Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption

Dongzhu Xu[+], Anfu Zhou[+], Xinyu Zhang[◇], Guixian Wang[+], Xi Liu[+]
Congkai An[+], Yiming Shi[+], Liang Liu[+], Huadong Ma[+]
[+]Beijing University of Posts and Telecommunications
{xdz9601, zhouanfu, wangguixian, 2016213522, ACK, syming111, liangliu, mhd}@bupt.edu.cn
[◇] University of California San Diego xyzhang@ucsd.edu

ABSTRACT

5G, as a monumental shift in cellular communication technology, holds tremendous potential for spurring innovations across many vertical industries, with its promised multi-Gbps speed, sub-10 ms low latency, and massive connectivity. On the other hand, as 5G has been deployed for only a few months, it is unclear how well and whether 5G can eventually meet its prospects. In this paper, we demystify operational 5G networks through a first-of-its-kind cross-layer measurement study. Our measurement focuses on four major perspectives: *(i)* Physical layer signal quality, coverage and hand-off performance; *(ii)* End-to-end throughput and latency; *(iii)* Quality of experience of 5G's niche applications (*e.g.*, 4K/5.7K panoramic video telephony); *(iv)* Energy consumption on smartphones. The results reveal that the 5G link itself can approach Gbps throughput, but legacy TCP leads to surprisingly low capacity utilization (<32%), latency remains too high to support tactile applications and power consumption escalates to $2-3\times$ over 4G. Our analysis suggests that the wireline paths, upper-layer protocols, computing and radio hardware architecture need to co-evolve with 5G to form an ecosystem, in order to fully unleash its potential.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Network performance analysis**;

KEYWORDS

5G, Network Measurement, Network Coverage, End-to-end Performance, TCP, Energy Efficiency

ACM Reference Format:
Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, Huadong Ma. 2020. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In *Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication (SIGCOMM '20), August 10–14, 2020, Virtual Event, NY, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3387514.3405882

SIGCOMM '20, August 10–14, 2020, Virtual Event, NY, USA

ACM ISBN 978-1-4503-7955-7/20/08…$15.00
https://doi.org/10.1145/3387514.3405882

1 INTRODUCTION

We are standing on the eve of the 5G era. Major US cellular operators such as Verizon and AT&T already rolled out their first 5G deployment in 2019. Meanwhile, China's three major mobile service providers officially launched commercial 5G services, and have deployed more than 150 thousand 5G base stations by the end of 2019 [77]. 13.7 million of 5G-enabled smartphones have been sold within less than a year [94]. It is widely reported that 5G represents a giant leap beyond 4G. It is expected to attain multi-Gbps wireless bit-rate for bandwidth-hungry applications like 4K/8K UHD video/VR transmission, ultra reliable and low latency communication (uRLLC) for auto-driving or telesurgery [92] and also the massive machine to machine communication for IoT [66, 87]. Overall, 5G is predicted to generate new economic revenue up to $12.3 trillion across a broad range of industries [73].

Despite the huge potential, one should be cautious that it takes time for 5G to evolve and mature. The most recent 5G standard (3GPP Release-15 [66], standardized in March 2019) focuses on enhancing network capacity, while low-latency and machine-type communication tasks are still in progress. Moreover, the current 5G deployment commonly follows the pragmatic Non-standalone (NSA) mode, reusing the legacy 4G infrastructure to reduce cost. On the other hand, whereas 5G optimization mainly resides on the edge (*i.e.*, the radio access network, fronthaul/backhaul and the cellular core network), the end-to-end performance of mobile applications also depends on the wireline paths, cloud servers and even the processing capacity of the mobile devices. All in all, at this early stage, one natural question is: How far away is 5G from its prospects and what does it take to reach the tipping point of the 5G ecosystem?

