c_1, c_2, \ldots, c_{n-1}$. Now consider n cells $c_1, c_2, \ldots, c_{n-1}, c_n$. Since $\Theta_{A3}^{i\to j}+\Theta_{A3}^{j\to i}\geq 0, \forall i\neq j$, any $1,2,\ldots,(n-1)$-cell loop freedom still retains among these cells. Then consider if n-cell loop $c_1\to c_2\to\cdots\to c_n\to c_1$ can happen for some $(\mathrm{SNR}_1, \mathrm{SNR}_2, \ldots, \mathrm{SNR}_n)$. To incur it, the following conditions should

be satisfied simultaneously

$$\begin{cases} \mathrm{SNR}_2 > \mathrm{SNR}_1 + \Theta_{A3}^{1\to 2} & (c_1 \to c_2) \\ \mathrm{SNR}_3 > \mathrm{SNR}_2 + \Theta_{A3}^{2\to 3} & (c_2 \to c_3) \\ \cdots & \cdots \\ \mathrm{SNR}_n > \mathrm{SNR}_{n-1} + \Theta_{A3}^{n-1\to n} & (c_{n-1} \to c_n) \\ \mathrm{SNR}_1 > \mathrm{SNR}_n + \Theta_{A3}^{n\to 1} & (c_n \to c_1) \end{cases}$$

summing up all conditions results in $\Theta_{A3}^{1\to 2} + \Theta_{A3}^{2\to 3} + \cdots + \Theta_{A3}^{n-1\to n} + \Theta_{A3}^{n\to 1} < 0$. But since $\Theta_{A3}^{1\to 2} + \Theta_{A3}^{2\to 3} \geq 0, \Theta_{A3}^{2\to 3} + \Theta_{A3}^{3\to 4} \geq 0, \ldots, \Theta_{A3}^{n-1\to n} + \Theta_{A3}^{n\to 1} \geq 0, \Theta_{A3}^{n\to 1} + \Theta_{A3}^{1\to 2} \geq 0$, summing up them results in $2(\Theta_{A3}^{1\to 2} + \Theta_{A3}^{2\to 3} + \cdots + \Theta_{A3}^{n-1\to n} + \Theta_{A3}^{n\to 1}) \geq 0$ and thus contradiction. So we conclude that no n-cell loop will occur for any SNR settings, and conclude the sufficiency by recursion. □

A.5 Proof of Theorem 4.3.3

Proof. We prove it by contradiction. Assume $\Theta_{A3}^{i\to j} + \Theta_{A3}^{j\to i} \geq 0, \forall i, j$ but a persistent loop $c_1 \to c_2 \to \cdots \to c_n \to c_1$ happen for some $(\mathrm{SNR}_1, \mathrm{SNR}_2, \ldots, \mathrm{SNR}_n)$. Regardless of any other non-SNR policies between $c_1, c_2, \cdots, c_n$ and how they are evaluated, Equation (A.4) will still hold and result in $\Theta_{A3}^{1\to 2} + \Theta_{A3}^{2\to 3} + \cdots + \Theta_{A3}^{n-1\to n} + \Theta_{A3}^{n\to 1} < 0$. But since $\Theta_{A3}^{1\to 2} + \Theta_{A3}^{2\to 3} \geq 0, \Theta_{A3}^{2\to 3} + \Theta_{A3}^{3\to 4} \geq 0, \ldots, \Theta_{A3}^{n-1\to n} + \Theta_{A3}^{n\to 1} \geq 0, \Theta_{A3}^{n\to 1} + \Theta_{A3}^{1\to 2} \geq 0$, summing up them results in $2(\Theta_{A3}^{1\to 2} + \Theta_{A3}^{2\to 3} + \cdots + \Theta_{A3}^{n-1\to n} + \Theta_{A3}^{n\to 1}) \geq 0$ and thus contradiction. □

APPENDIX B

Proofs of Theorems in Chapter 5

For the ease of referring to notations in the proofs, we reiterate the Table 5.1 in the submission draft here.

B.1 Proofs of Theorems for Preference-Based Policy

Proof of Proposition 5.3.1

Proof. Follow the static condition assumption, neither inter-carrier policy nor intra-carrier policy changes. Therefore the decision will be deterministically the same, and loop is persistent by definition. □

Proof of Lemma 5.4.1

Proof. **(Sufficiency $\Rightarrow$)** The sequence (*) $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$ by definition is an N-carrier loop.

(Necessity $\Leftarrow$) We show in three steps that, if an inter-carrier switch sequence contains an N-carrier loop, then it contains the sequence (*) $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$. Suppose the phone initially connects to carrier C_0's RAT_0. We denote the highest preference (carrier, RAT) combination in carrier C_i as $P^i_{max} = \max_j P_{i,j}$.

Step 1. We show that under any initial condition (C_0 that phone is connected to), the device will be served by C_1 after finite switch steps. We prove it by cases. If $C_0 = C_1$, then the conclusion holds. If $C_0 \neq C_1$, then an inter-carrier switch $C_0 \mapsto C_1$ occurs. This

Table B.1: Notations

C_i	Carrier i
RAT_j	Radio access technology j (e.g., 3G, 4G)
c^k / c_i^k	Cell k (in carrier C_i)
$P_{i,j} / P_i$	Inter-carrier preference on carrier C_i's RAT_j / C_i
$p(c^i)$	Intra-carrier priority of cell c^i
$M, M(C_i)$	Measure M (on C_i) for inter-carrier policy
$Q, q(c^j)$	Measure Q (on c^j) for intra-carrier policy
δ, θ, ϕ	Different inter-carrier thresholds (on carrier)
$\Delta^i, Thresh^{i,j}$	Different intra-carrier thresholds (on c^i / c^j)

is because $P^1_{max} \geq P^0_{max} \geq P_{0,0}$ and C_1 is the carrier with the smallest index. According to Policy 1, such switch will happen. Therefore, the device will always be served by C_1 initially or after finite steps.

Step 2. The inter-carrier switch from $C_1 \mapsto C_2$ must occur, given that an N-carrier loop exists. We can prove it by contradiction: If $C_1 \mapsto C_2$ does not happen, there are two possibilities: either (a) the inter-carrier logic decides to not switch at C_1, or that (b) $C_1 \mapsto C_i, i \neq 2$ occurs. For case (a), the conditions are that (1) C_1 has the highest preference and (2) C_1 is available. Such case does not hold, because though the first condition (C_1 has the highest preference) holds by assumption, the second condition does not hold. If it is true, then there is no other possibilities nor reason to switch out from C_1, therefore the N-carrier loop will not exist. For case (b), this does not hold because $P^1_{max} \geq P^2_{max} \geq P^j_{max}, j \in [3, n]$. Therefore if a switch out from C_1 happens, it will switch to C_2 according to Policy 1, not any other carrier $C_i, i \neq 2$. Therefore, it must be the case that the inter-carrier switch $C_1 \mapsto C_2$ happens.

Step 3. Similar to the above proof, inter-carrier switches $C_2 \mapsto C_3, C_3 \mapsto C_4, \ldots, C_N \mapsto C_1$ must occur and no other switch sequences may occur, otherwise the inter-carrier switch sequence will stop at any of the carriers $C_3, C_4, \ldots, C_N$ but no N-carrier loop exists. □

Proof of Theorem 5.4.1

Proof. **(Sufficiency $\Rightarrow$)** Suppose both conditions satisfy. Without loss of generality, suppose RAT_1 is RAT_H and RAT_2 is RAT_L. This further implies: (i) for the condition (a) stated in Theorem 5.4.1, $P_{max} = P^i_{max} = P_{i,1} \geq P_{i,j}, \forall i \in [1, N], \forall j \in [3, K]$, and that RAT_2's preference is always lower than RAT_1: $P_{i,2} < P_{i,1}, \forall i \in [1, N]$. (ii) for the condition (b), the intra-carrier logic in every carrier will prefer RAT_2: as long as the phone is not connected to RAT_2, intra-carrier logic will move phone to RAT_2. We next constructively prove that the inter-carrier switch sequence (*) occurs.

