Vivisecting Mobility Management in 5G Cellular Networks

Ahmad Hassan[†], Arvind Narayanan[†], Anlan Zhang[†], Wei Ye[†], Ruiyang Zhu[‡], Shuowei Jin[‡], Jason Carpenter[†], Z. Morley Mao[‡], Feng Qian[†], Zhi-Li Zhang[†]
[†]University of Minnesota – Twin Cities [‡]University of Michigan – Ann Arbor

ABSTRACT

With 5G's support for diverse radio bands and different deployment modes, *e.g.,* standalone (SA) *vs.* non-standalone (NSA), mobility management - especially the handover process - becomes far more complex. Measurement studies have shown that frequent handovers cause wild fluctuations in 5G throughput, and worst, service outages. Through a cross-country (6,200 km+) driving trip, we conduct in-depth measurements to study the current 5G mobility management practices adopted by three major U.S. carriers. Using this rich dataset, we carry out a systematic analysis to uncover the handover mechanisms employed by 5G carriers, and compare them along several dimensions such as (4G *vs.* 5G) radio technologies, radio (low-, mid- & high-)bands, and deployment (SA *vs.* NSA) modes. We further quantify the impact of mobility on application performance, power consumption, and signaling overheads. We identify key challenges facing today's NSA 5G deployments which result in unnecessary handovers and reduced coverage. Finally, we design a holistic handover prediction system Prognos and demonstrate its ability to improve QoE for two 5G applications *16K panoramic VoD* and *real-time volumetric video streaming*. We have released the artifacts of our study at https://github.com/SIGCOMM22-5GMobility/artifact.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Mobile networks**; **Network mobility**.

KEYWORDS

5G, Mobility Management, Handover, Mobility, Network Measurement, Energy, Coverage, Handover Prediction, Application Performance, Dataset

ACM Reference Format:
Ahmad Hassan[†], Arvind Narayanan[†], Anlan Zhang[†], Wei Ye[†], Ruiyang Zhu[‡], Shuowei Jin[‡], Jason Carpenter[†], Z. Morley Mao[‡], Feng Qian[†], Zhi-Li Zhang[†]. 2022. Vivisecting Mobility Management in 5G Cellular Networks. In *ACM SIGCOMM 2022 Conference (SIGCOMM '22), August 22–26, 2022, Amsterdam, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544216.3544217

Corresponding author: hassa654@cs.umn.edu.

SIGCOMM '22, August 22–26, 2022, Amsterdam, Netherlands

ACM ISBN 978-1-4503-9420-8/22/08…$15.00
https://doi.org/10.1145/3544216.3544217

1 INTRODUCTION

With 5G's support for diverse radio bands, mobility management becomes far more complex. Moreover, with generally smaller and denser cells compared to its predecessors, 5G *handovers* (HOs) between cells are more frequent. Given that 4Gand 5G are expected to co-exist, 3GPP has introduced a number of 5G *non-standalone* (NSA) deployment architectures and the 5G *standalone* (SA) mode [2]. All these further complicate the 5G HO procedure: besides *horizontal* HOs between cells within the same technology (*e.g.,* 5G-to-5G low-band, mid-band and high-band), there are also *vertical* HOs across the technologies (*e.g.,* 4G-to-5G and 5G-to-4G).

Previous studies in 4G/LTE [35, 43, 63, 66] and recently in 5G [50, 51, 53, 54, 65] have shown that frequent HOs can lead to wild fluctuations in 5G throughput, and in the worst case, complete service "outages". These impairments will translate to poor application performance in particular for low-latency applications that 5G is supposed to support, such as AR/VR, edge offloading, and vehicle-to-everything (V2X) communication. The impact will be further aggravated by improper HO configurations that are observed in 3G/4G [35, 36, 59].

Study Goal, Challenges, and Data Collection. Given the *importance* and *complexity* of 5G HOs, it is imperative to gain a thorough understanding of the current 5G HO mechanisms and practices adopted by commercial carriers. With this goal, we conduct – to our knowledge – *a first comprehensive, in-depth study of 5G mobility management.* Unlike in-lab experiments, measuring 5G HOs in the wild faces numerous challenges: How to obtain key control-plane signaling events from unrooted smartphones? How to thoroughly survey various 5G architectures (SA vs. NSA), radio bands, and carriers under limited human resources and budgets? How to orchestrate data collection tasks at different layers? How to accurately profile the HO effect on UE (user equipment) energy consumption? To overcome these challenges, we set up a measurement platform comprising of: (1) multiple 5G smartphones with access to three major 5G carriers in the U.S., (2) a custom-built software that captures mobility-related information on unrooted smartphones, (3) a professional measurement tool that collects cellular control-plane events, and (4) a physical power monitor with an external power bank for accurately profiling UE's battery drain.

Using this platform, we carry out a cross-country data collection field trip, conducting measurements along highways (5560 km+) and within several major cities (712 km+). With over 600GB+ of logs collected, we observe 47,000+ handovers in our datasets that span multiple dimensions: (1) carriers (denoted as OpX, OpY, and OpZ), (2) radio technologies (5G *vs.* 4G), (3) 5G architectures (NSA *vs.* SA), and (4) 5G bands – low-band, mid-band, mmWave (high-band). This constitutes – to our knowledge – the largest (in terms of the mileage) cross-layer driving test of commercial 5G networks.

Leveraging our unique driving dataset summarized in Table 1, we conduct a detailed analysis to obtain key insights regarding 5G HOs

Table 1: Driving Dataset Statistics.

	OpX	OpY	OpZ
# of unique cells (*i.e.,* towers)	3030	5535	3544
# of 5G-NR radio frequency bands	4	2	4
# of 4G/LTE radio frequency bands	5	9	6
City distance traveled (4 major cities)	697 km+	712 km+	652 km+
Inter-state distance traveled (freeways)	4855 km+	5560 km+	4855 km+
# of 4G/LTE handovers	7001	9500	7491
# of 5G-NSA mobility procedures	4611	11,107	6880
# of 5G-SA handovers	N/A	465	N/A
Cumulative 5G-NR (Low-band) traces	723 min	1532 min	1063 min
Cumulative 5G-NR (Mid-band) traces	15 min	1088 min	132 min
Cumulative 5G-NR (mmWave) traces	258 min	N/A	172 min
Cumulative 5G-NSA traces	996 min	2204 min	1366 min
Cumulative 5G-SA traces	N/A	416 min	N/A
Cumulative 4G/LTE traces	2412 min	1510 min	2038 min

and uncover their impacts. Our findings reveal that there indeed exist significant disparities among the HO mechanisms adopted by the major 5G carriers with considerable performance implications as detailed below.

How do 5G HOs Impact Applications? (§4) To study the impact of 5G HOs on application QoE (quality-of-experience), we consider three case studies: i) live video conferencing, ii) real-time 3D volumetric video streaming, and iii) cloud gaming. Our experiments suggest that 5G HOs adversely affect application QoE. For example, a HO event during a live video conferencing application causes the average frame loss-rate to increase by 2.24×, and the end-to-end latency increases by 2.26× (up to 14.5×). For 4K cloud gaming at 60 FPS, we observe an average 3.64× increase in dropped frames due to HOs.

Based on both our experimental results and prior studies of 3G/4G mobility [63, 66], we note that 5G HOs exert a far severe impact on application QoE than their 4G counterparts — the severity hinges on HO types, radio bands, and radio access technologies. For instance, most of today's 5G deployment is NSA that uses 4G as the control plane and 5G New Radio (5G-NR) as the high-speed data plane – referred to as *NSA-4C* thereafter. NSA-4C and 5G-NR incur separate HOs over 4G eNodeBs (eNB) and 5G gNodeBs (gNB) respectively, leading to more frequent HOs. In particular, due to the directionality and shorter range of mmWave radio, applications over mmWave 5G suffer far higher performance fluctuations compared to mid-band and low-band 5G due to mmWave HOs (between beams). On the positive side, applications employing the *dual mode* in NSA 5G, where user data can be delivered over both 4G and 5G, mitigate the negative impact of HOs, thanks to its flexible multi-radio paradigm.

What are the Key Characteristics of 5G HOs? (§5) Motivated by the above findings, we conduct an in-depth, measurement-driven investigation of 5G HOs to uncover their key characteristics. We focus on three aspects: *HO frequency, duration, and UE energy consumption.* We find that 5G HOs are indeed triggered frequently. While driving over freeways, we experience a 5G HO occurs every 0.4 km on average, compared to every 0.6 km for 4G. The HO frequency depends on the 5G architecture and band: HOs occur more frequently in NSA (every 0.4 km) compared to 5G low-band SA (every 0.9 km) due to NSA's separate HO procedures for NSA-4C and 5G-NR; 5G NSA HOs are particularly excessive in mmWave 5G (every 0.13 km) compared to mid/low-band 5G (every 0.35/0.4 km) given mmWave gNBs' much smaller coverage. In terms of HO duration, an average HO in NSA 5G takes 167 ms to complete, about 1.19× longer than a HO in 4G.

To understand why 5G HOs take a longer time, we break down a 5G HO into multiple stages. We find that the *HO preparation* stage – during which base stations make HO decisions (before executing them) – accounts for 41% of the overall HO duration in NSA 5G. Compared to 4G, NSA 5G causes on average a 48% increase in *HO preparation* stage. This increase contributes to a longer data-plane interruption time (1.4× longer than 4G). This points to the complexities of NSA 5G HOs that involve both gNBs and eNBs as the plausible culprit. Somewhat surprisingly, we also observe high preparation time in many SA 5G HOs, likely attributed to the technical immaturity of today's SA 5G that is still in its early stages of commercial deployment.

We also examine the UE energy overhead incurred by 5G HOs. This turns out to be non-trivial: a smartphone traveling at 130 km/h for 1 hour (without user data transmission or reception) can witness on average 553 5G HOs that drain 34.7 mAh energy. 4G HOs, on the other hand, only consume 3.4 mAh energy. This hints at the importance of reducing the number of HO-related signaling messages, which is found to be positively correlated with the increased energy consumption during 5G HOs.

What are 5G HOs' Implications on Carriers? (§6) Our analysis also sheds light on potential improvements on the carrier side. We highlight three key findings. First, our extensive drive test helps depict a landscape of 5G cell coverage that is closely relevant to HOs. We find that for NSA 5G, the average coverage (diameter) of a single 5G cell is 1.4 km, 0.73 km, and 0.15 km for low-band, mid-band, and mmWave, respectively. In particular, for low-band NSA 5G, although the data plane (5G-NR) operates on the low-band, its coupled control plane (NSA-4C) still uses the mid-band, which reduces the effective coverage of low-band 5G-NR since an NSA-4C HO always triggers a 5G-NR HO. Second, HOs are performed with the goal to improve the received signal strength of UE and hence its throughput. However, we find that a 5G→5G HO between two gNBs often worsens the performance, with a median bandwidth reduction of 14% after HOs. This is because NSA 5G does not support direct HOs between gNBs; the UE instead experiences a 5G→4G and then a 4G→5G HO where each HO is performed independently without accounting for the overall (5G→5G) signal strength improvement. Third, we find that for NSA HOs where the (origin or destination) gNB and eNB are co-located at the same physical tower, their duration is significantly shorter than HOs whose gNB and eNB are not co-located where the cross-tower communications incur delays. These findings not only identify new inefficiencies of NSA 5G, but also provide valuable hints on how NSA carriers can mitigate the impact of 5G HOs, such as accounting for 4G/5G antenna locations and considering the overall HO sequence when making HO decisions.