In this paper, we perform a measurement study on one of the world's earliest commercial 5G networks, deployed in an urban environment and running on the sub-6 GHz spectrum. Using 5G-enabled smartphones and custom-built tools, we conduct in-depth active-passive measurements to characterize 5G from the physical layer to application layer, with particular emphasis on its comparison against 4G LTE. Specifically, we build a software toolset to log 5G's physical layer information (*e.g.*, channel quality and bit-rate) and fine-grained energy consumption traces to enable passive diagnosis. In addition, we leverage high-bandwidth cloud servers to set up an application service pipeline that can take advantage of the massive capacity of 5G, so as to enable active probing on the interactions among applications, networking protocols and the radio layer.

Measurement perspectives. Our measurement aims to demystify 5G from four major perspectives:

(i) 5G coverage (Sec. 3). In theory, due to its usage of higher frequencies than 4G, the 5G links suffer from more severe attenuation and penetration loss, leading to poor coverage. To understand the coverage issues in practice, we develop a 5G channel analytics tool based on XCAL-Mobile [4] – a commercial 5G air-interface monitoring framework. As a result, we can profile a comprehensive set of physical layer metrics, including signal strength, bit-rate, hand-off timing, *etc*, separated on a per-cell basis.

(ii) End-to-end throughput and delay (Sec. 4). 5G claims to support Gbps bit-rate and sub-10 ms latency, through its New Radio (NR) technology and a more flat core network architecture. However, its practical performance faces many attrition factors, *e.g.*, limited capacity of wire-line paths, poor interaction across layers within the network stack and link quality disruptions due to frequent hand-off across cells (each with limited coverage). To understand how these factors manifest in practice, we measure the end-to-end performance of mainstream transport-layer protocols, along with a breakdown of network latency. Our measurement identifies the bottlenecks and sources of anomalies that prevent 5G from delivering its expected performance.

(iii) Application performance (Sec. 5). Besides network performance, the application quality of experience (QoE) also depends on other factors, especially the processing capabilities of end-user devices. We thus investigate the inter-play between the communication and computing factors, by implementing a 4K/5.7K panoramic real-time video delivery system and characterizing the feasibility and challenges of the much anticipated immersive technologies over 5G.

(iv) 5G smartphone energy consumption (Sec. 6). 5G's high bit-rate comes at the cost of power-hungry signal/packet processors and RF hardware. In this paper, we develop an energy profiling tool - *pwrStrip*, to quantitatively analyze the power consumption on a typical 5G smartphone. Our analysis can breakdown the energy cost across different hardware/software components, as well as different radio states.

Summary of insights. Our measurement campaign leads to several major insights, which we summarize as follows:

(i) Our measurement reveals that even though the current 5G is densely deployed (approximately 0.077 km^2 per base station), there still exist many coverage holes outdoor. In addition, 5G channel quality suffers from a sharp degradation when transitioning indoor, with a drop of 50.59%, in comparison to 20.38% for 4G. Remarkably, we find that the current 5G base stations are all co-sitting with 4G ones, implying that the densification potential of 5G deployment can be further exploited.

(ii) We find that the de facto loss/delay based transport protocols (*e.g.*, Cubic, Vegas) behave abnormally when running over 5G, with a bandwidth utilization below 32%. An in-depth analysis shows that the legacy core Internet routers tend to cause excessive packet drops under 5G workloads. We identify and verify two possible solutions - proper buffer sizing, and adopting loss/latency-insensitive probing based transport protocols. In term of network latency, we find that the current 5G NR reduces "in air" latency by only less than 1ms, while the flatten core network architecture reduces latency by 20ms. However, the end-to-end latency remains similar to 4G LTE, as it is dominated by the wire-line paths. The results hint that the legacy Internet infrastructure also needs to be retrofitted to meet the prospects of the low-latency 5G. On the other hand, mobility worsens 5G latency. We find the cross-cell hand-off takes around 108.4ms, 3.6× longer compared to 4G, mainly due to the use of the NSA architecture.

(iii) As for application performance, we find that 5G offers negligible benefits to mobile Web loading, whose latency is dominated by either page rendering time or TCP's transient behavior which severely under-utilizes the network bandwidth. For 4K panoramic video telephony, 5G can improve video quality and smoothness owing to its high throughput. However, the codec/processing latency tends to outweigh transmission time by 10×, *i.e.*, the computing modules become the bottleneck in such demanding 5G use cases.