Step 1. Starting from C_0 and RAT_0 initially, we show that phone will be connected to C_1 initially or in finite steps. If $C_0 = C_1$ then it is true already; otherwise, suppose $C_0 \neq C_1$ and there are two subcases: (a) if $RAT_0 \neq RAT_1$: according to Policy 1, since $P_{0,0} \leq P^0_{max} = P^1_{max} = P_{1,1}$, the inter-carrier switching will select $C_1.RAT_1$ and the phone will connect to C_1; (b) if $RAT_0 = RAT_1$: according to intra-carrier policy, phone will be reselected to RAT_2. Next, inter-carrier policy will select $C_1.RAT_1$ and switch to C_1, due to the same reason as (a).

Step 2. We show that the inter-carrier switches $C_i \mapsto C_{i+1}, \forall i \in [1, n-1]$ occur. We prove it by induction.

(Base case) First, the switch $C_1 \mapsto C_2$ will occur. After Step 1, phone is connected to C_1. Using assumption (iii), C_1's intra-carrier policy moves the phone from RAT_1 to RAT_2. However, since $P_{max} = P_{2,1} > P_{1,2}$ and that C_2 is unselected carrier, phone will switch to C_2 according to Policy 1. Moreover, C_1 will not switch to $C_3, C_4, \ldots, C_N$, because that all these carriers have larger index than C_2.

(Inductive step) Next, suppose that it is true for $k, k \in [2, n-2]$ (which means that $C_k \mapsto C_{k+1}$ occurs), we show that it is true for $k+1$. Since $C_k \mapsto C_{k+1}$ occurs, it means two things: (a) Inter-carrier logic chooses $C_k.RAT_1$ according to Policy 1; (b) $C_1, C_2, \ldots, C_k$ have been selected, while $C_{k+1}, C_{k+2}, \ldots$ have not been selected. Given assumption (ii),

intra-carrier logic at C_{k+1} moves to RAT_2. Since $P_{max} = P_{k+2,1} > P_{k+1,2}$, inter-carrier logic will perform switch. The switch target is C_{k+2}, because $C_1, \ldots, C_{k+1}$ have been selected so that C_{k+2} is the highest preference carrier which has not been selected, and has the smallest index among all possible carriers. Together, $C_i \mapsto C_{i+1}, \forall i \in [1, n-1]$ occur.

Step 3. We show that the inter-carrier switches $C_N \mapsto C_1$ occurs. Following Steps 1 and 2, we have selected all carriers with highest preference: $C_i, \forall i \in [1, n]$. Therefore, when the phone connects to C_N, all carriers are marked 'unselected' again following Policy 1. When C_N's intra-carrier logic moves phone from $C_N.RAT_1$ to $C_N.RAT_2$, an inter-carrier switch happens because $P_{max} = P_{1,1} = P_{N,1} > P_{N,2}$. Therefore, it will select $C_1.RAT_1$, since it has the highest preference and C_1 is not selected and has the smallest index.

Together, with Steps 1, 2 and 3, we prove that the sufficient condition will lead to the inter-carrier switch sequence (*) $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$. With the Lemma 5.4.1, the N-carrier loop occurs.

(Necessity $\Leftarrow$) We prove via contrapositive. The original statement is: if N-carrier loop happens, then both two conditions holds. We prove the contrapositive statement: if one of the conditions does not hold, N-carrier loop will not happen.

First, assume the condition (a) does not hold. We are proving: if some carriers have no RAT assigned with highest preference P_{max}, then no N-carrier loop may happen. It is easy to prove, because the carrier with no RAT assigned with highest preference will not get selected by inter-carrier Policy 1. Under this case, a K-carrier loop ($1 < K < N$) may happen, but not the N-carrier loop.

Second, assume the condition (b) does not hold. We are proving: if in some carriers, phone can stay in the RAT_H due to intra-carrier policy, then no N-carrier loop may happen. This is evident. When the phone stays in RAT_H, it satisfies both inter-carrier and intra-carrier preference. Hence, the inter-carrier switch will stop.

Therefore, we have prove that the necessary condition of N-carrier loop. Together, the

conditions (a) and (b) are necessary and sufficient conditions for N-carrier loop. □

Proof of Theorem 5.4.2

Proof. **(Sufficiency ⇒)** We will show that under such sufficient condition, N-carrier loop will happen. We prove it in three main steps.

Step 1. Similar to the proof in Lemma 5.4.1 and Theorem 5.4.1, under any initial condition (C_0 that phone is connected to), the device will be served by C_1 in finite steps. Therefore, we will always begin from C_1.

Step 2. We show that the inter-carrier switches $C_i \mapsto C_{i+1}, \forall i \in [1, n-1]$ occur. We prove it by induction.

(Base case) First, the switch $C_1 \mapsto C_2$ will occur. Similar to the proof of Step 2 for Lemma 5.4.1 and Theorem 5.4.1: first, C_1 will switch because C_1's intra-carrier logic will lead to an unavailable cell; second, C_1 will switch to C_2, not $C_2, C_3, \ldots, C_N$.

(Inductive step) Next, assume that it is true for $k, k \in [2, n-2]$ (which means that $C_k \mapsto C_{k+1}$ occurs), we show that it is true for $k+1$. Since $C_k \mapsto C_{k+1}$ occurs, it means two conditions: (a) C_k is unavailable, which is assumed by the sufficient condition. (b) $C_1, C_2, \ldots, C_{k-1}$ have been selected, therefore C_{k+1} is the highest preference carrier which has not been selected, and has the smallest index among all possible same preference carriers.

Therefore, given that C_{k+1} is also unavailable by intra-carrier logic, inter-carrier will perform switch. The switch target is C_{k+2}, because $C_1, C_2, \ldots, C_{k-1}, C_k$ have been selected so that C_{k+2} is the highest preference carrier which has not been selected, and has the smallest index among all possible same preference carriers. Together, it proves that $C_i \mapsto C_{i+1}, \forall i \in [1, n-1]$ occur.

Step 3. We show that the inter-carrier switches $C_N \mapsto C_1$ occurs. As C_N is unavailable assumed by the sufficient condition, it needs to perform inter-carrier switch. Following Steps 1 and 2, we have connected from all carriers $C_i, \forall i \in [1, n]$. Therefore, all carriers are

marked 'unselected' again following Policy 2. Therefore, it will select C_1, since P_1 is the highest preference and C_1 is not selected and has the smallest index.

Together, with Steps 1, 2 and 3, we prove that the sufficient condition will lead to an inter-carrier switching sequence $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$. With the Lemma 5.4.1, the N-carrier loop occurs.

(Necessity $\Leftarrow$) We prove by contrapositive. The original statement is: if N-carrier loop happens, then the necessary condition holds. Therefore we prove the contrapositive statement: if such necessary condition does not hold, N-carrier loop will not happen.

If the necessary condition does not hold, it means that at least one carrier $C_i, \exists i \in [1, n]$ will not move the device to an unavailable cell, so that the device has service in carrier C_i. Without loss of generality, i is the *first* carrier that will not move the device to an unavailable cell.