Can We Predict 5G HOs to Improve Application QoE? (§7) Last but not the least, we explore *5G HO prediction* to help applications to accommodate and mitigate the negative impact of frequent 5G HOs. For this, we develop a robust and effective 5G HO prediction framework (dubbed Prognos). It uses observed signal strength readings, UE-side measurement reports (MRs), and past HOs to predict future HOs and their types. Prognos can work with

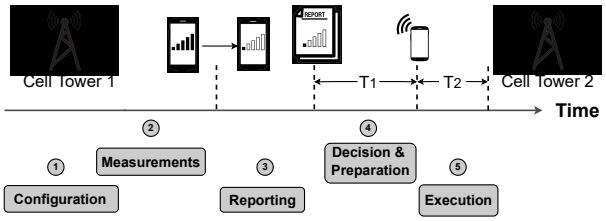


Figure 1: Logical view of handover procedure

any 3GPP-compliant 5G deployment without requiring proprietary information from the carrier. Prognos consists of a novel two-stage prediction pipeline. It first predicts the future signal strength that determines UE's MRs sent to the base station, and then learns the base station's HO logic that produces the HO decision based on the MRs. Compared to a monolithic model, decoupling the UE MR inference and network side decision logic learning reduces the model complexity and improves accuracy by eliminating indirect or unnecessary features.

We conduct extensive evaluation of Prognos using our dataset. Prognos achieves an F1-Score between 0.92–0.94 for predicting 4G/5G HOs, significantly outperforming existing HO prediction approaches developed for 4G/5G [49, 57] by 1.9×–3.8×. We incorporate Prognos into two applications, 16K panoramic video streaming and real-time volumetric video streaming, by modifying the throughput prediction algorithm used in the adaptive bitrate (ABR) adaptation modules. Prognos significantly boosts both applications' QoE compared to using the default throughput prediction algorithm: a 34.6%–58.6% reduction in stall time without degrading video quality for 16K streaming, and an 15.1%–36.2% increase in the content quality without prolonging stalls for volumetric video streaming.

Contributions. We summarize our contributions as follows: (1) creation of a large *cross-layer, multi-band, multi-carrier* dataset of 5G mobility management, (2) a first *comprehensive characterization* of mobility management in commercial 5G networks, and (3) a new methodology of *accurately predicting 5G HOs* and demonstrations of its efficacy on real-world applications over 5G.

Artifacts. To support future research, we make our dataset, source code of analysis/proposed techniques, and results publicly accessible through our project website: https://github.com/SIGCOMM22-5GMobility/artifact.

Ethics: This work does not raise any ethical issues.

2 MOBILITY MANAGEMENT TODAY

Cellular carriers dispense their services by laying out a blanket of cellular towers around an area. Cellular towers can manage multiple cells (antennas), each of which covers a geographical area. PCI (Physical Cell ID) is the identifier used for cells at the physical layer. For any mobile device, its *primary cell* is considered to be the backbone of cellular connection. It provides basic control plane signaling (*e.g.,* connection establishment, HO management, and security) along with data services to the user equipment (UE). In addition, a UE (*e.g.,* a smartphone device) can subscribe to multiple *secondary cells* for higher bandwidths. With the data flowing from a UE via a cellular tower to the 4G/5G core, mobility management procedures (*e.g.,* HOs, MRs, *etc.*) are employed to switch between cells and continuously report on the signal quality of UE.

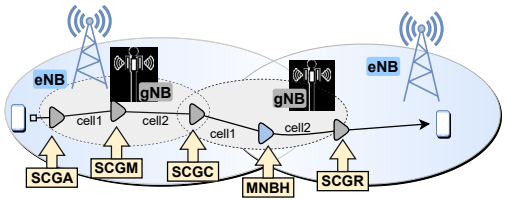


Figure 2: SCG HO procedures for mobility in NSA 5G.

HO Procedures. Fig. 1 depicts a basic HO procedure; the detailed description of all steps is in Appendix A.1. Carriers use multiple radio signal quality indicators such as Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Signal to Interference & Noise Ratio (SINR), *etc.* [8] to perform measurements based on the configurations received from the primary cell. We refer to these radio quality indicators as *RRS* (**R**SRP, **R**SRQ, **S**INR) for the rest of the paper. If any event trigger criterion is met, a measurement event is raised and its report is sent to the primary cell. The primary cell then decides a target cell based on carrier-specific HO logic and directs UE to perform HO with the target cell via an HO command (*RRC Connection Reconfiguration* [10]). Finally, the UE undergoes HO and performs link synchronization through Random Access Procedure [14].

Table 2: Handover terminology used in the paper

Procedure Type	Access Tech. Change	4G/5G HO	Acronym
SCG Addition	4G → 5G	5G	SCGA
SCG Release	5G → 4G	5G	SCGR
SCG Modification	5G → 5G	5G	SCGM
SCG Change	5G → 4G → 5G	5G	SCGC
MeNB HO	5G → 5G	4G	MNBH
MCG HO (SA)	5G → 5G	5G	MCGH
LTE HO (NSA)	5G → 5G	4G	LTEH
LTE HO (LTE)	4G → 4G	4G	LTEH

HOs in 5G: A Taxonomy. The classification of HOs has become complex in 5G; Table 2 summarizes the radio access technology change and 4G/5G HO category for each HO type used in the paper. In NSA 5G, all the cells associated with eNB constitute a master cell group (MCG). On the other hand, the group of cells linked to the gNB form a secondary cell group (SCG). A new category of HO procedures was introduced in 3GPP Release-15 [4] for SCG HO management. Fig. 2 provides an overview of SCG HO procedures used to add, modify and release 5G cells. *SCG Addition* adds 5G-NR cells to the existing LTE connection while *SCG Release* removes them. *SCG Modification* is used to switch 5G cells within the same SCG (or gNB). Unlike inter-eNB HO in LTE, NSA 5G does not have an option to perform a direct HO between two gNBs. Hence, the *SCG Change* procedure (a combination of SCG Release and Addition) is used to move the UE from one gNB to another. A master-eNB (MeNB) HO will change the LTE cell while keeping the gNB the same. In SA 5G, we only observe MCG HO that moves the UE from one 5G-NR cell to another.

3 MEASUREMENT METHODOLOGY

5G HO Measurement Tool. We extend 5G Tracker [52] to capture several key pieces of information relevant to mobility management in commercial 5G: PCIs, HOs, and radio bands. The above

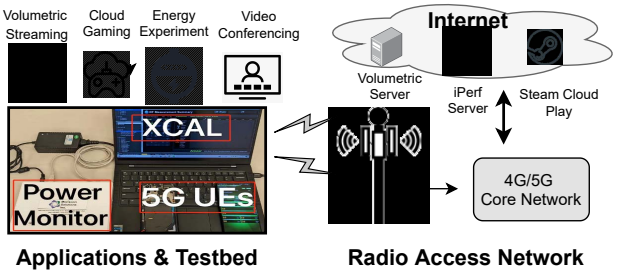


Figure 3: An overview of our measurement setup.

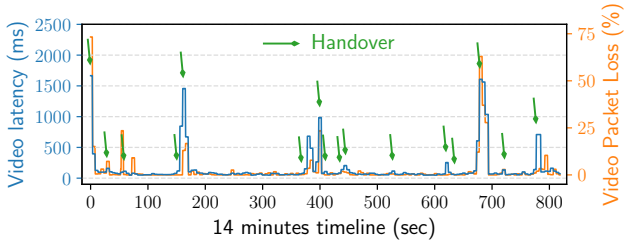


Figure 4: Video conferencing latency and packet loss during HOs in NSA 5G (Low-Band).

information is extracted from 5G-specific APIs introduced in Android 11 [16]. Regarding the last item, we use the **onDisplayInfoChanged()** API of Android TelephonyManager to identify the radio band (low-band *vs.* mmWave) of the UE. Our app also logs additional information such as UE's geolocation, radio technology (4G/LTE *vs.* 5G), ping measurements, *etc.*

5G UE and Other Measurement Tools. We use two UE models: Samsung Galaxy S21 Ultra 5G/SM-G998U (S21U) and Samsung Galaxy S20 Ultra 5G/SM-G988U (S20U). A total of four mobile phones (three S21U and one S20U) are used in our study. They are equipped with the Qualcomm Snapdragon 888 and 865 chipsets, respectively [25, 26]. The radio hardware profile of these chipsets represent the state-of-the-art, and the measurement findings hold true for other 5G smartphone models, especially Qualcomm models. To ensure a fair comparison among carriers, we place multiple smartphones side-by-side to concurrently conduct experiments and make external factors (*e.g.,* driving speed, location, *etc.*) remain consistent. Acquiring and parsing lower layer information from smartphones requires access to *Diag* (diagnostic interface), which needs special licenses and tools [23]. Therefore, we rely on a professional tool called *Accuver XCAL* [15] to read Qualcomm *Diag*. This tool runs on a laptop and can collect physical layer radio KPIs (*e.g.,* PCI, RRS values) and RRC layer signaling messages [10] (such as HO commands, event configurations, measurement reports, *etc.*). For power measurements, we use Monsoon Power Monitor [22] to power a high-end S20U smartphone. Note that all experiments except power measurements use S21U.

5G and 4G Networks. Our analysis focuses on three dimensions: *(1) 5G Carriers:* We collected data across three major U.S. 5G carriers (OpX, OpY, OpZ). *(2) Radio Access Technologies (RAT):* We compare different radio technologies (LTE *vs.* NSA 5G *vs.* SA 5G). At the time of this study, both OpX and OpZ had deployed their 5G services in NSA while OpY was in both SA and NSA modes. *(3) Radio Frequency Bands:* The bands considered in this study were dictated by how carriers rolled out their services in the areas we covered. In 5G-NR, we capture mmWave and low-band data for OpX and OpZ. For OpY, we collect data from their mid-band and low-band 5G deployments. Additionally, the 4G/LTE dataset contains low-band and mid-band ranges for all carriers.

Drive Tests. To conduct drive-tests across major cities and interstate freeways in the U.S., we tether three S21U smartphones - one for each carrier - to a laptop running XCAL via *USB3* cables (Fig. 3). As summarized in Table 1, our field trip covers a total travel distance of 6,200 km+. The city data mostly comprises of dense deployments and mmWave 5G coverage, while the inter-state data loosely represents suburban deployments and Low-Band 5G coverage. This helps us understand key mobility configurations employed by commercial 5G networks and their impacts in a large scale. Most of the data is collected while driving. For analysis where walking data is used, we mention it before discussing the results.

Profiling Applications under Mobility. In order to understand the impact of mobility on application QoE, we utilize three existing mobile applications shown in Fig. 3: (i) *real-time volumetric video streaming* leverages a state-of-the-art system (ViVo) [40], (ii) *cloud gaming* adopts three popular games cloud-powered on *Steam Remote Play* [28], and (iii) *live video conferencing* utilizes a popular application, *Zoom* [31]. The detailed experimental setup can be found in Appendix A.2. All the applications are tested with OpX (NSA Low-Band, NSA mmWave, and LTE) while driving.

UDP/TCP Experiments. Using a bulk transfer application iPerf3 [12], we study the impact of mobility on transport layer performance. We use two flavors of TCP congestion control: CUBIC [30] and BBR [29]. The iPerf server runs on an AWS EC2 instance (g4dn.2xlarge | 8vCPUs | 32GB | Ubuntu 18.04) with 3 Gbps+ network bandwidth. The server captures iPerf logs, packet traces *(pcap)* and *socket statistics* (ss) logs [21]. On the UE, we run the iPerf client (cross-complied within 5G Tracker) and collect its logs.