(iv) We find the 5G module results in alarmingly high power consumption, 2−3× over 4G and 1.8× over screen display which used to dominate the 4G phone power budget [42]. More interestingly, such high power consumption is intrinsic to the 5G radio hardware and DRX state machine, which makes standard power-saving schemes ineffective. Our trace-driven simulation shows that an oracle sleep scheduling mechanism can only reduce 5G power consumption by 16.02%, 12.24% and 11.17% for web browsing, video telephony and bulk file transfer, respectively; Whereas our heuristic-based scheme, which opportunistically offloads certain traffic to 4G, can achieve 25.04% power saving compared to the 4G module.

Our contributions. To our knowledge, this work represents the first cross-layer study of *operational 5G New Radio (NR, or sub-6 GHz) networks*, through a comprehensive measurement toolset. Our main contributions can be summarized as follows: *(i)* Quantitative characterization of 5G's coverage in comparison to 4G, which offers hints for optimizing deployment and hand-off/mobility management. *(ii)* Identifying an alarming TCP anomaly that severely underutilizes 5G capacity, diagnosing the root causes and proposing practical solutions. *(iii)* A breakdown analysis of the 5G end-to-end latency which pinpoints the bottleneck and space for improvement. *(iv)* Implementation and profiling of a 5G immersive media application to explore the feasibility and underlying challenges. *(v)* A detailed accounting of the power budget on 5G smartphones, along with pragmatic mechanisms to improve 5G energy efficiency. *(vi)* We have released our dataset and measurement tools to the public [68] for facilitating the future study.

2 MEASUREMENT METHODOLOGY

5G network. Our measurement is conducted in a densely populated city, which is one of the first regions with 5G coverage worldwide (*i.e.*, launched in April 2019). Most of our experiments focus specifically on a $0.5km \times 0.92km$ campus, where 6 5G base stations (gNBs) are deployed, surrounded by tall buildings, trees and heavy human activities. The 5G deployment adopts the NSA infrastructure, wherein a 5G gNB is co-located with an existing 4G base station (eNB). Under NSA, the 5G radio only operates within the data plane (or user plane), and relies on the legacy 4G LTE for control plane operations, as shown in Fig. 1. Both 5G gNBs and 4G eNBs share the same 4G Evolved Packet Core (EPC) network infrastructure [52]. To our knowledge, all the existing commercial 5G services are provided under the NSA architecture, due to its

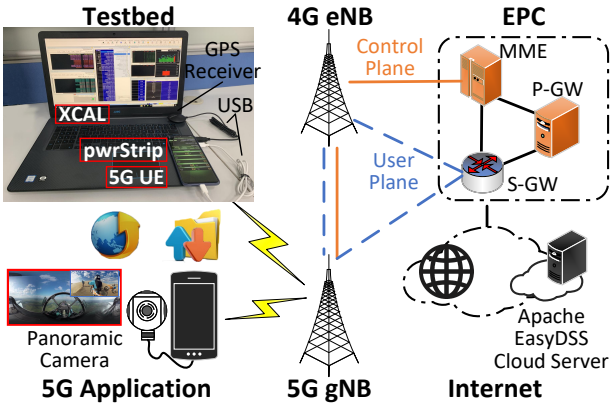

Figure 1: An overview of the measurement setup.

easy deployment and low cost.

Our 5G network operates at 3.5 GHz frequency, also referred to as New Radio (NR) or sub-6 GHz band (*i.e.* 0.45 GHz to 6 GHz) following 3GPP Rel-15 TS 38.104 [8]. The alternative 5G millimeter-wave (operating at a much higher frequency, *i.e.* 24.25 GHz to 52.6 GHz) has not been deployed by our local 5G operators, and hence it is beyond the scope of this work. One may refer to recent work in [60, 61] for a measurement study of 5G millimeter-wave networks. In Sec. 8, we will provide more details on the 5G spectrum usage.