Step 1. We first show the inter-carrier switching sequence $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$ will not occur under this condition. Following the similar proof to the Lemma 5.4.1, it holds. The reason is that inter-carrier switching sequence $C_1 \mapsto \cdots \mapsto C_i, \exists i \in [1, n]$ will happen, but inter-carrier switching will stop at carrier C_i. The assumption states that C_i is available while all carriers $C_1, \ldots, C_{i-1}$ whose preference higher or equal to C_i's are unavailable. Following Policy 2, all carriers $C_1, \ldots, C_{i-1}$ would have been selected when the serving carrier is C_i. Therefore, the highest preference among unselected carriers will be $P_{i+1} \leq P_i$. Since C_i is available, Policy 2 will decide that staying in C_i ($i \leq n$), so that the inter-carrier switching sequence $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$ does not happen.

Step 2. Following the Lemma 5.4.1, since the inter-carrier switch sequence $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_N \mapsto C_1$ does not occur, no N-carrier loop will happen. □

Proof of Corollary 5.7.1

Proof. Due to the similarity of the proof to that of Theorem 5.4.1, we show a proof sketch

here.

($\Rightarrow$) Construct the sequence (*) using both conditions. Without loss of generality, suppose RAT_1 is RAT_H. Further assume that the RAT_2 (RAT_L in Theorem 5.4.1) has the highest intra-carrier priority in all carriers. Step 1, C_1 is chosen initially or after finite steps, same reasoning as in Theorem 5.4.1. Step 2, $C_i \mapsto C_{i+1}, \forall i \in [1, n-1]$ occur. Prove it by induction. The key is, intra-carrier policy always select to C_i's RAT_2 by Assumption 5.7.1 because highest priority RAT_2 is guaranteed to be selected, so $C_i \mapsto C_{i+1}$ happens following condition (a) and Policy 1. Step 3, $C_N \mapsto C_1$ occurs because all carriers are marked 'unselected' again following Policy 1. By Lemma 5.4.1, an N-carrier loop happens since the sequence (*) occurs.

($\Leftarrow$) We prove via contrapositive. First, negate condition (a): if some carriers have no RAT_H assigned with highest preference, then no N-carrier loop. It holds because such carrier does not have highest preference, and will not be selected by Policy 1. Second, negate condition (b): if at least in one carrier, most preferred RAT is the same for inter-carrier and intra-carrier policy, then no N-carrier loop. It is true because the inter-carrier policy will not further move away, hence it stops. Under both negations, a K-carrier loop ($1 < K < N$) may happen, but not the N-carrier loop. □

B.2 Proofs of Theorems for Threshold-Based Policy

Without the loss of generality, we assume $M(C_1) \geq M(C_2) \geq \cdots \geq M(C_N)$. Given the problem setting and policy, we have the following Lemma regarding switch loop.

Lemma B.2.1. *If threshold-policy incurs k-loop ($2 \leq k \leq N$), then it must be $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_k \mapsto C_1$.*

Proof of Theorem 5.5.1

Proof. Assume inter-carrier policy takes Criterion *F1* with threshold θ, we prove loop will

occur. Based on our problem setting, the threshold must be a reasonable value such that there is chance for any carrier's measure to be greater than the threshold. Therefore, consider all N carriers have measure greater than threshold, θ. Without the loss of generality, assume $M(C_1) \geq M(C_2) \geq \cdots \geq M(C_N) > \theta$. Since $M(C_i) > \theta (1 \leq i \leq N)$ is satisfied for all carriers at the same time, the phone will keep switching among those carriers. □

Proof of Theorem 5.5.2

Proof. We prove this theory for criterion *F2-F4* respectively.

F2. **(Necessary condition.)** By setting the measure of any carrier equal to the lowest measure among all its cells, we prove loop-freedom is guaranteed. Once switch $C_i \mapsto C_j$ occurs, then there must be $M(C_j) \geq \phi$. Given $M(C_j) - M^{min}(C_j) \leq \phi - \theta$, no matter which cell the intra-carrier handoff leads to, the cell's measure must be no less than $M^{min}(C_j) \geq M(C_j) + \phi - \theta \geq \theta$. As a result, as long as a carrier is selected as the switch target and the phone switches to that carrier, then the phone will not trigger any switch. Loop-freedom is achieved here.

(Sufficient condition.) Here the measure of the carrier cannot always satisfy $M(C_j) - M^{min}(C_j) \leq \phi - \theta$, then we prove the inter-carrier policy cannot be loop-free. Consider only top-k carriers have the measure M no less than threshold ϕ, i.e., $C_1, C_2, \cdots, C_k$ and $M(C_1) \geq M(C_2) \geq \cdots \geq M(C_k) \geq \phi$. In addition, each top-k carrier k has $M^{min}(C_j) < \theta$, which is possible because $M(C_j) - M^{min}(C_j) \leq \phi - \theta$ is not always guaranteed. Since the intra-carrier policy is based on a different measure Q independent of M, in any carrie $C_j (1 \leq j \leq k)$, the phone could be moved to that cell with measure less than θ. Initially, assume the phone is connected to a carrier C_1. Then, based on the inter-carrier switch mechanism and intra-carrier handoff, switch $C_i \mapsto C_{i+1} (1 \leq i < k)$ and $C_k \mapsto C_1$ would happen sequentially. By now, a switch loop is formed in static case.

F3, F4. Similar to the proof above. **Sufficient condition.** Since $M(C_j) - M^{min}(C_j) > \delta$ is possible for any carrier, we assume there are two carriers C_1, C_2 which satisfy this condition.

When the phone stays on C_1 or C_2, intra-carrier handoff will move the phone to the cell with the lowest measure less than θ. In addition, assume $M(C_1) = M(C_2)$ and other carriers are unavailable. Under this condition, loop will happen between C_1 and C_2 either *F3* or *F4* is used.

Necessary condition. Prove by contradiction. Assume $M(C_j) - M^{min}(C_j) \leq \delta$ holds for any carrier, any time. Meanwhile, k-loop $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_k \mapsto C_1$ occurs. According to Lemma B.2.1, we have $M(C_1) \geq M(C_2) \geq \cdots \geq M(C_k)$. Then prove $M(C_1) \geq M(C_2) > M^{min}(C_1) + \delta$, which leads to contradiction. □

Proof of Theorem 5.5.3

Proof. We consider *F2* and *F4* separately.

F2. Prove loop-freedom is guaranteed if all conditions in Theorem 5.5.3 are violated. We first prove that, if carrier switch $C_{j_0} \mapsto C_j$ occurs, then the phone would not switch out of C_j in static case. Given $C_i \mapsto C_j$, we get $M(C_j) \geq \phi$. After switching to C_2, the phone initially camps on the cell $c_j^{u_0}$ with the maximum measure among all cells in C_j, so we have $M(c_j^{u_0}) \geq M(C_j) \geq \phi$. Finally, the phone is stably connected to cell $c_j^{u_l}$. So there exists a cell path $c_2^{u_0} \rightarrow c_2^{u_1} \rightarrow \cdots \rightarrow c_2^{u_l}$ indicating a sequence of cells selected by intra-carrier policy, from the initial cell $c_j^{u_0}$ till the terminate $c_j^{u_l}$. Note that handoff may not happen, and l is possibly equal to 0. Given $M(c_j^{u_0}) \geq \phi$, we prove any cell in the cell path has measure no less than θ if conditions regarding *F2* in Theorem 5.5.3 are violated by C_j. Prove this by induction. The hypothesis is, for all $0 \leq k < l$, if any $c_j^{u_i} (0 \leq i \leq k)$ has $M(c_j^{u_i}) \geq \theta$ then $M(c_j^{u_{k+1}}) \geq \theta$. (a) $k = 0$. We have $M(c_j^{u_0}) \geq \phi \geq \theta$. (b) $k = 1$. If handoff $c_j^{u_0} \rightarrow c_j^{u_1}$ takes criteria of *absolute-value comparison* or *indirect comparison*, then we have $M(c_j^{u_1} > Thresh1_j^{0,1})$ or $M(c_j^{u_1} > Thresh3_j^{0,1})$. In both cases, $M(c_j^{u_1}) > \theta$. (c) $k \geq 2$. Suppose for any $0 \leq i < k$, $M(c_j^{u_i}) > \theta$ holds. If handoff $c_j^{u_{k-1}} \rightarrow c_j^{u_k}$ takes criteria of *absolute-value comparison* or *indirect comparison*, then we have $M(c_j^{u_k})$ similar to case (b). Otherwise, handoff $c_j^{u_{k-1}} \rightarrow c_j^{u_k}$ takes criteria of *direct comparison*. Analyze the following different cases