4 IMPACT OF MOBILITY ON APPLICATION PERFORMANCE

In this section, we use a combination of latency-sensitive and bandwidth-hungry applications to understand QoE fluctuation during mobility. We exclude SA 5G from our analysis as it is not fully mature to achieve high downlink throughput (similar to recent findings in [54]) required by applications under study.

4.1 Quantifying App QoE under Mobility

We consider the following three applications as case studies.

Live Video Conferencing. We run *Zoom* while driving around a loop in a downtown area with NSA 5G coverage. Fig. 4 shows a representative trace collected during our study. We extract a 1-second time window around the UE's HO timestamps (HOs annotated using green arrows). We find the average latency is 2.26× higher compared to no-handover periods (up to 14.5× higher in the worst case). Likewise, the average packet loss rate increases by 2.24×. Prior studies show that Zoom requires a minimum bandwidth of 0.6-0.95 Mbps for a one-on-one call as in our case [34, 47]. Low-band NSA 5G offers much higher bandwidth than what *Zoom* requires. Despite this, we show that video conferencing over today's 5G

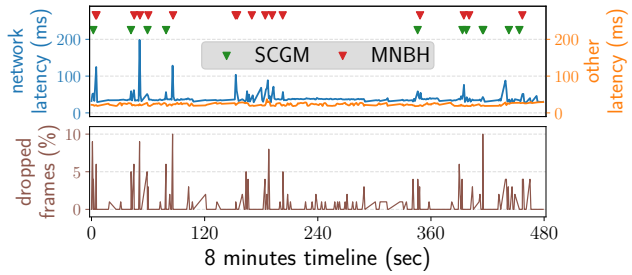


Figure 5: Cloud gaming latency and frame drop rate during HOs in NSA 5G.

remains challenging, especially during mobility as frequent HOs cause network fluctuations and increase latency impacting the QoE. Additionally, NSA 5G requires the UE be connected to both the eNB and the gNB. This causes HOs to occur on both radios. In today's 5G, the frequency of HOs are far higher than LTE (§5.1), thus the impact is amplified.

Real-time Cloud Gaming. Using a cloud-gaming application, we show the impact of HO type on QoE. We select two key metrics: (i) network (or transmission) latency, and (ii) dropped frames. The other latency (encoding, decoding, rendering, *etc.*) stays at the same level and the network latency dominates the overall latency during experiments. In our setup, the game fetches the streams at 4K@60FPS, thus in addition to being latency-sensitive, our setup also had high bandwidth requirement. As shown in Fig. 5, the network latency increases by an average 2.26× (up to 14.5×) during HOs. Likewise, HOs increase the dropped frame rate by 2.6× for a game running at 60FPS.

Considering NSA handles 4G and 5G radios at the same time, both NSA-4C (defined in §1) and 5G-NR HOs can be triggered on the UE. 5G-NR HOs in NSA 5G *e.g.,* **SCG M**odification (SCGM) have lower impact on the QoE than NSA-4C HOs, *e.g.,* **MeNB H**O (MNBH): compared to SCGM, MNBH averages 16.8ms higher network latency and a 65% increase in the number of dropped frames (see Fig. 5). Since SCGM only involves a HO between gNB cells over 5G, whereas MNBH changes the LTE primary cell (see Table 2), the QoE degradation of SCGM is relatively less than MNBH. This is also observed in volumetric video streaming experiments. Hence, we conclude that the QoE fluctuation level depends on the HO type in NSA 5G.

Volumetric Video Streaming. 5G-NR supports a wide range of radio frequencies (up to 100 GHz). The diversity of bands has a cascading impact on application performance especially under user mobility. To quantitatively capture this impact, we consider a volumetric streaming application (ViVo [40]), which is a key building block of telepresence [24], and compare HOs across two 5G-NR bands (low-band and mmWave). We focus on two key performance metrics: video bitrate and network latency. From our experiments, we note that high frequency bands usually cause more QoE degradation than low frequency bands. Fig. 6 contrasts the perceived QoE metrics between the radio frequency bands. Although low-band HOs result in a lower video quality, the degradation is significantly higher for mmWave HOs. In the median case, the video bitrate reduces by 31% for low-band HOs whereas it degrades by 58% for mmWave HOs. Similarly, the network latency increases by 41% for

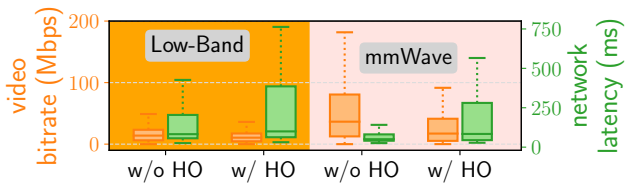


Figure 6: Impact of HOs and radio band on the QoE of volumteric video streaming.

low-band HOs while mmWave HOs see a stark 107% increase in latency. The mmWave 5G performance fluctuates wildly especially during HOs, sometimes incurring a ~2Gbps drop in throughput (see §6.2). On the other hand, the throughput degradation during HOs is comparatively lower for low-band 5G [65]. All in all, the above results suggest that the level of QoE fluctuation under mobility is determined by a combination of HO type, radio access technology, and radio frequency band.

4.2 *5G-only vs. dual* traffic mode in NSA

In NSA, 5G-NR radio resources (such as radio data bearers) are added to the ongoing 4G/LTE connection to increase data plane bandwidth for users. The user data can be exchanged on the LTE radio interface, 5G interface, or both. The NSA deployment scheme of a carrier typically decides the proportion of data arriving on each interface. A *dual mode* (MCG Split bearer [4]) splits the traffic across both 4G and 5G radio interfaces. In contrast, the *5G-only mode* employs the 5G interface for all data traffic (SCG bearer [4]). During mobility, the NSA traffic mode can differ from one area to another.

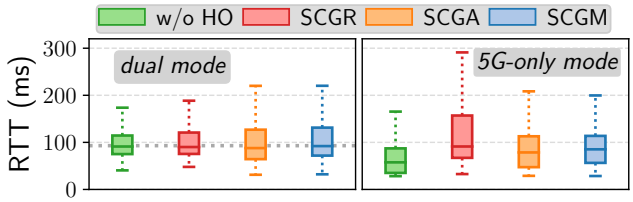


Figure 7: TCP (BBR) RTT during HOs in two NSA deployment modes.

To understand how HOs and traffic mode affect network performance, we use a simple TCP application and measure its round-trip-time (RTT). We conduct a driving experiment in areas with two different traffic modes. The traffic mode information is extracted from the PDCP layer messages [9]. There are three key takeaways from the results in Fig. 7. First, *5G-only mode* results in a comparatively lower RTT than *dual mode* when there is no HO (w/o HO case). Second, the median RTT does not change significantly during HOs in *dual mode* as 4G radio is not impacted by 5G-NR HO interruptions. This allows 4G radios to continue transmission during HOs. In the median case, we only observe a 1-4% change in RTT for *dual mode* which can be due to HO latencies [63, 65]. Finally, in *5G-only mode*, HOs have a relatively higher impact on RTT since there is no secondary interface. To be precise, the RTT can inflate by up to 37-58% in the median case. Although the results are only shown for TCP BBR, Cubic also behaves in a similar

manner. Notably, the *dual mode* absorbs HO fluctuations while the *5G-only mode* does not. However, the *dual mode* has comparatively lower performance (higher RTT) when there is no HO. In *dual mode*, the core network first sends 5G data to the eNB which is then forwarded to the gNB (before getting transmitted to the UE). Whereas, in *5G-only mode*, the 5G data is directly sent to the gNB from the core network, resulting in lower RTT compared to *dual mode*. We believe that a combination of *5G-only* and *dual modes* can get carriers the best of both worlds; they can employ *dual mode* where core network sends 5G data directly to the gNB. This can lead to a similar performance as *5G-only mode* while also minimizing HO fluctuations.

5 CHARACTERISTICS OF 5G HANDOVERS

Motivated by our findings in §4, we systematically investigate the key characteristics of handovers (HOs) in 5G using our large dataset. We focus on three key aspects that affect the UE performance: HO frequency, HO duration, and HO energy consumption by UE.

5.1 Handover Frequency

We use our drive test data to quantify the frequency of HOs across radio access technologies (4G *vs.* 5G), architectures (SA *vs.* NSA), and bands (low-band *vs.* mid-band *vs.* mmWave). Our findings suggest that compared to 4G, HOs become more frequent in NSA 5G. Specifically, in our freeway drive tests (Table 1), NSA 5G HOs are triggered every 0.4 km on average, in contrast to every 0.6 km for 4G HOs. As NSA uses 4G as control plane and 5G as data plane, both NSA-4C and 5G-NR HOs are triggered on the UE. This leads to more frequent HOs in NSA 5G when compared to 4G. On the other hand, SA 5G experiences an HO every 0.9 km. This suggests that SA realizes the performance benefits promised by 5G and reduces HO overheads [61]. For different bands within NSA, mmWave 5G sees a HO every 0.13 km, mid-band every 0.35 km, and low-band every 0.4 km. The frequency of HOs in NSA mmWave is particularly high due to the small coverage of mmWave 5G cells (§6.1). This leads to high energy inefficiency as will be measured in §5.3.

We also compare HO-related signaling overheads across all radio access technologies (LTE *vs.* NSA *vs.* SA) and bands (low-band *vs.* mmWave). Specifically, we include three message types for RRC Layer (*Measurement Reports*, *RRC Reconfiguration*, and *RRC Reconfiguration Complete* [10]). We also consider Random Access (RACH) procedure on MAC layer [5] and SSR measurements (defined in §2) on PHY layer. We find that SA 5G reduces HO-related signaling messages by a factor of ~3.8× when compared to LTE because of lower HO frequency. Additionally, HO-related signaling, especially PHY-layer procedures, increases significantly (over a 5-fold increase) in NSA mmWave compared to low-band, again due to the small mmWave cell coverage and beam management procedures [2, 53].

5.2 Handover Duration

Our application-layer study in §4 identified long 5G HOs to be a leading cause of application performance degradation during user mobility. This is also confirmed by previous studies in LTE [63, 65]. We now conduct an in-depth investigation of 5G HO duration. Overall, we find that HO duration increases significantly in NSA 5G. The average HO duration in NSA 5G is 167 ms, a 119% increase compared to 76 ms for 4G/LTE HOs. SA 5G HOs, on the other hand, are significantly shorter (110 ms) on average despite its high variation. To explain the above results, we split the HO into two stages based on the radio resource control (RRC) processes involved: (i) *preparation stage* (T_1) during which the carrier decides a new cell for HO, and (ii) *execution stage* (T_2) in which the actual HO is performed and the UE connects to a new cell.

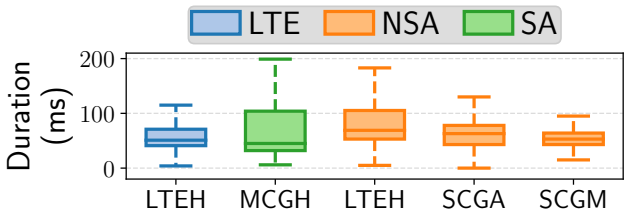


Figure 8: HO preparation stage (T_1) for OpY in NSA 5G *vs.* SA 5G *vs.* LTE.