5G user equipment (UE). Our measurement involves three 5G phone models: ZTE Axon10 Pro (Qualcomm Snapdragon TM855), HUAWEI Mate20 X (Hisilicon Kirin 980) and HUAWEI Mate30 Pro (Mali G76 | Hisilicon Kirin 990). Unless otherwise mentioned, we use Axon10 in most experiments. When we carried out our measurement campaign, only a few 5G smartphone models are available on the market. ZTE Axon10 Pro represents mainstream 5G smartphones with powerful communication (SDX M50 5G modem), computing (Snapdragon TM855) and storage (256 GB) capabilities. Axon10 also adopts the mainstream Qualcomm chip-set, allowing for a flexible diagnostic mode wherein certain PHY and MAC layer information can be extracted, to enable the low-layer measurement. We believe the computing and radio hardware profile of Axon10 represent the state-of-the-art, and the measurement findings hold true for other 5G smartphone models, particularly for these with the Qualcomm chip-set. In terms of network performance at TCP and application level, the phone model does not matter much. We have also used two HUAWEI 5G UEs to measure TCP and application performance, which show consistent results.

Cloud server. We deploy some of our measurement tools (mentioned later) in a HUAWEI cloud server (8vCPUs | 64GB | Ubuntu 18.04) with *1000 Mbps bandwidth* to match the 5G wireless bit-rate. It is noteworthy that such a Gbps cloud service incurs a high cost ($36.43 per hour), which hints that 5G services, particularly the bandwidth-hungry applications, may be too expensive for end-users at this stage. The server is located in the same city with our campus, and the geographical distance is about dozes of kilometers (due to our communication with the could service technicians). In addition, we utilize 20 SPEEDTEST [5] servers for end-to-end delay measurement (Sec. 4.4), and they are located 1 km to 3400 km away, as listed in Appendix C.

Measurement tools. We collect five types of trace information using existing or custom-built tools: *(i)* The 5G and 4G signaling information and physical-layer key performance indicators (KPIs) from a commercial software - XCAL-Mobile [4]; *(ii)* The TCP/UDP traffic traces generated by iperf3 [58] and captured by Wireshark [3]; *(iii)* The end-to-end delay along with per-hop latency extracted from traceroute [15]; *(iv)* Page loading time and video frame information gained from Google Chrome developer tools and our custom-built 5G application called *360TEL*. *(v)* The smartphone energy consumption traces obtained by a custom-built tool named *pwrStrip*. We proceed to provide more details on this tool set.

XCAL-Mobile runs on a laptop tethered to the Axon10 smartphone via a USB3 cable (Fig. 1). It can monitor the smartphone's diagnostic interface, including *(i)* basic cellular network KPIs on the physical and MAC layers, such as PCI, RSRP, RSRQ, SINR, CQI, MCS and the PRB allocation configurations. *(ii)* Signaling messages on the cellular control plane, allowing us to monitor hand-off start/end time, RRC state machine transitions, *etc.* We note that the advantage of XCAL-Mobile over other wireless network analyzers like MobileInsight [55] mainly lie in its capabilities in decoding the 5G signaling messages following the 3GPP Rel-15. To our knowledge, no other analysis tools possess similar capabilities to date.

Our TCP/UDP experiments are conducted using iperf3, of which the server-side is deployed in the cloud and the client-side executes on our 5G-devices. We set the receiver's buffer size to 25 MB, which is enough to avoid the small initial receiving window problem [46]. For each experiment, we log the traffic traces by Wireshark for off-line analysis.

For examining application-level performance, we develop a UHD panoramic live video Android APP (360TEL) based on the Insta360 ONEX's open APIs [76]. 360TEL runs on the 5G phone, which connects to an external panoramic camera with an H.264 hardware codec (Fig. 1). It can stream the captured 360° videos to the cloud server running EasyDSS RTMP [17], with up to 5.7K resolution and 30 FPS frame rate. 360TEL can be considered as a real-time panoramic video telephony system to enable immersive experience sharing.