based on which handoff criteria is used by $c_j^{u_{k-2}} \to c_j^{u_{k-1}}$: (c1) It takes either criteria of *absolute-value comparison* or *indirect comparison*, then we have $M(c_j^{u_k} > Thresh1_j^{k-2,k-1} + \Delta_j^{k-1})$ or $M(c_j^{u_k} > Thresh3_j^{k-2,k-1} + \Delta_j^{k-1})$. In both cases, $M(c_j^{u_k}) > \theta$. (c2) It takes either criteria of *direct comparison*. In this case, $M(c_j^{u_k}) > M(c_j^{u_{k-2}}) + \Delta_j^{k-2} + \Delta_j^{k-1}$. Since intra-policy is assumed loop-free here, based on [Li, Sigmetrics'16] we know $\Delta_j^{k-2} + \Delta_j^{k-1} \geq 0$. So we have $M(c_j^{u_k}) > M(c_j^{u_{k-2}}) \geq \theta$. By now, we prove that every cell $c_j^{u_i}$) in the sequence has $M(c_j^{u_i}) \geq \theta$. Therefore, we have $M(c_j^{u_l}) \geq \theta$ so the phone will not switch out of carrier C_j. Since any switch will lead the phone to stay on a new carrier without any more switch, loop would not occur.

F4. We prove this by contradiction. Assume conditions in Theorem 5.5.3 are violated and there exists a k-loop. According to Lemma B.2.1, the loop is $C_1 \mapsto C_2 \mapsto \cdots \mapsto C_k \mapsto C_1$. Within carrier C_1, assume the handoff sequence is $c_1^{u_0} \to c_1^{u_1} \to \cdots \to c_1^{u_l}, l \geq 0$. $c_1^{u_0}$ is the initial cell with $m(c_1^{u_0}) \geq M(C_1)$. Moreover, if handoff happens ($l > 0$), then the last handoff must be based on the criterion of *direct comparison*. Otherwise, the phone ends up with a cell whose measure is no less than θ and it will not switch out. Next, we do case analysis on the length of handoff path. Case (a). $l = 0$. In this case, we have $M(C_2) \leq M(C_1) \leq M(c_1^{u_0})$. Then the phone will not switch out, so this case is impossible. Case (b). $l = 1$. In this case, we know handoff $c_1^{u_0} \to c_1^{u_1}$ is based on *direct* comparison. Then, $M(c_1^{u_1}) + \delta > M(c_1^{u_0}) + \Delta^{u_0} + \delta \geq M(c_1^{u_0}) \geq M(C_1) \geq M(C_2)$ shows the phone will not switch out either because the criterion is not satisfied. Case(c). $l > 1$. In the handoff sequence, assume $c_1^{u_i}$ is the first cell after which all handoffs are based on *direct comparison* criterion. Based on previous analysis, we know $i \leq l - 1$. Here, if $i = l - 1$ then handoff $c_1^{u_{l-2}} \to c_1^{u_{l-1}}$ is either based on *absolute-value comparison* or *indirect comparison*, so we have $M(c_1^{u_i}) > \theta$. In this case, we get either $M(c_1^{u_l}) > M(c_1^{u_{l-1}}) + \Delta_1^{u_{l-1}} > Thresh1_1^{u_{l-2},u_{l-1}} + \Delta_1^{u_{l-1}} \geq \theta$ or $M(c_1^{u_l}) > M(c_1^{u_{l-1}}) + \Delta_1^{u_{l-1}} > Thresh3_1^{u_{l-2},u_{l-1}} + \Delta_1^{u_{l-1}} \geq \theta$. Both indicate the phone will not switch out carrier C_1. So we only have one case left, that is $i \leq l - 2$. In this case, $M(c_1^{u_l} > M(c_1^{u_i} + \sum_{x=i}^{l-1} \Delta_1^{u_x}$. Since intra-carrier handoff is assumed loop-free here, and based

on [Li, Sigmetrics'16] we have $\sum_{x=i}^{l-1} \Delta_1^{u_x} \geq 0$. Therefore, $M(c_1^{u_l} > M(c_1^{u_i}$. Now we know either $i = 0$ or not, $M(c_1^{u_l} > M(c_1^{u_i} \geq \min\{\theta, M(c_1^{u_0})\}$. Again, the phone will not switch out. Contradiction. □

Proof of Theorem 5.5.4

Proof. Assume the switch loop is $C_1 \mapsto C_2 \mapsto \cdots C_k, k \geq 2$. First we prove $M(C_1) \geq M(C_2) \geq \cdots \geq M(C_k)$. Then, we prove C_1 satisfies the condition in theorem by contradiction. If C_1 violates the condition, then we show $C_1 \mapsto C_2$ would not happen after $C_k \mapsto C_1$. The phone switches to C_1, and initially camps on cell $c_1^{u_0}$. Based on intra-carrier cell selection policy, the initial cell $c_1^{u_0}$ has the highest measure among all cells in C_1. Then intra-carrier handoff may happen and finally move the cell to cell $c_1^{u_l}$. Next we prove $M(c_1^{u_0}) \leq M(c_1^{u_l}) + \delta$. (1) If c_1^u and $c_1^{u_0}$ are the same cell, then the condition holds. (2) Otherwise, there is a handoff sequence $c_1^{u_0} \rightarrow c_1^{u_1} \rightarrow \cdots c_1^{u_l}$. Each handoff in the sequence is based on criterion of direct comparison or indirect comparison. Then we use $\delta + \sum_{j=0}^{l-1} h(c_1^{u_j} \rightarrow c_1^{u_{j+1}}) \geq 0$ to prove $M(c_1^{u_0}) \leq M(c_1^{u_l}) + \delta$. Therefore, we have $M(C_2) \leq M(C_1) \leq M(c_1^{u_0}) \leq M(c_1^{u_l}) + \delta$. That means switch $C_1 \mapsto C_2$ will not happen because criterion *F3* is not fulfilled. Now we get contradiction. □

B.3 Proofs of Theorems For Hybrid Policy

Proof of Theorem 5.6.1

Proof. Based on Theorem 5.5.1, if criterion *F1* is used for switch $C_i \mapsto C_j$ and $C_j \mapsto C_i$ at the same time, then loop will occur. As a result, if *F1* is applied to switch between carriers with equal preference, there will be loop. Similarly, if *F1* is applied to both switch to higher preference or switch to lower preference, loop will happen too. So far, we prove combination (1) and (2) are loop-prone. Next, suppose switch to higher preference takes criterion *F1* and

switch to lower preference takes criterion *F3*. Then we show loop could occur regardless of configuration of threshold. Consider two carriers C_1 and C_2 with $P_1 > P_2$ and other carriers are unavailable at the current location. Initially the phone stays on C_2. When $M(C_1) \geq \theta$, carrier switch $C_2 \mapsto C_1$ occurs. In carrier C_1, the phone is stably connected to cell c_1^u, while cell c_1^v has the maximum measure among all local cells. When $M(C_2) > M(c_1^v) + \delta$, carrier switch $C_1 \mapsto C_2$ also occurs because $M(C_2) > M(c_1^v) + \delta \geq M(c_1^u) + \delta$. In static case, the phone will keep switching back and forth between C_1, C_2, which forms loop.