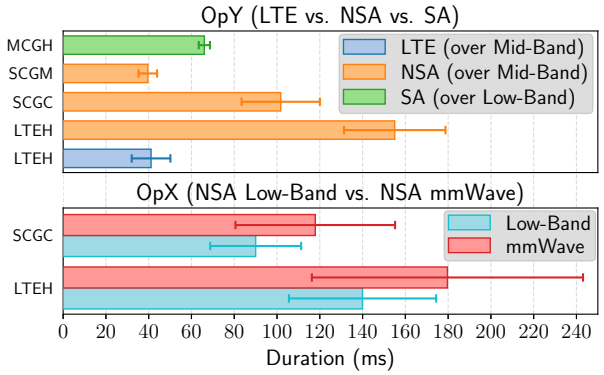


Figure 9: Comparison of HO execution stage T_2 across access technologies (NSA 5G *vs.* SA 5G *vs.* LTE) and radio bands (Low-Band *vs.* mmWave).

[T_1] HO Preparation Stage. T_1 is key when deciding and preparing a new cell for HO, and it accounts for 41% of the overall HO duration in NSA 5G. Once the primary cell is notified about a measurement event via MR, it uses the carrier-specific HO logic to decide whether to perform a HO. If yes, the source cell requests the target cell to allocate radio resources for the incoming UE [11]. As HO is performed when UE's signal strength is bad, a long T_1 duration causes the UE to stay in worse network conditions for a prolonged time. Fig. 8 shows the time consumed by OpY in T_1 stage across their deployments: LTE *vs.* NSA *vs.* SA. We clearly notice that NSA 5G takes on average 92 ms (which is almost 48%) more time than LTE. This delay in NSA 5G is very likely due to additional signaling overheads. For instance, HOs in NSA 5G involve communication between distributed identities (eNB and gNB) that may or may not be co-located [4, 60]. On the other hand, the median time spent on T_1 phase by SA 5G is comparable and to some extent slightly better than LTE. But, SA 5G still experiences large variance in the time spent on T_1. We suspect that SA 5G is still in rudimentary stages leading to high variations in HO duration. However, due to limited visibility into carrier's network, we are unable to confirm this. Later, we also explore how carriers can reduce T_1 by intelligently configuring their HO decision logic (§6.3).

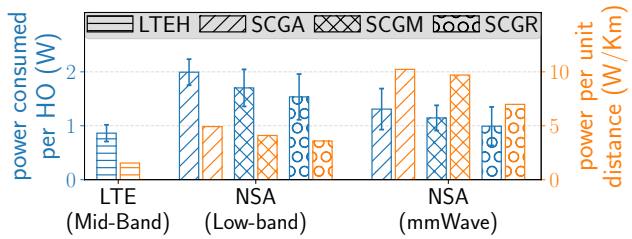


Figure 10: Comparing power consumption of HOs in Low-Band NSA 5G *vs.* mmWave NSA 5G *vs.* Mid-Band LTE.

[T_2] HO Execution Stage. Compared to T_1, T_2 has more direct impacts on upper-layer performance, and it accounts for ~59% of the overall HO duration in NSA 5G. During this phase, the HO from a source cell to a target cell is performed. Additionally, the data plane operations are completely halted[1], hence the duration spent on T_2 is critical to upper-layer application performance and user QoE. The HO ends with a successful completion of RACH procedure. Due to additional signaling overheads [10, 65], NSA 5G leads to a higher T_2 that is 1.4-5.4× compared to LTE. Within NSA 5G, mmWave band incurs 42~45 % larger T_2 time than low-band, despite the fact that the RACH procedure (part of T_2) takes less time in mmWave when compared to low-band due to shorter PRACH formats [7]. We suspect that beam management procedures involved in performing the complex beam tracking, searching, selection, *etc.* result in higher T_2 in mmWave 5G [2, 53].

Overall, the above decomposition highlights the complexities involved in 5G HOs. In particular, in NSA, the dependency of 5G on 4G's control plane results in the exchange of additional signaling messages between eNB and gNB that leads to longer HO duration.

5.3 Handover Energy Consumption

We quantify the energy overhead for NSA 5G HOs and compare our results with 4G HOs. We use 5G Tracker, *XCAL*, and Monsoon Power Monitor (MPM) introduced in §2 to conduct drive tests in areas with OpX NSA 5G (low-band and mmWave) and LTE coverage. Here, we focus on NSA HOs that bear higher HO frequency and in general smaller cell coverage compared to SA HOs.

Data Collection Methodology. To precisely calculate the energy consumption of HOs, ideally we need two pieces of data: (i) lower-layer measurement events, reports, and HOs information that can be precisely obtained from *XCAL*, and (ii) the actual power readings during HOs. We use MPM to profile the power consumption of a high-end smartphone (Samsung Galaxy S20 Ultra 5G). A practical challenge is that *XCAL* and MPM cannot be used simultaneously as the smartphone will draw current from the tethered *XCAL* laptop, making the MPM's power reading meaningless. To address this challenge, we first survey 42 km+ using *XCAL* to identify spots where a HO is triggered repeatedly by a single measurement event. Then, we drive 6 loops around identified spots with 5G Tracker and *XCAL* to establish the ground-truth of HOs. Specifically, we verify that the HO, radio technology, and band information reported by 5G Tracker's Android APIs is exactly same as *XCAL* data. Finally, we drive 10 loops with 5G Tracker (which does not require laptop tethering) and MPM to collect HO power measurements. To keep the UE in RRC connected state [10], we send a 32-byte ping packet every 5 seconds[2]. To exclude the ping transmission power, we take a +1s window starting from the time when a ping packet was transmitted and remove the corresponding measurements. We set the phone brightness to 25% for consistency and subtract baseline power from the total when presenting results. The baseline power is calculated when there is no HO and the UE is stationary. The transmission power of PING packets is also subtracted.

HO Energy Results. We calculate the battery drain for a typical smartphone using NSA 5G low-band. We find that a smartphone traveling at 130 km/h for 1 hour can witness on average 553 5G HOs. This will result in ~34.7 mAh energy drain. 4G HOs, on the other hand, only consume ~3.4 mAh energy. Similarly, NSA mmWave can experience 998 HOs and drain ~81.7 mAh energy using the same settings.

Intuition suggests that when the device is in RRC connected state [10] (transmitting or receiving data), the data-plane energy consumption overwhelms the control-plane (HO) energy, but our experiments tell a different story for commercial 5G. We compare the HO energy consumption with the data-plane energy consumption. Narayanan *et al.* [54] present the power consumption per byte for the same smartphone model as ours *i.e.,* S20U. In particular, we use the slopes of *Throughput-Power curves* presented in Table 8 of Narayanan *et al.*'s [54] work. We find that S20U using NSA low-band can download ~4.3 GB data (or upload ~2.0 GB data) with 34.7 mAh worth of battery capacity. Likewise, NSA mmWave can download ~75.4 GB data (or upload ~14.5 GB data) using 81.7 mAh energy. These results indicate the non-trivial energy footprint for 5G HOs, in particular small form-factor devices such embedded IoT devices that relatively have lesser and limited power resources.

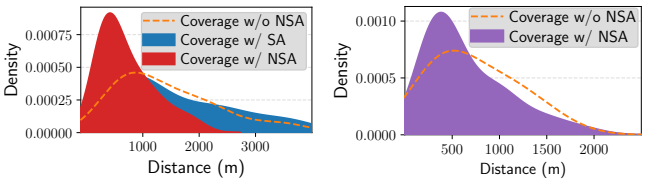
Fig. 10 provides further details of our HO energy experiments. The figure shows two metrics: (i) the power consumption of a single HO (left y-axis), and (ii) the energy consumption per unit distance (right y-axis). To compute energy per-unit distance, we take into account the frequency of HOs measured in §5.1. We separately plot the HO power consumption of 4G/LTE mid-band (left), NSA low-band (middle), and NSA mmWave (right). As shown in Fig. 10, HOs in NSA 5G consume 1.2-2.3× more energy when compared to HOs in 4G/LTE. The HO energy consumption is higher for NSA 5G HOs because both 4G and 5G radio are involved in the HO process. Surprisingly, a single mmWave HO in NSA 5G is 54% more energy efficient than a single low-band HO. This is likely because the improved RACH procedure in mmWave [7] results in lower HO energy consumption. Despite this, since HOs are highly frequent in NSA mmWave bands (§5.1), they cumulatively incur a greater energy footprint. For instance, we find that NSA mmWave HOs result in 1.9-2.4× more energy consumption per-unit distance compared to low-band HOs.

6 IMPLICATIONS OF 5G HANDOVERS ON CARRIERS

This section takes a network-side look at HOs in 5G. We: (1) present a 5G coverage landscape and highlight a coverage issue in NSA 5G, (2) discuss the impact of 5G HOs on network throughput, and

[1] In NSA, 5G HOs do not affect 4G/LTE data plane, however, 4G HOs interrupt data activity on 5G radio as well.

[2] 5 seconds is the shortest RRC tail timer [54, 65] observed in our survey.



(a) Low-Band **(b) Mid-Band**

Figure 11: Comparison of tower's effective coverage footprint (diameter): with NSA *vs.* without NSA.

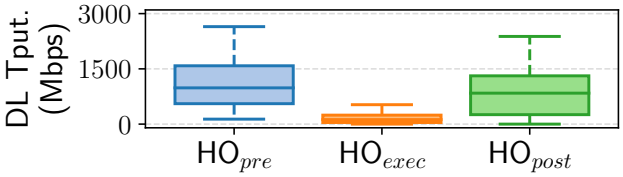


Figure 12: Impact of SCGC on network bandwidth in 5G mmWave.

(3) reveal challenges faced by NSA 5G HOs regarding the co-location of eNB and gNB.

6.1 Coverage Landscape in 5G

In cellular networks, the coverage of a cell determines when a HO will be performed. Since we did not have the tower (or cell) locations, we estimate the coverage of a cell by finding the continuous distance a UE travels while being connected to the same cell (*i.e.,* the UE does not connect to a new PCI). Essentially, the estimation calculates the average diameter of a cell. Leveraging our extensive drive test, we first present the coverage landscape in 5G. Then we discuss how the effective coverage of a 5G cell can be affected by NSA.

We find that for NSA 5G, the coverage of a single 5G cell is 1.4 km, 0.73 km, and 0.15 km for low-band, mid-band, and mmWave, respectively. Notably, coverage reduces by 48% from low-band to mid-band. The mmWave coverage is 3.9× and 8.3× lower than mid-band and low-band coverage, respectively. The signal attenuation is frequency dependent in radio networks. This means higher frequency bands are more attenuated than lower ones, thus reducing cell coverage.

Reduction of effective coverage in NSA 5G. Our study collects data under both NSA and SA deployments of 5G. A key observation we make is that the coverage of the low-band NSA cell effectively reduces as compared to low-band SA. In our dataset, this reduction is found to be between 1.2 to 2×. Fig. 11(a) shows the effective coverage for low-band NSA (red shaded area) and SA (blue shaded area). The dashed lines correspond to the hypothetical (ideal) scenario of low-band NSA coverage, assuming the UE to be in the same coverage as long as the same PCI of 5G gNB is observed. We find that UE can travel over 2000m without a HO when using low-band (n71) SA 5G. Under NSA 5G using the same n71 band, the UE on average will experience a HO within 1000m only, thus effectively reducing the coverage by half. This nullifies NSA low-band's advantage of extended coverage and infrequent HOs. To explain this, note that for low-band NSA 5G, although the data plane (5G-NR) operates on the low-band, its coupled control plane (NSA-4C) still uses the mid-band. As a result, a NSA-4C HO in NSA will always trigger 5G HO (SCGR), therefore reducing the effective coverage of a 5G cell. A similar case is found for mid-band (Fig. 11(b)) where NSA 5G's effective coverage also slightly reduces when compared to the ideal scenario where NSA-4C's impact is not considered. The above findings suggest that HOs in NSA 5G not only incur wild QoE fluctuations (§4) and long HO durations (§5.2), but also have implications on cell coverage.