To obtain the energy consumption statistics on smartphones, we tried many existing software solutions (*e.g.*, Battery Historian [48], Emmagee [49]). However, these methods either cannot provide fine-grained time series or fail to read 5G signaling information. On the other hand, the hardware solution, *e.g.*, Monsoon Power Monitor [59] typically used in 4G test, needs to remove the phone's battery, which is not feasible for our all-in-one 5G smartphones. Therefore, we custom-build an energy consumption logging and analysis tool, called *pwrStrip* by directly reading battery status (time-stamp, instant current, voltage, *etc*) from the Android kernel, at a fine-grained 100ms interval.

Following the above methodology, we perform extensive investigation over the 5G networks via both active and passive measurements, which lead to a dataset of 2.1 TB (366.5 GB for analysis after removing irrelevant payload) through 7 months. Leveraging on such a first-of-its-kind measurement campaign, we conduct a comprehensive and in-depth analysis to profile the current 5G networks. Moreover, we have publicly released our dataset along with the measurement tools at GitHub [68], for facilitating in-depth 5G exploration in our community.

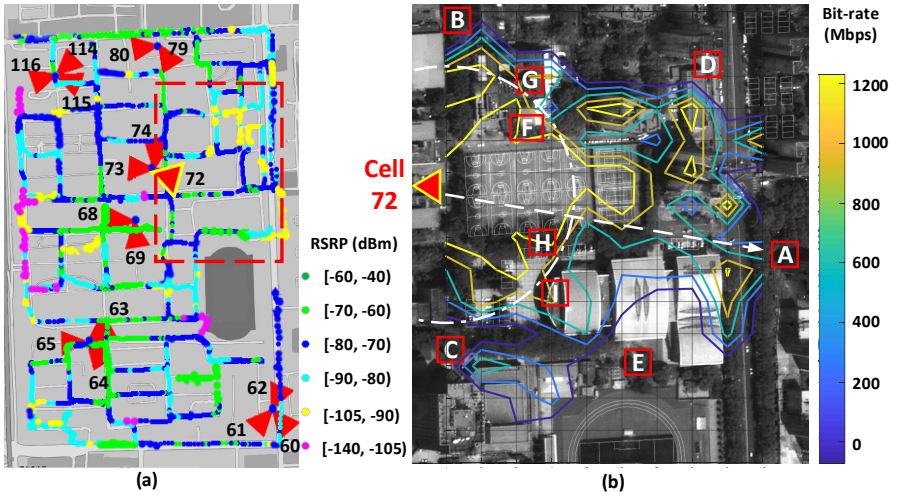

Figure 2: 5G network coverage. (a) The whole campus' RSRP map under a blanket road test; (b) The bit-rate contour of the gNB cell 72.

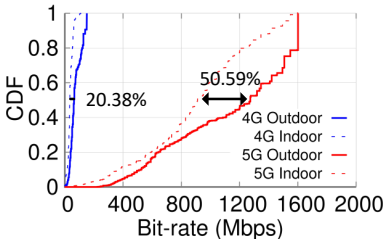

Figure 3: In-outdoor bit-rate.

Table 1: Basic physical info.

Info.	4G	5G
DL Band	1840	3500
(MHz)	~1860	~3600
# Cells	34	13
RSRP	-84.84	-84.03
(dBm)	± 8.72	± 11.72

Table 2: The statistic of RSRP distribution.

RSRP (dBm)	4G	5G	4G (6 eNBs)
[-60, -40)	6 (0.13%)	44 (0.95%)	6 (0.13%)
[-70, -60)	257 (5.56%)	378 (8.15%)	245 (5.29%)
[-80, -70)	1092 (23.60%)	1246 (26.88%)	1012 (21.86%)
[-90, -80)	1814 (39.20%)	1825 (39.37%)	1795 (38.77%)
[-105, -90)	1376 (29.74%)	769 (16.59%)	1390 (30.02%)
[-140, -105)	84 (1.77%)	374 (8.07%)	178 (3.84%)

3 5G COVERAGE

We first perform a blanket measurement within the campus region, and zoom in on the coverage of a single 5G gNB. We then study the indoor-outdoor capacity gap and finally examine mobile handoff across cells. In all experiments, we contrast the 5G with its counterpart 4G.