Similarly, we can prove it is also loop-prone to apply *F1* to switch to lower preference and *F3* to switch to higher preference. □

B.4 Dynamic Policy Updates

Proof of Proposition 5.9.1

Proof. We prove each type of update.

(1) Preference update. We prove for RAT-aware preference update here. RAT-oblivious preference update is a special case for this proof.

Suppose the policy update is *safe*, then the inter-carrier preference values P_{old} given a fixed intra-carrier policy *before* the update is loop-free by definition. According to Theorem 5.4.1, P_{old} and the given intra-carrier policy *must not* satisfy both conditions at the same time: (a) every carrier has one or more RATs (denoted RAT_H) assigned with equal, highest preference; and (b) each carrier's intra-carrier priority and threshold result in reselection from RAT_H to a different RAT_L.

Since updating the inter-carrier preference to P_{new} will not affect the given intra-carrier policy, condition (b) is not affected in any case. When the top-preferred RAT_H is given a higher preference, condition (a) will not be satisfied in any case, by enumeration. Therefore, after the update, two conditions still do not satisfy *at the same time*, thus the loop will not

incur by Theorem 5.4.1. It means that the loop-freedom is still ensured *after* the policy update assuming (1).

(2) Threshold update. We prove the update rule is *safe* for criterion *F2* respectively. Proof for other criteria is similar.

F2. Suppose the policy is loop-free before update. To update thresholds, we can only decrease θ or increase ϕ or do both. Denote θ', ϕ' as new values, so we have $\theta' \leq \theta, \phi' \geq \phi$. Next we prove that carrier switch which does not happen before update will not happen afterwards either. Consider the phone does not switch from C_i to C_j before. Denote c_i^u as the cell selected as the final serving cell by intra-policy. So the switch criterion is not satisfied, either the $M(c_i^u) \geq \theta$ or $M(C_j) < \phi$. Then, after threshold update, we still have $M(c_i^u) \geq \theta \geq \theta'$ or $M(C_j) < \phi \leq \phi'$. Therefore, switch $C_i \mapsto C_j$ still cannot happen. Then we know, if there exists no loop before threshold update, loop will not happen afterwards as long as the update rule is followed. □

APPENDIX C

Supporting Materials for Chapter 6

C.1 Proof of Theorem 6.2.1

Proof. We first prove (ii) by the equivalent statement: *for k, if we have* $\forall p(1 \leq p \leq P), k \neq k_p$, *then* $\vec{\mathbf{h}}_k = \mathbf{0}$. In this case, for any $p, 1 \leq p \leq P$, we have:

$$\mathcal{F}(k\Delta\tau, k_p\Delta\tau) = \sum_{k'=0}^{M-1} e^{j2\pi(k\Delta\tau - k_p\Delta\tau)k'\Delta f} = \frac{1 - e^{j2\pi(k-k_p)}}{1 - e^{j\frac{2\pi}{M}(k-k_p)}}$$

Since $k \neq k_p$, we have $\mathcal{F}(k\Delta\tau, k_p\Delta\tau) = 0$ for any p. Hence,

$$\vec{\mathbf{h}}_{k,l} = \sum_{p=1}^{P} h_p e^{-j2\pi\tau_p\nu_p} \mathcal{F}(k\Delta\tau, k_p\Delta\tau)\mathcal{G}(l\Delta\nu, \nu_p) = 0$$

Next, for any path p', we have $\forall p \neq p', k_p \neq k_{p'}$. Thus, $\forall p \neq p', \mathcal{F}(k_{p'}\Delta\tau, k_p\Delta\tau) = 0$. In addition, $\mathcal{F}(k_{p'}\Delta\tau, k_{p'}\Delta\tau) = M$. Finally, we get

$$\vec{\mathbf{h}}_{k_{p'},l} = M h_{p'} e^{-j2\pi\tau_{p'}\tau_{p'}} \mathcal{G}(l\Delta\nu, \nu_{p'})$$

□

C.2 Proof: Expressions of ι_p (Equation 6.3 and 6.4)

Proof. We compute the modulus of numbers on both side in (6.2):

$$|\vec{\mathbf{h}}_{k_p,l}| = M h_p \left| \frac{1 - e^{-j2\pi(l-\iota_p)}}{1 - e^{-j\frac{2\pi}{N}(l-\iota_p)}} \right|$$

Denote $\alpha_l = \frac{\pi}{N}(l - \iota_p)$, and thus:

$$\begin{aligned}\frac{|\vec{\mathbf{h}}_{k_p,l}|}{Mh_p} &= \left|\frac{1 - e^{-j2N\alpha_l}}{1 - e^{-j2\alpha_l}}\right| \\ &= \left|\frac{\sin N\alpha_l}{\sin \alpha_l}\right| \cdot \left|\frac{\sin N\alpha_l - j\cos N\alpha_l}{\sin \alpha_l - j\cos \alpha_l}\right| = \left|\frac{\sin N\alpha_l}{\sin \alpha_l}\right|\end{aligned}$$

N is even number as a multiple of symbol numbers in a sub-frame. Therefore, there exists:

$$\left|\frac{\vec{\mathbf{h}}_{k_p,0}}{\vec{\mathbf{h}}_{k_p,N/2}}\right| = \left|\frac{\cos\frac{\pi}{N}\iota_p}{\sin\frac{\pi}{N}\iota_p}\right| = \left|\cot\frac{\pi}{N}\iota_p\right|, \text{ and } \left|\frac{\vec{\mathbf{h}}_{k_p,0}}{\vec{\mathbf{h}}_{k_p,1}}\right| = \left|\sin\frac{\pi}{N}\cot\frac{\pi}{N}\iota_p - \cos\frac{\pi}{N}\right|$$

□

C.3 Proof of the expression of a_i (6.5)

Proof. We prove the derivation based on series representations. Note that

$$\frac{|\vec{\mathbf{h}}_{k_p,l}|}{Mh_p} = \left|\frac{\sin\pi(l - \iota_p)}{\sin\frac{\pi}{N}(l - \iota_p)}\right| = \left|\sin(\pi\iota_p)\csc\frac{\pi}{N}(l - \iota_p)\right|$$

Thus we only need to prove

$$\sum_{l=0}^{N-1}\left|\sin(\pi\iota_p)\csc\frac{\pi}{N}(l - \iota_p)\right|^2 = N^2$$

According to [Ber14], we have

$$\csc^2(z) = \sum_{r\in\mathbb{Z}}\frac{1}{(z - r\pi)^2} \quad (\text{for all } z \in \mathbb{C}\setminus\{r\pi\colon r\in\mathbb{Z}\})$$

Through the same expansion, we have

$$\begin{aligned}\sum_{l=0}^{N-1}\csc^2\frac{\pi}{N}(l - \iota_p) &= \sum_{l=0}^{N-1}\sum_{r\in\mathbb{Z}}\frac{1}{(\frac{\pi}{N}(l - \iota_p) - r\pi)^2} \\ &= N^2\sum_{r\in\mathbb{Z}}\sum_{l=0}^{N-1}\frac{1}{(-\pi\iota_p + (l - Nr)\pi)^2} \\ &= N^2\csc^2(\pi\iota_p)\end{aligned}$$