6.2 Impact of 5G HOs on Bandwidth

Horizontal HOs are supposed to boost the network performance by associating a UE to a new tower with better signal strength. However, we find that oftentimes this is not the case in NSA 5G. We next describe our findings and explain the root causes.

To get insights on the impact of HOs over the network bandwidth, while walking a 35+ minute loop, we perform bulk download using iPerf3 (§2) in areas with OpX's 5G mmWave coverage. For each type of HO, we then measure the throughput in three phases: (i) Pre-HO (HO_{pre}), which captures the throughput just 1-second before the HO procedure starts, (ii) During-HO (HO_{exec}), which captures the throughput during the execution of HO procedures, and (iii) Post-HO (HO_{post}), which denotes the perceived throughput 1-second after the HO procedures are complete. Fig. 12 compares the throughput in the three phases for inter-gNB (SCGC) handovers. We observe that the average post-HO throughput reduces by 14% compared to the average pre-HO throughput. This is counter-intuitive because inter-gNB HOs are supposed to improve the received signal strength of UE and hence its throughput. While prior literature identifies one reason to be suboptimal signal strength threshold settings [65], we identify a new reason in the 5G context, as detailed next.

As explained in §2, NSA 5G does not support direct HOs between gNBs. Instead, an SCGC HO ($5G \rightarrow 5G$) comprises of $5G \rightarrow 4G$ and $4G \rightarrow 5G$ HOs, and each of the latter two HOs is performed independently without accounting for the overall (5G→5G) signal strength improvement. As a result, an SCGC HO oftentimes shows no overall signal strength improvement. To mitigate this issue, NSA carriers may need to improve their inter-gNB HO logic by considering the overall HO sequence.

Besides SCGC HOs, using the same experimental methodology descried above, we find that other types of HOs also exhibit different throughput change patterns for the above three phases. The details can be found in Appendix A.3. Such patterns can be leveraged as features for HO prediction, as to be detailed in §7.4.

6.3 Impact of eNB and gNB Co-location

In NSA, the UE connects to both eNB and gNB, which may not be co-located at the same cell tower. To identify such co-location, we find that when the NSA-4C eNB and 5G-NR gNB are co-located at the same physical tower, their 4G and 5G PCIs are the same; on the other hand, their PCIs are typically different if they are not co-located[3]. Using this heuristic, we find that the NSA-4C eNB and 5G-NR gNB are co-located only in 5%–36% of the NSA low-band samples in our dataset across the three carriers.

We find that the non-co-located NSA-4C eNB and 5G-NR gNB incur a major side effect. Specifically, we find that for NSA HOs where the (origin or destination) gNB and eNB are co-located, their

[3]To verify this, we use 4G and 5G PCIs to construct convex hulls. Using a simple algorithm [20], we identify the overlapping convex hulls for 4G and 5G PCIs.

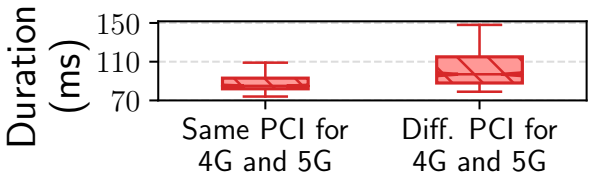


Figure 13: Handover Duration (T_1 + T_2) with same (vs. different) 4G-LTE PCI and 5G-NR PCI.

duration is significantly shorter than HOs whose gNB and eNB are not co-located. This can be clearly seen in Fig. 13 which shows that an NSA HO with same NSA-4C and 5G-NR PCI saves 13ms on average over a NSA HO with different PCIs. The additional latency is attributed to the cross-tower communication between NSA-4C and 5G-NR towers [60]. This finding suggests that NSA carriers can mitigate the impact of 5G HOs by facilitating NSA-4C and 5G-NR towers' co-location, or at least take into account for 4G/5G antenna locations when making HO decisions.

7 4G/5G HANDOVER PREDICTION

In this section, we first introduce the HO prediction problem for cellular networks (LTE/5G). We then discuss the design of our system (Prognos) along with an overview of its components. As part of the evaluation, we compare the performance of Prognos against two existing approaches. Finally, we show the advantage of utilizing HO predictions for two mobile applications: 16K panoramic video-on-demand (VoD) and real-time volumetric video streaming.

7.1 Challenges and Goals

The design of Prognos is inspired by practical mobility management concerns. The "black-box" policy-based HO logic employed by cellular carriers (*e.g.,* to choose a target cell for HO) depends on the carrier's deployment strategy for a geographical area. Moreover, the HO policies can change from one geographical region to another depending on the carrier's goal. On the other hand, we observe low temporal variation in HO policies for our collected data. In general, these insights confirm previous LTE studies [37]. Finally, the policy-based HO logic is unique for each HO type and can be formulated as a sequence of measurement reports (MRs) preceding a HO. For example, $[A2, A5, LTEH_{inter}]$ translates to an A2 MR followed by an A5 MR that eventually triggered an LTE inter-frequency HO. The trigger of MRs, in the first place, depends on mobility configurations and signal strength values of serving and neighboring cells.

We seek to overcome these challenges and build a system that can learn such carrier-specific HO policies. More specifically, our goal is to build a light-weight, scalable, context-aware, and explainable system for HO prediction. An **explainable** system can help understand the "black-box" nature of HO policies and apply sanity checks during prediction process. A **transferable** scheme can help the system scale well by enabling us to transfer models with similar geographic properties and/or carrier's deployment strategies. Any solution involving offline training will rely on the collected dataset to learn HO policies and may not generalize to the unseen mobility scenarios. A **light-weight** system avoids unnecessary overhead of real-time prediction on energy-constrained mobile devices. As the UE moves, a **reactive** system must respond to the changing radio environment. In addition to predicting HOs, a **context-aware** approach can consider factors such as radio access technology (LTE, 5G) and bands to inform applications about the possible improvement or deterioration of network conditions in future.

We realize our goal and its design principles by adopting an incremental learning scheme that extends system's knowledge as more data arrives. Compared to offline training, our approach is more adaptive. Prognos adapts to all mobility scenarios, geographic locations, and cellular carriers. The HO logic learned by Prognos sheds light on carrier-specific HO decisions. It also facilitates sanity checks during prediction, and reduction of action space. For example, an SCGM HO prediction cannot be made when a device is using LTE. Finally, Prognos outputs a meaningful value *ho_score* for applications, which specifies the expected change in network capacity due to HO. We leverage the domain knowledge of cellular networks to design a system that predicts all HO types.

7.2 Design

Prognos is a holistic system for HO prediction and provides meaningful information about network fluctuations caused by HOs. The system consists of three key components (see Fig. 17 in Appendix A.4). The *report predictor* module considers mobility configurations and signal strength qualities to predict MRs. The *decision learner* module learns the carrier-specific HO decision logic by leveraging ideas from sequential pattern mining. Finally, the *handover predictor* module uses the sequence of predicted MRs and learned HO logic to forecast the HO type.

Measurement Report Prediction. Using MRs after they have been triggered only leaves a few milliseconds – 70 ms in the median case – for the application to take any decision proactively. Therefore, *report predictor* helps predict the HOs earlier while leaving enough time for applications to minimize QoE degradation during HO. To decide if a measurement event will be triggered and reported to the serving cell, we observe three factors: (i) configurations (threshold, time-to-trigger (*TTT*) *etc.*) received from the serving cell for a measurement event, (ii) predicted RRS of serving cell, and (iii) predicted RRS of neighbor cell. To predict the RRS of serving and neighbor cells in next prediction window, the RRS values in the last history window are fed into a linear regression model, which is light-weight. A triangular kernel-based method [46] is used for signal smoothing in order to eliminate the variations caused by small scale fading and measurement noise. Based on the configurations received from the serving cell and predicted RRS, we forecast if the triggering condition[4] of an event will be satisfied in next prediction window or not. If a triggering condition is met for *TTT* amount of time, the *report predictor* module sends this prediction to the *handover predictor* module.

Policy-based Handover Logic Learning. The *decision learner* learns the up-to-date HO logic employed by the carrier. The input for *decision learner* module is a continuous stream of MRs and HO commands delivered on the *RRC* layer. We split the input stream into phases — each *phase* consists of MR(s) followed by a HO command. In Prognos, we call the learned decision logic a *pattern* which is a unique sequence of MRs repeatedly triggering a specific type of HO. The goal of the HO decision learning algorithm is to learn up-to-date

[4]The triggering condition for each measurement event is described in Table 4 and details can be found in 3GPP specifications [1].

patterns for each HO type. This sequence-based formulation of HO decision logic takes motivation from sequential pattern mining [33]. We make modifications to prefixSpan algorithm [58] making it learn patterns in an online fashion. At the end of each phase, the online learning algorithm may decide to take one of the following two actions; (i) increment the support count[5] of a pattern if an old sequence is observed or (ii) add a pattern if a new sequence is encountered. The algorithm evicts old patterns according to a freshness threshold as well. Here, freshness simply means how recent a pattern was. The eviction process also makes sure that the number of learned HO patterns do not grow excessively. Finally, the *phase count* is incremented, and we wait for a new HO to process the next *phase*.

Handover Prediction. To predict the HO, we consider the sequence of predicted MRs received so far in the current *phase*. This predicted sequence is matched against all the learned HO patterns sent by *decision learner*. If no pattern is found among the candidates, a "*no HO*" prediction is made by the *handover predictor*. Otherwise, the HO type is predicted based on the pattern which has the highest similarity. The similarity of a pattern is a function of its support count, length and freshness. Finally, based on the predicted HO type and current radio technology, Prognos generates a *ho_score* $\in (0, \infty)$. This value represents expected improvement or degradation in throughput (*e.g., ho_score*=0.4 indicates 60% degradation in throughput, while a score of 1 indicates no HO or no degradation). It is empirically calculated from results reported in Fig. 16. Specifically, we calculate the median change in network capacity using the ratio of throughput before and after HO. Most of time, *ho_score* is 1, representing "*no HO*", thus no expected change in throughput due to HO.

7.3 Performance Evaluation

We evaluate Prognos using trace-driven emulation. We collect logs from operational cellular networks using the methodology outlined previously (§3) and replay the traces.

Dataset. We collect two datasets. D1 consists of 7× traces representing a 35-min. walking loop of a tourist area. D2 is collected by walking a 25 mins loop 10× in the city's downtown area. Both datasets are collected for OpX logged @ 20 Hz. The major difference between the two is that D1 only has 5G mmWave and LTE Mid-Band coverage while D2 has 5G Low-Band coverage as well. They also represent two different U.S. cities. We observe a total of over 320 and 840 HOs in D1 and D2, respectively. The data has imbalanced classes (*i.e.,* HOs only cover 0.4% of the total data points). We therefore evaluate the performance on metrics oblivious to class imbalance such as F1-Score, precision, and recall.

Comparative Approaches. We compare Prognos with two recent 5G HO prediction techniques: 1) a *Gradient Boosting Classifier* (GBC) method used by Mei et al. [49] which uses lower layer information such as signal strength qualities of serving and neighboring cells for HO prediction and 2) a *stacked long-short-term memory* (LSTM) model [57] that predicts HOs by utilizing the location information of mobile device. Unlike these approaches, Prognos does not involve any offline training. Unless otherwise noted, we used 60% of our corpus as the training set for both these approaches; we used the remaining 40% as a test set for all prediction methods. In totality, our test set comprises of over 3.5+ hours of cellular traces. To report the results, we choose a prediction and history window of 1s for all approaches.