3.1 Campus Coverage

We traverse all road segments (6.019 km in total) within the campus region at a normal walking speed of about 4 - 5 km/h, while carrying an XCAL-equipped laptop and a smartphone to monitor the physical-layer information of both 5G and 4G. Meanwhile, we use a GPS receiver on the laptop to record the sampling locations. Through the blanket survey, we identify 6 5G gNBs on the campus. Each gNB consists of 2 or 3 sectors (cells) facing towards different directions, as marked with ▲ in Fig. 2(a). For instance, the bottom-right gNB has 3 cells with physical cell indicators 60, 61 and 62. The gNB cells operate at 3.5 GHz carrier frequency with 100 MHz bandwidth, in contrast to the 1.8 GHz carrier and 20 MHz bandwidth in 4G. Tab. 1 summarizes the general physical layer profile of the co-located 4G and 5G networks. Note that the RSRP (*i.e.*, reference signal received power) in Tab. 1 represents the average value across all sampling points. According to 5G standard Rel-15 TS 38.104 [8], the 3.5 GHz frequency belongs to the n78 band, which uses the TDD (Time Division Duplexing). In contrast, the 4G network uses the b3 band with FDD, *i.e.*, the uplink (UL) and downlink (DL) channels use orthogonal frequencies.

In addition, the 5G gNB density on our campus is 12.99/km^2, which is on a similar scale as the average density across our city's urban region (7000 eNBs in 667/km^2, *i.e.*, 10.49/km^2) [20]. Therefore, we believe the campus coverage profile is representative.

Despite the high deployment density, many coverage holes still exist as marked by pink dots in Fig. 2(a). These are the areas with the lowest level of RSRP [-140, -105] dBm, unable to initiate communication services. According to Rel-15 TS 36.211 [10], if the RSRP is less than -105 dBm, the communication service cannot be triggered. We randomly sample 4630 locations, and summarize the number of locations (and percentage) in each RSRP category in Tab. 2. We observe that: *(i)* The fraction of 5G coverage holes is non-negligible, *i.e.*, 8.07% locations have RSRP lower than -105 dBm, in contrast to only 1.77% for 4G. We find that a 5G gNB is always co-sitting with a 4G eNB due to the NSA. However, not all 4G eNBs have 5G companions, implying that the 5G deployment has not fully matured yet. Overall, the 4G eNB density is much higher (13 base stations in total, 28.14/km^2), which partly explains its better coverage. *(ii)* Even under the same deployment density (*i.e.*, only focusing on the 6 4G eNBs co-sitting with 5G counterparts), the fraction of 4G coverage holes is only 3.84%, still much lower than 5G. This may be due to the higher carrier frequencies used by 5G, and hence higher attenuation loss over the same distance. On the other hand, it implies that, for both 4G and 5G, the deployment has to be highly redundant to fix all the coverage holes. It is noteworthy that we carry out the experiments under the scenario with the daily human, bicycle, and vehicular traffic on campus. The coverage holes exist consistently irrespective of such environment dynamics.

3.2 Cell Coverage

We now characterize the coverage profile around one typical 5G gNB. We force the 5G phone to lock on a specific cell by turning on the frequency-lock switch (PCI 72, 1850 MHz in our setting) in the diagnostic mode. Notably, this is locking on the frequency of the master 4G eNB, because under NSA, the RRC configuration messages from the 5G UE must pass through the corresponding master 4G eNB before reaching the 5G gNB. Afterward, we partition the gNB's nearby region into 20 m^2 grids and sample 154

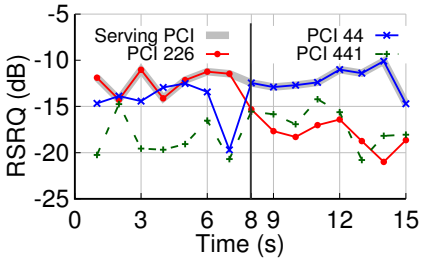

Figure 4: Signal quality evolution during hand-off (at 8s).