This concludes the proof when ι_p is not an integer since $\sin^2(\pi\iota_p)\csc^2(\pi\iota_p) = 1$. When ι_p is the integer, we can calculate that $\vec{\mathbf{h}}_{k_p,l} = 0$ for $l \neq \iota_p$ and $|\vec{\mathbf{h}}_{k_p,l}| = h_p MN$ if $l = \iota_p$ with similar proof shown in §C.1. □

C.4 Proof of Theorem 6.3.1

Proof. Denote the subsets selected by MinFreqNum as $\{\mathbf{S_{i_1}}, \mathbf{S_{i_2}}, \ldots, \mathbf{S_{i_t}}\}$, in the order of being chosen. Since obviously $t \leq Q$, the problem is equivalent to prove that $t \leq \left\lfloor \frac{log\left(Q - \max_{1\leq j\leq c}|\mathbf{S}_j|\right)}{log\frac{c}{c-k_m}} \right\rfloor + 2$. Define $\mathbf{R}_j$ as the set of uncovered elements after the first j subsets are chosen, i.e., $\mathbf{R}_j = \mathbf{S} \setminus \bigcup_{j'\leq j} \mathbf{S}_{i_{j'}}$. Let $\alpha_j = \mid \mathbf{R}_j \mid$. After selecting the first $j-1$ subsets, MinFreqNum aims the remaining α_{j-1} elements. Given k_m as the minimum occurrence of each element, i.e., $k_m = \min_{1\leq i\leq Q}\{k_i\}$, there must exist a subset with at least $\frac{k_m\alpha_{j-1}}{c}$ elements in R_{j-1}. Then given the greedy selection done by MinFreqNum, we must have:

$$\alpha_{j-1} - \alpha_j \geq \frac{k_m\alpha_{j-1}}{c} \Longleftrightarrow \alpha_j \leq \frac{c-k_m}{c}\alpha_{j-1}$$

By induction, we can further get

$$1 \leq \alpha_{t-1} \leq (Q-s)\left(\frac{c-k_m}{c}\right)^{t-2}$$

where $s = \max_{1\leq j\leq c}|\mathbf{S}_j|$. Thus, we have the upper bound of t when $s < Q-1$:

$$t \leq \left\lfloor \frac{log(Q-s)}{log\frac{c}{c-k_m}} \right\rfloor + 2,$$

Obviously, $t = 2$ when $s \leq Q-1$. □

REFERENCES

[3GP06] 3GPP. "TS25.331: Radio Resource Control (RRC).", 2006.

[3GP11] 3GPP. "TS36.300: E-UTRA and E-UTRAN; Overall description; Stage 2.", 2011.

[3GP12a] 3GPP. "TS25.304: User Equipment (UE) Procedures in Idle Mode and Procedures for Cell Reselection in Connected Mode.", 2012.

[3GP12b] 3GPP. "TS36.322: Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification.", Sep. 2012.

[3GP14] 3GPP. "TS36.321: Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification.", Mar. 2014.

[3GP15] 3GPP. "TS36.331: Radio Resource Control (RRC).", Mar. 2015.

[3GP17a] 3GPP. "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception.", Jul. 2017.

[3GP17b] 3GPP. "Study on new radio access technology Physical layer aspects.", 2017.

[3GP17c] 3GPP. "TS36.211: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation.", 2017.

[3GP17d] 3GPP. "TS38.322: Technical Specification Group Radio Access Network; NR; Packet Data Convergence Protocol (PDCP) specification.", Jun. 2017.

[3GP19a] 3GPP. "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception.", Oct. 2019.

[3GP19b] 3GPP. "TS36.141: Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Conformance Testing.", Oct. 2019.

[3GP19c] 3GPP. "TS36.304: Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode.", Jun. 2019.

[3GP19d] 3GPP. "TS38.104: NR; Base Station (BS) Radio Transmission and Reception.", Oct. 2019.

[3GP19e] 3GPP. "TS38.211: 5G NR; Physical channels and modulation.", Jun. 2019.

[3GP19f] 3GPP. "TS38.214: 5G NR; Physical layer procedures for data.", Jun. 2019.

[3GP19g] 3GPP. "TS38.321: 5G NR; Medium Access Control (MAC) protocol specification.", Jun. 2019.

[3GP19h] 3GPP. "TS38.331: 5G NR: Radio Resource Control (RRC).", Jun. 2019.

[3GP20] 3GPP. "TS38.300: 5G NR: Overall description; Stage-2.", Jan. 2020.

[3GP21] 3GPP. "TS37.340: NR; Multi-connectivity; Overall description; Stage-2.", 2021. V16.5.0.

[3GP22] 3GPP. "TS36.413: S1 Application Protocol (S1AP).", Apr. 2022.

[Ame19] 5G America. "Global 5G: Implications of a Transformational Technology." https://www.5gamericas.org/wp-content/uploads/2019/09/2019-5G-Americas-Rysavy-Implications-of-a-Transformational-Technology-White-Pa pdf, 2019.

[App] Apple. "Apple SIM." http://www.apple.com/ipad/apple-sim/.

[Arc15] Archclearing. "ARCH Insight: Alternative Roaming Providers spring up in China.", 2015. http://www.archclearing.com/html/News/Views/ALTERNATIVE/ALTERNATIVE.htm.

[Bak19] Bakshi, Arjun and Mao, Yifan and Srinivasan, Kannan and Parthasarathy, Srinivasan. "Fast and Efficient Cross Band Channel Prediction Using Machine Learning." In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom)*, p. 37. ACM, 2019.

[Bel63] Philip Bello. "Characterization of randomly time-variant linear channels." *IEEE transactions on Communications Systems*, **11**(4):360–393, 1963.

[Ber14] Christian Berg. *Complex analysis*. Citeseer, 2014.

[BMV10] Aruna Balasubramanian, Ratul Mahajan, and Arun Venkataramani. "Augmenting Mobile 3G Using WiFi." In *ACM MobiSys*, 2010.

[ca 22] "Demo: Carrier aggregation on Flora.", 2022. https://youtu.be/36PUAmyAPs0.

[Cab] CableFree. "RSRP and RSRQ Measurement in LTE.".

[Chi20] China's First 5G-Covered High-Speed Railway Switches On. "Caixin Global." https://www.caixinglobal.com/2020-01-13/chinas-first-5g-covered-high-speed-railway-switches-on-101503588.html, Jan 2020.

[CHO04] Nicolai Czink, Markus Herdin, Hüseyin Özcelik, and Ernst Bonek. "Number of multipath clusters in indoor MIMO propagation environments." *Electronics letters*, **40**(23):1498–1499, 2004.

[CNR15] Andrei Croitoru, Dragos Niculescu, and Costin Raiciu. "Towards Wifi Mobility without Fast Handover." In *USENIX NSDI*, 2015.

[CRR09] Mostafa Zaman Chowdhury, Won Ryu, Eunjun Rhee, and Yeong Min Jang. "Handover between macrocell and femtocell for UMTS based networks." In *IEEE ICACT 2009*, 2009.

[DKP21] Chamitha De Alwis, Anshuman Kalla, Quoc-Viet Pham, Pardeep Kumar, Kapal Dev, Won-Joo Hwang, and Madhusanka Liyanage. "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research." *IEEE Open Journal of the Communications Society*, **2**:836–886, 2021.

[DLG20] Haotian Deng, Kai Ling, Junpeng Guo, and Chunyi Peng. "Unveiling the Missed 4.5G Performance In the Wild." In *HotMobile'20*, March 2020.

[DLH20] Haotian Deng, Qianru Li, Jingqi Huang, and Chunyi Peng. "Icellspeed: Increasing cellular data speed with device-assisted cell selection." In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pp. 1–13, 2020.