[5]Support count quantifies the number of times a pattern is observed.

Table 3: Performance evaluation on D1 and D2 datasets.

Dataset	Method	F1-Score	Precision	Recall	Accuracy
D1	GBC	0.475	0.403	0.577	0.936
D1	Stacked LSTM	0.284	0.190	0.562	0.857
D1	Prog. (ours)	**0.919**	**0.928**	**0.917**	0.917
D2	GBC	0.396	0.346	0.463	0.867
D2	Stacked LSTM	0.241	0.144	0.732	0.420
D2	Prog. (ours)	**0.936**	**0.946**	**0.926**	0.931

Results. As mentioned in §7.2, the *report predictor* module enables us to predict the HO before a MR has been raised. On average, it allows us to predict HOs 931 ms earlier with a slight 1.2% loss in accuracy (see Fig. 18 in Appendix A.4). Table 3 compares the performance of Prognos with other approaches on D1 and D2. Although the GBC and stacked LSTM models can achieve high accuracy sometimes, their F1-Score is low highlighting the inefficacy of "blind" machine learning techniques to produce reliable HO predictions. On the other hand, our system performs well on all metrics without any training. Our system achieves higher performance by decoupling the HO prediction task into two phases: (i) MR inference and (ii) carrier-specific HO decision logic. We find this decoupling not just helps increase our confidence in building the model but more importantly also helps improve accuracy by reducing model complexity. Additionally, our system scales well as it not only learns new HO patterns, but also removes the old (not recently observed) ones. For our datasets D1 and D2, new HO patterns are learned at a rate of 9.1 ±2.3 per hour, while old HO patterns are evicted @ 8.3 ±3.1 per hour. The eviction process makes sure that the number of learned patterns do not grow excessively, and prediction accuracy remains stable.

7.4 Prognos Use Cases

We demonstrate the usability of Prognos by considering two resource-demanding applications (*16K panoramic VoD* and *real-time volumetric video streaming*). We make minor tweaks to their rate adaptation algorithms to use HO prediction.

Trace Collection. We collect bandwidth traces by saturating the downlink channel of a mobile device while driving. We feed these traces into Mahimahi network emulation tool [55]. Concurrently, we use *XCAL* to collect cellular logs *i.e.,* RRS values, MRs and HO commands *etc.* We post-process the collected logs to generate 40+ traces (each spanning 240 seconds) using a sliding window across the data. All traces are collected for OpX and include 5G (Low-Band and mmWave) and LTE (Mid-Band) coverage. To avoid situations where quality level selection is trivial, we only consider traces with an average bandwidth less than 400 Mbps (and minimum bandwidth above 2 Mbps) following the approach used by Mao et al. [48].

Experimental Setup. For 16K panoramic VoD, our evaluation uses a custom 16K panoramic video encoded with H.264/MPEG-4 at 6 quality levels (720p, 1080p, 2K, 4K, 8K, 16K). Additionally, the video is divided into 60 chunks and has a total length of 120 seconds. We extend the setup outlined by Pensieve [48] to leverage HO prediction information delivered by Prognos. Real-time volumetric video streaming, on the other hand, makes use of ViVO system described

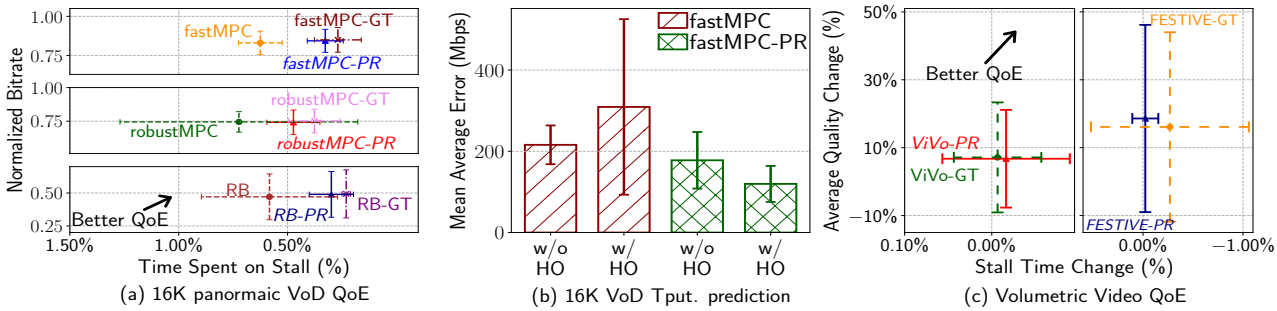


Figure 14: QoE improvement due to Prognos for *16K panoramic VoD* and *real-time volumetric video streaming*.

earlier in §3. We disable ViVo's visibility-aware optimizations for a fair comparison and modify its codebase to make it operable with our trace-driven emulation. A 3-min volumetric video compressed with Draco [19] is encoded at 5 point-cloud density levels (corresponding to bitrates in {43, 77, 110, 140, 170} Mbps).

Modified Rate Adaption Algorithm. For both applications, we correct the throughput prediction generated by their rate adaption algorithms. Specifically, we scale up or down the predicted throughput by multiplying it with the *ho_score* received from Prognos. Our system only intervenes when a HO is expected; we do not change anything in "*no HO*" situations. For evaluation, we modify 2-3 rate adaptation algorithms for each application. The same approach can be applied to any rate adaptation scheme.

Next, we demonstrate how HO-aware rate adaptation can improve the QoE of both applications. We evaluate three type of algorithms: (i) original rate adaption algorithms such as fastMPC (ii) algorithms that use ground-truth HO prediction such as fastMPC-GT, and (iii) algorithms that use HO predictions generated by Prognos (*e.g.,* fastMPC-PR). The main purpose here is to show the effectiveness of our system; we do not compare the performance of rate adaption schemes.

• **16K Panoramic VoD.** Fig. 14a and 14b compare the performance of ABR algorithms to the HO prediction-enhanced versions of rate-based (RB), fastMPC and robustMPC [48, 67]. There are three key takeaways from these results. First, the throughput prediction accuracy of the original ABR schemes degrades by an average 37.14-43.22% during HOs. Fig. 14b shows the mean average error in throughput prediction for fastMPC. Second, Prognos can improve throughput prediction during HOs by 52.42-61.29% depending on the ABR scheme (Fig. 14b). Finally, we find that our system can boost the QoE for all the ABR schemes and mobility traces. As shown in Fig. 14a, Prognos reduces stall by 34.6%-58.6% and increases the video quality by 1.72% on average. In absolute terms, the QoE is within 0.05-0.10% of the ground-truth for stall and 0.60%-0.99% for video quality.

• **Real-time Volumetric Video Streaming.** We evaluate the performance of ViVo [40] and FESTIVE [41] against the modified algorithms that use HO prediction. In Fig. 14c, we only plot the improvement brought by HO-aware (ground-truth and Prognos) rate adaptation algorithms when compared to the original rate adaptation algorithms. The improvement is shown for two metrics: video bitrate quality and stall time. Fig. 14c indicates that Prognos improves video quality by 15.1%-36.2% while also reducing stall time by 0.24%-3.67%. The QoE improvement, in absolute terms, is within 0.01%-0.25% of the ground-truth for stall time and 0.39%-2.49% for video quality.

In summary, the evaluation shows the effectiveness of our system in improving the QoE for two applications with different workloads. Additionally, we employ the same technique to improve throughput prediction for both applications.

8 RELATED WORK

Several studies [37, 45, 66] have been conducted on characterizing real-world HO configurations and provide suggestions on improving mobility management for cellular networks. However, these studies were done in the context of LTE/4G [37, 45], 3G [37, 43]. Li *et al.* [45] observe that persistent HO loops exist in operational cellular networks and provide methods to identify the persistent instability. Deng *et al.* [37] conducted large-scale HO analysis over LTE/3G/2G networks and verified the complexity and diversity of HO configurations deployed by the carriers. [43] studies the instability of mobility management in operational mobile networks by analyzing the signaling messages [44] and other low-level network information. More recently, several measurement studies [50, 51, 53, 54, 65] have been carried out to characterize 5G cellular technology. Xu *et al.* [65] focus on 5G in sub-6 GHz bands, revealing the impact of HOs on radio signal strength and TCP throughput. Narayanan *et al.* [53] perform a measurement-driven analysis of mmWave deployments for two commercial 5G carriers. Our work performs a more comprehensive study, specifically on 5G mobility management. We also take a deeper look at the implications of HOs on energy efficiency and application QoE.

Leveraging HO prediction to proactively adapt to the changing network conditions is a promising direction to explore in 5G. Several works [38, 42] developed simple HO prediction techniques for 3G and LTE systems by utilizing user's mobility pattern. Similarly, Ozturk *et al.* [57] exploited temporal correlation to do location-based HO prediction. Mei *et al.* [49] adopted a GBC approach to predict HO using lower layer information. However, all these works did not show the usability of their HO prediction schemes for real-world applications. Thus, we used two applications as case study to demonstrate the usefulness of our system.

9 DISCUSSION

Mobile systems, especially 5G, exist at the intersection of many potentially impactful variables that operate within the control of cellular carriers. Furthermore, 5G is a maturing technology and may experience major changes in architecture, structure, and capabilities

over the next few years. In this section, we discuss the limitations of our work, and the impact of our study in the context of future 5G.

Limitations of measurement scope. Our work represents a rigorous examination of 5G mobility with respect to HOs. However, there are some factors we did not explore due to scope limitations or limited visibility into the carrier's network. Regarding data-plane energy consumption, existing 5G studies investigate the observable differences by smartphone model [54, 65]. Our HO energy results compliment the existing work and our insights will hold true in general, regardless of model type. We conducted our study without any cooperation from cellular carriers. Hence, we did not explore disparity across base station vendors or manufacturers. Xie *et al.* show that the time of the day impacts user density [64], and thus the fair-share of bandwidth for each user. By experimenting at several locations (spatial diversity), and across multiple weeks (temporal diversity) and time-of-day (including night time: 12am-4am), we reduce the impact of crowds and congestion that may confound our QoE measurements.

Likewise, the impact of mobility speed and tower density on TCP performance, application QoE, and power consumption is well-explored by previous LTE studies [32, 37, 39, 63]. 5G mobility management is far more complex than LTE; HOs are more frequent and lead to higher QoE fluctuations (§4). Therefore, impact of factors such as speed and tower density intensify in 5G.

Applicability of measurement findings to future 5G and beyond. The current 5G infrastructure is still maturing, with much of the existing deployments being NSA 5G using LTE's control plane. NSA 5G deployments are here to stay at least for a few years, but will eventually be replaced by SA 5G or future NSA 5G deployments. However, as these transitions happen, future NSA 5G will also evolve such that the control plane will be 5G, with LTE acting as data plane only. For example, 5G deployment option 4 enables carriers to continue using legacy 4G equipment while connecting to the 5G core [13]. Our findings will be relevant for these new NSA 5G deployments too. Moreover, our HO prediction system (Prognos) supports all 4G and 5G HO types, and therefore, can predict HOs for SA 5G deployments as well. Additionally, multiple 5G deployment options (*e.g.,* NSA, SA, *etc.*) have been defined by the 3GPP to allow flexible (and easy) transition from 4G to 5G. In hindsight, studies like our work will help provide valuable insights in understanding the implications of adopting such transition strategies in future (*e.g.,* 5G to 6G).