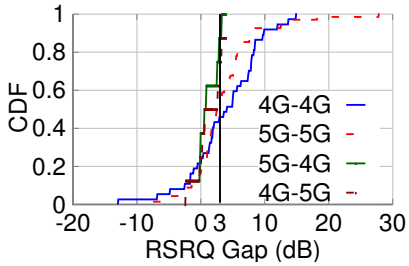

Figure 5: RSRQ gap before and after hand-off.

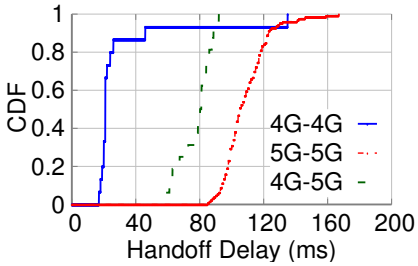

Figure 6: Comparison of the hand-off latency.

locations across all the grids. The contour plot in Fig. 2 (b) connects the sample locations with the same bit-rate. Our observations are as follows: *(i)* The contour lines obviously deviate from the ideal sector/circle shape, mostly biased by building blockage and multi-path reflections. *(ii)* To examine the impact of signal path-loss, we walk along a line-of-signal (LoS) path between cell 72 and location A in Fig. 2(b). We find that the 5G becomes disconnected due to too weak signal strength and only the 4G master eNB remains connected when reaching A (230m away from the gNB). This phenomenon occurs at other locations with similar distance from the sector, which implies that the coverage radius of one gNB is approximate 230m in dense urban areas like our campus. In contrast, typical 4G link distance is much longer, at around 520m, on the same campus. *(iii)* The limited field-of-view (FoV) of gNB cells, along with building occlusions, inevitably cause coverage defects. The gNBs commonly use sectionalized antennas with a fan-shaped gain pattern, and hence a narrow FoV. Location B and C outside the FoV are thus not covered. Whereas location D and E fall within the gNB's FoV, they cannot reach the gNB either, due to building blockage. A *deliberate arrangement of the gNB locations* may help maximize the coverage with minimum cost. For decades, such cellular network planning problems have been solved using blanket war-driving [31, 34], which becomes a daunting task as the 5G network density and parameter space grows. A more intelligent planning mechanism, which leverages the 3D terrain and building map information, may help automate and optimize the 5G deployment.

3.3 Indoor-outdoor Gap

In the single-cell measurement (Fig. 2 (b)), we also observe a huge bit-rate gap between indoor and outdoor cases. In particular, near the location F, G, H, I (100m from the base station 72), we use XCAL-Mobile to measure the bit-rate in the immediately adjacent indoor and outdoor spots. The resulting CDF plot in Fig. 3 shows that, *on average, 5G has more than 2× indoor-outdoor bit-rate gap, i.e., 50.59%, in contrast to only 20.38% in 4G.* We ascribe the larger bit-rate drop to the higher-frequency 5G signals (3.5 GHz), which suffer more from penetration loss than the 4G signals (1.85 GHz) [72]. Note that the buildings on our university campus have brick and concrete walls, which are common in urban scenarios. The observation on the indoor-outdoor bit-rate gap will hold true for similar environments. Drywalls and wood construction may experience lower penetration loss. One may refer to existing channel sounding work (*e.g.*, [50]) for a comparison across different material types.

Considering the meager coverage of 5G indoor, we believe combining micro-cells in residential/office buildings with the current macro-cells, such as the 5G/WiFi coexistence based on 3GPP ATSSS (Access Traffic Steering, Switching and Splitting) service [2], will lead to more seamless connectivity. Currently, a commercial 5G macro gNB equipped with 3 sectors costs $28,833.40, in contrast to $360.42 for a micro router (*i.e.*, CPE [64]) with about 120 m^2 coverage, which hints to an acceptable deployment cost. We will experimentally validate the performance of a typical indoor CPE in Sec. 8.