[DNS11] Supratim Deb, Kanthi Nagaraj, and Vikram Srinivasan. "MOTA: Engineering an Operator Agnostic Mobile Service." In *ACM MobiCom*, pp. 133–144, 2011.

[DPF18a] Haotian Deng, Chunyi Peng, Ans Fida, Jiayi Meng, and Charlie Hu. "Mobility Support in Cellular Networks: A Measurement Study on Its Configurations and Implications." In *ACM Internet Measurement Conference*, IMC'18, 2018.

[DPF18b] Haotian Deng, Chunyi Peng, Ans Fida, Jiayi Meng, and Y Charlie Hu. "Mobility Support in Cellular Networks: A Measurement Study on Its Configurations and Implications." In *Proceedings of the Internet Measurement Conference 2018*, pp. 147–160. ACM, 2018.

[DWC13] Ning Ding, Daniel Wagner, Xiaomeng Chen, Abhinav Pathak, Y. Charlie Hu, and Andrew Rice. "Characterizing and Modeling the Impact of Wireless Signal Strength on Smartphone Battery Drain." In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '13, pp. 29–40, New York, NY, USA, 2013. ACM.

[Eng16] Engadget. "Samsung's next smartwatch comes with an e-SIM.", 2016. `https://www.engadget.com/2016/02/18/samsung-gear-s2-esim/`.

[Eri21] Ericsson. "Carrier aggregation in 5G." `https://www.ericsson.com/en/ran/carrier-aggregation`, 2021.

[Flo] "Flora - Flexible Mobile Network Platform." http://metro.cs.ucla.edu/flora.html.

[For17] Small Cell Forum. "Small cells market status report (release 10.0).", 2017. `http://scf.io/en/white_papers/Market_status_report_December_2017_Special_edition.php`.

[GMC17] Weidong Gao, Lin Ma, and Gang Chuai. "Joint optimization of component carrier selection and resource allocation in 5G carrier aggregation system." *Physical Communication*, **25**:293–297, 2017.

[Goo] Google. "Google Fi." `https://fi.google.com/`.

[Goo18] Google. "Project Fi Coverage Map.", 2018. `https://fi.google.com/coverage`.

[GPR18] Marco Giordani, Michele Polese, Arnab Roy, Douglas Castor, and Michele Zorzi. "A tutorial on beam management for 3GPP NR at mmWave frequencies." *IEEE Communications Surveys & Tutorials*, **21**(1):173–196, 2018.

[GR01] Lixin Gao and Jennifer Rexford. "Stable Internet Routing Without Global Coordination." *IEEE/ACM Trans. Netw.*, **9**(6):681–692, December 2001.

[GSM] GSMA. "The SIM for the next Generation of Connected Consumer Devices." `https://www.gsma.com/esim/`.

[GW99] Timothy G. Griffin and Gordon Wilfong. "An Analysis of BGP Convergence Properties." In *ACM SIGCOMM*, 1999.

[han22] "Demo: Handovers on Flora.", 2022. https://youtu.be/-R5dfjVLfeQ.

[HM18] Ronny Hadani and Anton Monk. "OTFS: A new generation of modulation addressing the challenges of 5G." *arXiv preprint arXiv:1802.02623*, 2018.

[HQG12] Junxian Huang, Feng Qian, Alexandre Gerber, Zhuoqing Mao, Subhabrata Sen, and Oliver Spatscheck. "A Close Examination of Performance and Power Characteristics of 4G LTE Networks." In *ACM MobiSys*, 2012.

[HRK18] Ronny Hadani, Shlomo Rakib, Shachar Kons, Michael Tsatsanis, Anton Monk, Christian Ibars, Jim Delfeld, Yoav Hebron, Andrea J Goldsmith, Andreas F Molisch, et al. "Orthogonal Time Frequency Space Modulation." *arXiv preprint arXiv:1808.00519*, 2018.

[HRT17] Ronny Hadani, Shlomo Rakib, Michail Tsatsanis, Anton Monk, Andrea J Goldsmith, Andreas F Molisch, and R Calderbank. "Orthogonal time frequency space modulation." In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6. IEEE, 2017.

[Hua16] Huawei. "Intra-RAT Mobility Management in Connected Mode Feature Parameter Description." `https://www.honorcup.ru/upload/iblock/164/6.pdf`, 2016.

[JBK15] Kiran Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. "WiDeo: Fine-grained Device-free Motion Tracing using RF Backscatter." In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI'15)*, pp. 189–204, 2015.

[JEA21] Roghayeh Joda, Medhat Elsayed, Hatem Abou-Zeid, Ramy Atawia, Akram Bin Sediq, Gary Boudreau, and Melike Erol-Kantarci. "QoS-aware joint component carrier selection and resource allocation for carrier aggregation in 5G." In *ICC 2021-IEEE International Conference on Communications*, pp. 1–6. IEEE, 2021.

[JLG14] Xin Jin, Hongqiang Harry Liu, Rohan Gandhi, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Jennifer Rexford, and Roger Wattenhofer. "Dynamic Scheduling of Network Updates." In *ACM SIGCOMM*, 2014.

[Kal10] Kaltenberger, Florian and Jiang, Haiyong and Guillaud, Maxime and Knopp, Raymond. "Relative Channel Reciprocity Calibration in MIMO/TDD Systems." In *2010 Future Network & Mobile Summit*, pp. 1–10. IEEE, 2010.

[KSR16] Parishad Karimi, Ivan Seskar, and Dipankar Raychaudhuri. "Achieving high-performance cellular data Services with multi-network access." In *Global Communications Conference (GLOBECOM), 2016 IEEE*, pp. 1–6. IEEE, 2016.

[KVB17] S. Kanugovi, S. Vasudevan, F. Baboescu, J. Zhu, S. Peng, J. Mueller, and S. Seo. "Multiple Access Management Services.", 2017. RFC Internet Draft.

[LCC14] Hong-Sheng Liao, Po-Yu Chen, and Wen-Tsuen Chen. "An efficient downlink radio resource allocation with carrier aggregation in LTE-advanced networks." *IEEE Transactions on Mobile Computing*, **13**(10):2229–2239, 2014.

[LDL16] Yuanjie Li, Haotian Deng, Jiayao Li, Chunyi Peng, and Songwu Lu. "Instability in Distributed Mobility Management: Revisiting Configuration Management in 3G/4G Mobile Networks." In *The 42nd ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'16)*, Antibes Juan-les-Pins, France, June 2016.

[LDP16] Yuanjie Li, Haotian Deng, Chunyi Peng, Zengwen Yuan, Guan-Hua Tu, Jiayao Li, and Songwu Lu. "iCellular: device-customized cellular network access on commodity smartphones." In *USENIX NSDI)*, pp. 643–656, 2016.

[LDX16] Yuanjie Li, Haotian Deng, Yuanbo Xiangli, Zengwen Yuan, Chunyi Peng, and Songwu Lu. "In-device, runtime cellular network information extraction and analysis: demo." In *ACM MobiCom*, pp. 503–504. ACM, 2016.

[LLZ20] Yuanjie Li, Qianru Li, Zhehui Zhang, Ghufran Baig, Lili Qiu, and Songwu Lu. "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*, pp. 344–358, 2020.

[LMJ98] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. "Internet Routing Instability." *IEEE/ACM Trans. Netw.*, **6**(5):515–528, October 1998.

[LMP14] Fei Liu, Petri Mahonen, and Marina Petrova. "A handover scheme towards downlink traffic load balance in heterogeneous cellular networks." In *Communications (ICC), 2014 IEEE International Conference on*, pp. 4875–4880, 2014.