Delayed HO predictions during startup. Our system learns new HO patterns in real-time. In order to make reliable predictions, it first needs to collect a few initial HO patterns. The prediction score during the startup phase is typically low. From our analysis, the time until reliable prediction depends on multiple factors including but not limited to the density of cell towers, radio capability of the mobile device, and mobility speed. For our dataset D1 and D2, we observe that the prediction F1-Score goes above 0.9 after 14 and 11 mins, respectively. However, there are ways to improve predictions during the startup phase. For instance, bootstrapping the system with the most frequent pattern for each HO type can make predictions reliable. The most frequent patterns can be found empirically from our collected dataset. Fig. 15 uses a sample trace from dataset D1 to depict the benefit of bootstrapping Prognos with the most frequent pattern. It shows that F1-Score is typically low

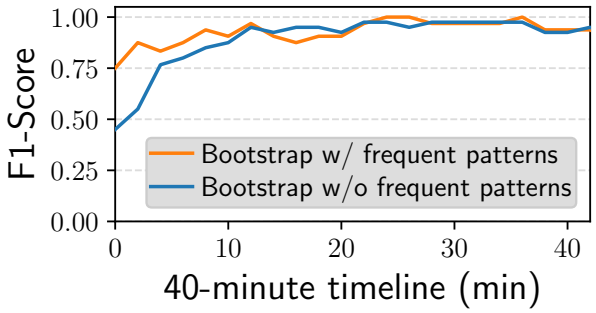


Figure 15: Impact of bootstrapping with most frequent pattern during startup phase of Prognos.

at the start if Prognos is not bootstrapped with frequent patterns. On the other hand, bootstrapping boosts the F1-Score to 0.8 within 1.5 mins. Another solution is to simply avoid making predictions during the startup phase and only learn the HO patterns for a while. Regardless, the question of how to orchestrate reliable HO predictions during the startup phase still remains open, and is left for future investigation.

The need for cross-layer communication for future 5G. Our work with Prognos rely upon information spanning several layers of the mobile network stack that is not accessible in its entirety without using special tools. Previous studies also used external tools to decode the lower layer information. Few examples are USRP-based control channel decoders [64], professional tools such as *Accuver XCAL* [15], and in-device solutions like MobileInsight [44], Mobilyzer [56], and LiveLab [62]. In future, the 5G Multi-access Edge Computing (MEC) will be able to gather and distribute control plane information through Radio Network Information (RNI) APIs [6]. We argue that exposing lower layer information through Android API calls can bring immense benefit to the mobile applications. This information can be leveraged for applications such as throughput and latency prediction, loss recovery, energy modeling, handover prediction, and more.

10 CONCLUDING REMARKS

We have carried out a first comprehensive measurement study that uncovers 5G mobility management. We conduct extensive drive tests for 6,200 km+, covering both urban and rural areas in the U.S. Our measurement findings offer deep insights into the performance, energy, cross-technology impact, upper-layer implications, and operational issues of 5G HOs. We design a holistic HO prediction system to improve application QoE during mobility. Our research also identifies key research directions on improving 5G mobility management. We have released our datasets to the research community.

ACKNOWLEDGMENTS

We thank our shepherd Kyle Jamieson and the anonymous reviewers for their suggestions and feedback. This research was in part supported by NSF under Grants CNS-1836772, CNS-1901103, CNS-1915122, CMMI-2038559, CNS-2106771, CNS-2128489, CNS-2112562, CMMI-2038215, CNS-1930041, an MnRI (Minnesota Robotics Institute) Seed Grant, and a Cisco University Research Grant.

REFERENCES

[1] 2018. 3GPP TS 38.101: NR; User Equipment (UE) radio transmission and reception; Part 3: Range 1 and Range 2 Interworking operation with other radios; Stage-3 (V15.2.0).

[2] 2018. 3GPP TS 38.912: Study on New Radio (NR) access technology (V15.0.0).

[3] 2018. *Hitman2.* Retrieved January 2022 from https://store.steampowered.com/app/863550/HITMAN_2/

[4] 2019. 3GPP TS 37.340: NR; Multi-connectivity; Overall description; Stage-2 (V15.3.0).

[5] 2019. 3GPP TS 38.321: Medium Access Control (MAC) protocol specification (V15.6.0).

[6] 2019. ETSI GS MEC 012: Multi-access Edge Computing (MEC); Radio Network Information API (V2.1.1). https://www.etsi.org/deliver/etsi_gs/MEC/001_099/012/02.01.01_60/gs_mec012v020101p.pdf

[7] 2020. 3GPP TS 38.211: Physical channels and modulation (V15.8.0).

[8] 2020. 3GPP TS 38.215: 5G NR: Physical layer measurements (V16.2.0).

[9] 2020. 3GPP TS 38.323: Packet Data Convergence Protocol (PDCP) specification (V16.2.0).

[10] 2021. 3GPP TS 36.331: Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (V16.3.0).

[11] 2021. 3GPP TS 36.423: Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP) (V16.8.0).

[12] 2021. *iperf3 – iperf3 3.9 documentation.* Retrieved January 2021 from https://software.es.net/iperf/

[13] 2022. *5G Deployment Options.* Retrieved June 2022 from https://devopedia.org/5g-deployment-options

[14] 2022. *5G(NR): Random Access Procedure.* Retrieved June 2022 from https://www.5gfundamental.com/2020/04/topic-5gnr-random-access-procedure.html

[15] 2022. *Accuver XCAL.* Retrieved January 2022 from https://www.accuver.com/sub/products/view.php?idx=6

[16] 2022. *Add 5G capabilities to your app.* Retrieved January 2022 from https://developer.android.com/about/versions/11/features/5g

[17] 2022. *Brawlhalla.* Retrieved January 2022 from https://store.steampowered.com/app/291550/Brawlhalla/

[18] 2022. *CSGO.* Retrieved January 2022 from https://store.steampowered.com/app/730/CounterStrike_Global_Offensive/

[19] 2022. *Draco 3D Data Compression.* Retrieved January 2022 from https://google.github.io/draco/

[20] 2022. *Intersection of convex polygon algorithm.* Retrieved January 2022 from https://www.swtestacademy.com/intersection-convex-polygons-algorithm/

[21] 2022. *linux socket statistics.* Retrieved January 2022 from https://man7.org/linux/man-pages/man8/ss.8.html

[22] 2022. Monsoon power monitor. https://www.msoon.com/LabEquipment/PowerMonitor/.

[23] 2022. *Qualcomm QXDM Professional(TM) Tool Quick Start.* Retrieved January 2022 from https://www.qualcomm.com/media/documents/files/qxdm-professional-qualcomm-extensible-diagnostic-monitor.pdf

[24] 2022. *Realtime 3D HOLOGRAPHIC TELEPRESENCE.* Retrieved January 2022 from https://www.omnivor.io/telepresence

[25] 2022. *Snapdragon 865 5G Mobile Platform.* Retrieved January 2022 from https://www.qualcomm.com/products/snapdragon-865-5g-mobile-platform

[26] 2022. *Snapdragon 888 5G Mobile Platform.* Retrieved January 2022 from https://www.qualcomm.com/products/snapdragon-888-5g-mobile-platform

[27] 2022. *Steam Link.* Retrieved January 2022 from https://store.steampowered.com/app/353380/Steam_Link/

[28] 2022. *Steam remote play.* Retrieved January 2022 from https://partner.steamgames.com/doc/features/remoteplay

[29] 2022. *TCP BBR.* Retrieved January 2022 from https://datatracker.ietf.org/doc/html/draft-cardwell-iccrg-bbr-congestion-control

[30] 2022. *TCP Cubic.* Retrieved January 2022 from https://datatracker.ietf.org/doc/html/rfc8312

[31] 2022. *Zoom.* Retrieved January 2022 from https://zoom.us/

[32] A. A. M. K. Abuelgasim and K. M. Yusof. 2020. High Speed Mobility Management Performance in a Real LTE Scenario. *Engineering, Technology amp; Applied Science Research* 10 (2020), 5175–5179. https://doi.org/10.48084/etasr.3245

[33] Charu C. Aggarwal and Jiawei Han. 2014. *Frequent Pattern Mining.* Springer Publishing Company, Incorporated.

[34] Hyunseok Chang, Matteo Varvello, Fang Hao, and Sarit Mukherjee. 2021. *Can You See Me Now? A Measurement Study of Zoom, Webex, and Meet.* Association for Computing Machinery, New York, NY, USA, 216–228. https://doi.org/10.1145/3487552.3487847

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

[36] Haotian Deng, Kai Ling, Junpeng Guo, and Chunyi Peng. 2020. Unveiling the Missed 4.5 G Performance In the Wild. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications.* 86–91.

[37] Haotian Deng, Chunyi Peng, Ans Fida, Jiayi Meng, and Y. Charlie Hu. 2018. Mobility Support in Cellular Networks: A Measurement Study on Its Configurations and Implications. In *Proceedings of the Internet Measurement Conference 2018.* 147–160.

[38] Huaining Ge, Xiangming Wen, Wei Zheng, Zhaoming Lu, and Bo Wang. 2009. A History-Based Handover Prediction for LTE Systems. In *2009 International Symposium on Computer Network and Multimedia Technology.* 1–4. https://doi.org/10.1109/CNMT.2009.5374706

[39] Lucas Chavarria Gimenez, Maria Carmela Cascino, Maria Stefan, Klaus I. Pedersen, and Andrea F. Cattoni. 2016. Mobility Performance in Slow- and High-Speed LTE Real Scenarios. In *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring).* https://doi.org/10.1109/VTCSpring.2016.7504347

[40] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware mobile volumetric video streaming. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking.* 1–13.

[41] Junchen Jiang, Vyas Sekar, and Hui Zhang. 2012. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies.* 97–108.

[42] Tae-Hyong Kim, Qiping Yang, Jae-Hyoung Lee, Soon-Gi Park, and Yeon-Seung Shin. 2007. A Mobility Management Technique with Simple Handover Prediction for 3G LTE Systems. In *2007 IEEE 66th Vehicular Technology Conference.* 259–263. https://doi.org/10.1109/VETECF.2007.68

[43] Yuanjie Li, Haotian Deng, Jiayao Li, Chunyi Peng, and Songwu Lu. 2016. Instability in Distributed Mobility Management: Revisiting Configuration Management in 3G/4G Mobile Networks. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science.* 261–272.

[44] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Jiayao Li, Haotian Deng, and Tao Wang. 2016. MobileInsight: Extracting and Analyzing Cellular Network Information on Smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking.* 202–215.

[45] Yuanjie Li, Jiaqi Xu, Chunyi Peng, and Songwu Lu. 2016. A first look at unstable mobility management in cellular networks. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications.* 15–20.

[46] Xiaobo Long and Biplab Sikdar. 2008. A Real-Time Algorithm for Long Range Signal Strength Prediction in Wireless Networks. In *2008 IEEE Wireless Communications and Networking Conference.* 1120–1125. https://doi.org/10.1109/WCNC.2008.202

[47] Kyle MacMillan, Tarun Mangla, James Saxon, and Nick Feamster. 2021. *Measuring the Performance and Network Utilization of Popular Video Conferencing Applications.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3487552.3487842

[48] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication.* 197–210.