[LPY16] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Jiayao Li, Haotian Deng, and Tao Wang. "MobileInsight: Extracting and Analyzing Cellular Network Information on Smartphones." In *The 22nd ACM Annual International Conference on Mobile Computing and Networking (MobiCom'16)*, New York, USA, October 2016.

[LWZ13] Hongqiang Harry Liu, Xin Wu, Ming Zhang, Lihua Yuan, Roger Wattenhofer, and David Maltz. "zUpdate: Updating Data Center Networks with Zero Loss." In *ACM SIGCOMM*, 2013.

[LXL18] Li Li, Ke Xu, Tong Li, Kai Zheng, Chunyi Peng, Dan Wang, Xiangxiang Wang, Meng Shen, and Rashid Mijumbi. "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*, pp. 161–175. ACM, 2018.

[LXP16] Yuanjie Li, Jiaqi Xu, Chunyi Peng, and Songwu Lu. "A First Look at Unstable Mobility Management in Cellular Networks." In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications*, pp. 15–20. ACM, 2016.

[LYP17] Yuanjie Li, Zengwen Yuan, and Chunyi Peng. "A Control-Plane Perspective on Reducing Data Access Latency in LTE Networks." In *ACM Mobicom*, Snowbird, Utah, USA, October 2017.

[MAT18] MATLAB. "Channel Estimation." `https://www.mathworks.com/help/lte/ug/channel-estimation.html`, 2018.

[MHT16] Anton Monk, Ronny Hadani, Michail Tsatsanis, and Shlomo Rakib. "OTFS-orthogonal time frequency space." *arXiv preprint arXiv:1608.02993*, 2016.

[NGM15] NGMN. "NGMN 5G white paper.", 2015. `https://www.ngmn.org/5g-white-paper/5g-white-paper.html`.

[Ope17] OpenSignal. "State of Mobile Networks: USA (Regional Performance).", Aug 2017. `https://opensignal.com/reports/2017/08/usa/state-of-the-mobile-network`.

[ope19] "OpenAirInterface." `https://gitlab.eurecom.fr/oai/openairinterface5g/wikis/home`, 2019.

[Ope20] OpenSignal. "Analyzing AT&T's spectrum usage to understand its 5G rollout plans." https://www.opensignal.com/2020/03/09/analysing-atts-spectrum-usage-to-understand-its-5g-rollout-plans, 2020.

[PLK20] Jae-Joon Park, Juyul Lee, Kyung-Won Kim, and Myung-Don Kim. "28-GHz high-speed train measurements and propagation characteristics analysis." In *2020 14th European Conference on antennas and propagation (EuCAP)*, pp. 1–5. IEEE, 2020.

[QWP17] Zafar Ayyub Qazi, Melvin Walls, Aurojit Panda, Vyas Sekar, Sylvia Ratnasamy, and Scott Shenker. "A high performance packet core for next generation cellular networks." In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, pp. 348–361. ACM, 2017.

[RPH18] Patchava Raviteja, Khoa T Phan, Yi Hong, and Emanuele Viterbo. "Interference cancellation and iterative detection for orthogonal time frequency space modulation." *IEEE Transactions on Wireless Communications*, **17**(10):6501–6515, 2018.

[RPH19] Patchava Raviteja, Khoa T Phan, and Yi Hong. "Embedded pilot-aided channel estimation for OTFS in delay–Doppler channels." *IEEE transactions on vehicular technology*, **68**(5):4906–4917, 2019.

[San] I. Santamaria. "Fading Channels: Capacity, BER and Diversity.".

[SDA19] Wenqian Shen, Linglong Dai, Jianping An, Pingzhi Fan, and Robert W Heath. "Channel estimation for orthogonal time frequency space (OTFS) massive MIMO." *IEEE Transactions on Signal Processing*, **67**(16):4204–4217, 2019.

[SG] R. Mumtaz S. A. Busari and J. Gonzalez. "Multi-connectivity in 5G new radio standards.".

[Sin19] Singular value decomposition (SVD). "Wkipedia." https://en.wikipedia.org/wiki/Singular_value_decomposition, 2019.

[SMR14] Peng Sun, Ratul Mahajan, Jennifer Rexford, Lihua Yuan, Ming Zhang, and Ahsan Arefin. "A Network-State Management Service." In *ACM SIGCOMM*, 2014.

[srs] "srsRAN." https://www.srslte.com/.

[Sta16] StackOverflow. "Setting WiFi priority on Android.", Feb 2016. https://stackoverflow.com/questions/35451980/setting-wifi-priority-on-android-lollipop-or-later-via-settings-ui.

[Tan14] Yanchao Tang. "The Research On LTE Coverage Solutions on High-Speed Railway." *Designing Techniques of Posts and Telecommunications*, **12**:20–23, 2014.

[TLL18] Zhaowei Tan, Yuanjie Li, Qianru Li, Zhehui Zhang, Zhehan Li, and Songwu Lu. "Enabling Mobile VR in LTE Networks: How Close Are We?" In *ACM SIGMETRICS*, 2018.

[TV05] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[use] "How Many Smartphones Are in The World?" https://www.bankmycell.com/blog/how-many-phones-are-in-the-world.

[Vas16] Vasisht, Deepak and Kumar, Swarun and Rahul, Hariharan and Katabi, Dina. "Eliminating Channel Feedback in Next-Generation Cellular Networks." In *Proceedings of the 2016 ACM SIGCOMM Conference*, pp. 398–411. ACM, 2016.

[WGA15] Cheng-Xiang Wang, Ammar Ghazal, Bo Ai, Yu Liu, and Pingzhi Fan. "Channel measurements and models for high-speed train communication systems: A survey." *IEEE communications surveys & tutorials*, **18**(2):974–987, 2015.

[Wik18] Wikipedia. "Dual SIM phone.", 2018. `https://en.wikipedia.org/wiki/Dual_SIM`.

[Wik19a] Wikipedia. "Coherence time (communications systems)." `https://en.wikipedia.org/wiki/Coherence_time_(communications_systems)`, 2019.

[Wik19b] Wikipedia. "High-speed rail in China." `https://en.wikipedia.org/wiki/High-speed_rail_in_China`, 2019.

[WZN19] Jing Wang, Yufan Zheng, Yunzhe Ni, Chenren Xu, Feng Qian, Wangyang Li, Wantong Jiang, Yihua Cheng, Zhuo Cheng, Yuanjie Li, et al. "An Active-Passive Measurement Study of TCP Performance over LTE on High-speed Rails." In *The 25th Annual International Conference on Mobile Computing and Networking*, pp. 1–16. ACM, 2019.

[WZZ17] Teng Wei, Anfu Zhou, and Xinyu Zhang. "Facilitating Robust 60GHz Network Deployment by Sensing Ambient Reflectors." In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI'17)*, pp. 213–226, 2017.

[XNM19] Shichang Xu, Ashkan Nikravesh, and Z Morley Mao. "Leveraging Context-Triggered Measurements to Characterize LTE Handover Performance." In *PAM*, 2019.

[YLL18] Zengwen Yuan, Qianru Li, Yuanjie Li, Songwu Lu, Chunyi Peng, and George Varghese. "Resolving Policy Conflicts in Multi-Carrier Cellular Access." In *The 24th ACM Annual International Conference on Mobile Computing and Networking (MobiCom'18)*, New Delhi, India, October 2018.

[zte] "ZTE Handover Description." https://tinyurl.com/o8enuz9.

[ZTE19] ZTE and China Telecom: 5G network test on a high speed train. "IEEE ComSoc Technology Blog." https://techblog.comsoc.org/2019/11/30/zte-and-china-telecom-5g-network-test-on-a-high-speed-train-uplink-enhancement-fast-verification/, Nov 2019.