[49] Lifan Mei, Jinrui Gou, Yujin Cai, Houwei Cao, and Yong Liu. 2021. Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks. *CoRR* abs/2104.12959 (2021). arXiv:2104.12959 https://arxiv.org/abs/2104.12959

[50] Arvind Narayanan, Eman Ramadan, Jason Carpenter, Qingxu Liu, Yu Liu, Feng Qian, and Zhi-Li Zhang. 2020. A first look at commercial 5G performance on smartphones. In *Proceedings of The Web Conference 2020.* 894–905.

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

[52] Arvind Narayanan, Eman Ramadan, Jacob Quant, Peiqi Ji, Feng Qian, and Zhi-Li Zhang. 2020. 5G Tracker: A Crowdsourced Platform to Enable Research Using Commercial 5g Services. In *Proceedings of the SIGCOMM '20 Poster and Demo Sessions* (Virtual event) *(SIGCOMM '20).* Association for Computing Machinery, New York, NY, USA, 65–67. https://doi.org/10.1145/3405837.3411394

[53] Arvind Narayanan, Muhammad Iqbal Rochman, Ahmad Hassan, Bariq S. Firmansyah, Vanlin Sathya, Monisha Ghosh, Feng Qian, and Zhi-Li Zhang. 2022. A Comparative Measurement Study of Commercial 5G mmWave Deployments. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications.* 800–809. https://doi.org/10.1109/INFOCOM48880.2022.9796693

[54] Arvind Narayanan, Xumiao Zhang, Ruiyang Zhu, Ahmad Hassan, Shuowei Jin, Xiao Zhu, Xiaoxuan Zhang, Denis Rybkin, Zhengxuan Yang, Zhuoqing Morley Mao, et al. 2021. A Variegated Look at 5G in the Wild: Performance, Power, and QoE Implications. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference.* 610–625.

[55] Ravi Netravali, Anirudh Sivaraman, Somak Das, Ameesh Goyal, Keith Winstein, James Mickens, and Hari Balakrishnan. 2015. Mahimahi: Accurate Record-and-Replay for HTTP. In *2015 USENIX Annual Technical Conference (USENIX ATC 15).* USENIX Association, Santa Clara, CA, 417–429. https://www.usenix.org/conference/atc15/technical-session/presentation/netravali

[56] Ashkan Nikravesh, Hongyi Yao, Shichang Xu, David Choffnes, and Z. Morley Mao. 2015. Mobilyzer: An Open Platform for Controllable Mobile Network Measurements. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services* (Florence, Italy) *(MobiSys '15)*. Association for Computing Machinery, New York, NY, USA, 389–404. https://doi.org/10.1145/2742647.2742670

[57] Metin Ozturk, Mandar Gogate, Oluwakayode Onireti, Ahsan Adeel, Amir Hussain, and Muhammad A. Imran. 2019. A novel deep learning driven, low-cost mobility prediction approach for 5G cellular networks: The case of the Control/Data Separation Architecture (CDSA). *Neurocomputing* 358 (2019), 479–489. https://doi.org/10.1016/j.neucom.2019.01.031

[58] Jian Pei, Jiawei Han, B. Mortazavi-Asl, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. 2001. PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings 17th International Conference on Data Engineering*. 215–224. https://doi.org/10.1109/ICDE.2001.914830

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

[60] Samsung. 2017. *4G-5G Interworking*. Retrieved January 2021 from https://images.samsung.com/is/content/samsung/p5/global/business/networks/insights/white-paper/4g-5g-interworking/global-networks-insight-4g-5g-interworking-0.pdf

[61] Samsung. 2021. *5G Standalone Architecture*. Retrieved January 2022 from https://images.samsung.com/is/content/samsung/assets/global/business/networks/insights/white-papers/0107_5g-standalone-architecture/5G_SA_Architecture_Technical_White_Paper_Public.pdf

[62] Clayton Shepard, Ahmad Rahmati, Chad Tossell, Lin Zhong, and Phillip Kortum. 2011. LiveLab: Measuring Wireless Networks and Smartphone Users in the Field. *SIGMETRICS Perform. Eval. Rev.* 38, 3 (jan 2011), 15–20. https://doi.org/10.1145/1925019.1925023

[63] Jing Wang, Yufan Zheng, Yunzhe Ni, Chenren Xu, Feng Qian, Wangyang Li, Wantong Jiang, Yihua Cheng, Zhuo Cheng, Yuanjie Li, et al. 2019. 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*. 1–16.

[64] Yaxiong Xie, Fan Yi, and Kyle Jamieson. 2020. PBE-CC: Congestion Control via Endpoint-Centric, Physical-Layer Bandwidth Measurements *(SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3387514.3405880

[65] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. 2020. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 479–494.

[66] Shichang Xu, Ashkan Nikravesh, and Z Morley Mao. 2019. Leveraging context-triggered measurements to characterize lte handover performance. In *International Conference on Passive and Active Network Measurement*. Springer, 3–17.

[67] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A control-theoretic approach for dynamic adaptive video streaming over HTTP. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 325–338.

A APPENDICES

Appendices are supporting material that has not been peer-reviewed.

A.1 Handover Procedure in Detail

Mobility support is considered to be a critical service for cellular networks, and HO is the most fundamental procedure that enables mobility. Fig. 1 depicts a simple handover procedure which consists of five steps. At step ①, the UE receives handover-related configuration (*e.g.*, thresholds, offsets, *etc.*) from the primary cell. In step ②, the UE starts monitoring the radio signal quality of neighboring cells based on the received configurations. Eventually, an "event" is raised when a (carrier-configured) handover criterion is met on UE. The summary of the raised event is sent to the primary cell in the form of a measurement report in step ③. Some common event types and their trigger criteria are listed in Table 4. Upon receiving the report, a handover decision is made by the primary cell in step ④. Once a target cell is selected from the set of neighboring cells, the primary cell requests the target cell to reserve radio resources for the incoming UE. We name this time period as T_1 or *HO preparation stage*. Finally, the primary cell sends a reconfiguration message to the UE via RRC (Radio Resource Control) signaling; once the handover procedure is complete, the UE responds with an RRC reconfiguration complete message and performs RACH procedure [5]. This time period of *HO execution* is T_2.

Table 4: LTE/NR Measurement Events.
($M = Measurement, P = Serving\ Cell, N = Neighboring\ Cell$)

Event Type	Event Description	Trigger Condition
A1	Serving cell becomes better than a threshold	$M_S > \Phi_{A1}$
A2	Serving cell becomes worse than a threshold	$M_P < \Phi_{A2}$
A3 (A6)	Neighboring cell becomes offset better than serving cell	$M_N > M_P + \Delta_{A3}$
A4 (B1)	Inter-RAT neighboring cell becomes better than a threshold	$M_N > \Phi_{A4}$
A5	Serving cell becomes worse than a threshold (Φ^1) and neighboring cell becomes better than another threshold (Φ^2)	$M_P > \Phi^1_{A5}$ $M_N > \Phi^2_{A5}$
P	Periodic reporting of cell conditions	N/A

A.2 Setup for Application QoE Experiments

In the following, we provide the detailed setup for our application QoE experiments.

• **Real-time volumetric video streaming** experiments leverage a state-of-the-art system (ViVo) [40]. To serve volumetric videos, we set up a university-hosted server (Intel Xeon | 8vCPUs | 64GB | Ubuntu 18.04) with 1Gbps+ network bandwidth. The volumetric video is encoded at 30 FPS using 5 different bitrate levels (43-170 Mbps). The ViVo client runs on an Android phone (S21) which is tethered to the *XCAL* laptop. While driving, we also replay user viewport traces collected by ViVo. We also modify ViVo to collect per-frame QoE logs on the mobile device.

• **Cloud gaming** is envisioned to be a killer app in 5G. We adopted several popular latency-sensitive games: Brawlhalla, CSGO, Hitman2 [3, 17, 18]. These games are cloud-powered on *Steam Remote Play* [28], a popular gaming platform. We spin up an AWS EC2 VM (g4dn.2xlarge | 8vCPUs | 32GB | NVIDIA T4 GPU | Windows10 | 25Gbps) to host and render the game. The user plays through *Steam Link* app [27] under 4K@60FPS settings. Although S21U only supports up to 2K resolution, the 4K frames are fetched from the cloud server and downscaled to 2K during rendering. The performance and streaming information is collected by *Steam's* built-in logging system.

• For **live video conferencing**, we run a popular application, *Zoom* [31], on a smartphone tethered to the *XCAL* laptop. Following the approach used by recent studies [47], we conduct a one-on-one video call from a stationary laptop to the mobile UE. We collect *Zoom* video latency and packet loss statistics during the experiment.

A.3 Impact of 5G HOs on Network Bandwidth

In Section 6.2, we only discussed the impact of SCG Change (SCGC) HO on network bandwidth. Here, we illustrate the bandwidth fluctuation of all HO procedures, and also depict the measurement event(s) that trigger each HO.

Fig. 16 clearly illustrates that a successful completion of SCG Addition can, increase the throughput by ~17× on average. SCG Release, on the other hand, reduces the throughput by 7× after the

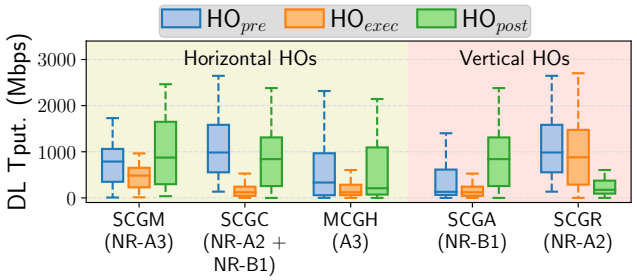


Figure 16: Impact of 5G HOs on network bandwidth in mmWave NSA 5G.

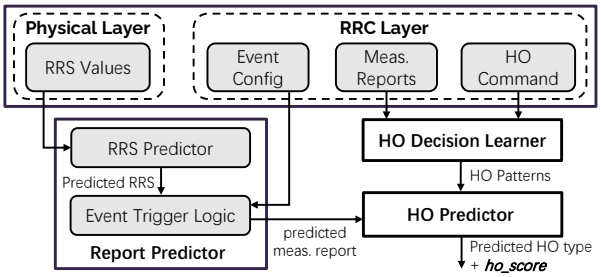


Figure 17: Design of HO prediction system Prognos.

HO. We also quantify the impact of intra-gNB HO (SCGM), inter-gNB HO (SCGC), and LTE HO (LTEH) on 5G mmWave's throughput in the shaded region highlighted as "horizontal handovers". As shown, the throughput can reduce 1.5× to 4.8× on average during all these horizontal HOs. Moreover, SCGM, on average, results in 43% throughput increase after HO, and LTE HO shows a slight 4% throughput decrease in average case.

A.4 Supporting Material for HO Prediction

In this section, we provide additional figures and results for our HO prediction system Prognos.

Figure 17 illustrates the design of Prognos. It shows the basic components (*report predictor*, *decision learner*, and *handover predictor*) of our system described in §7.2.

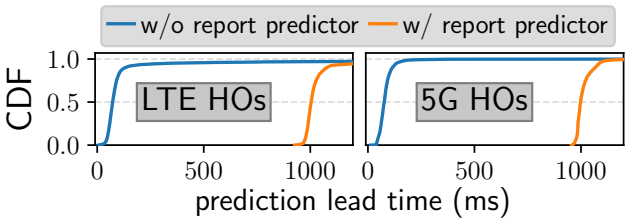


Figure 18: Lead time improvement in HO prediction due to *report predictor* approach.

Figure 18 represents how early we can predict LTE and 5G HOs if *report predictor* is used in conjunction with *decision learner*.