3.4 Hand-off across Cells

Due to smaller coverage, 5G hand-off (HO) is expected to become more frequent than in 4G. In this section, we first uncover the 5G HO strategy and evaluate its effectiveness, then we quantify the HO latency. The analysis is based on 407 HO events collected on the campus region and other areas in our city, during an 80 minutes measurement study at a walking/bicycling speed of 3~10 *km/h*. Among them, 387 are horizontal HOs, *i.e.*, switching between two 5G cells; 20 are vertical HOs (5G-4G and 4G-5G).

HO strategy evaluation. We use XCAL-Mobile to monitor the HO-related control signaling messages, *i.e.*, Radio Connection Reconfiguration messages with embedded HO configuration measurement report of the eNB/gNB. We first find that, although the smartphone reports 5 kinds of HO-related measurement events (*i.e.*, 21.98% A1, 0.18% A2, 67.25% A3, 9.19% A5, and 1.40% B1, as defined in [66]), the gNB only responds to the A3 event due to the ISP's configuration, and then triggers the HO procedure. For the definition of each hand-off trigger event, please refer to Tab. 5 in Appendix A. We thus focus on the A3 event, which indicates that the signal quality of the neighboring cell is higher than that of the current serving cell for a certain period [44], *i.e.*,

$$Mn + Ofn + Ocn - Hys > Ms + Ofs + Ocs + Off, \quad (1)$$

where Mn/Ms is the RSRQ value of the neighbor/serving cell. Ofn/Ofs is the frequency offset (default 0 dB). OCn/OCs is the cell specific offset (default 0 dB). Off is the intra frequency offset, and Hys is the frequency hysteresis (1 dB, 3 dB in our measurement, respectively). In order to prevent the frequent switching caused by signal fluctuations, two rules are applied in A3 event: *(i)* The RSRQ gap ($Mn - Ms$) between the serving cell and the neighboring cell must meet a threshold (3dB in the 5G gNB configurations, as we calculate from the above parameters using XCAL-Mobile) to successfully trigger a cell hand-off. *(ii)* A *timetotrigger* parameter is used as the time hysteresis, *i.e.*, a HO is triggered only when the above condition (Eq. (1)) holds true for *timetotrigger* (324ms in existing configuration).

To examine the effectiveness of such a HO strategy, we present

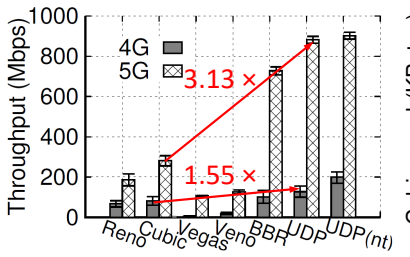

Figure 7: UDP and TCP throughputs.

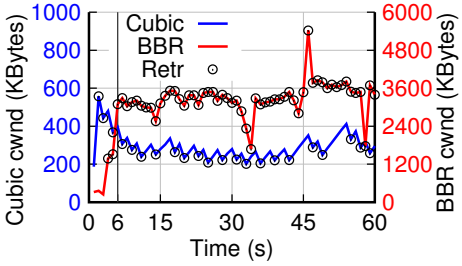

Figure 8: TCP cwnd evo. under 5G.

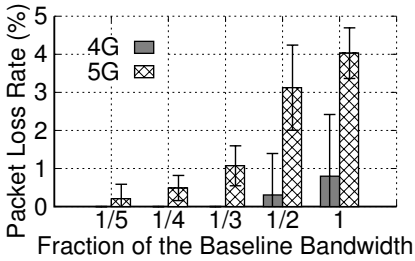

Figure 9: Packet loss ratio.

a case study in Fig. 4, where the UE switches from cell-226 to cell-44. If no HO occurs, the quality of the old cell will deteriorate. However, after switching to a new cell, the link quality does not necessarily get better. Statistically, we analyze and plot the changes in instantaneous RSRQ before and after each HO in Fig. 5. We observe that only 75% HOs have more than 3 dB RSRQ gain on average (80% for 4G-4G, 84% for 5G-5G, 75% for 5G-4G and 61% for 4G-5G). *The result reveals that the current empirical HO strategy in 5G has a non-negligible probability (i.e., 25%) of worsening link performance. A more intelligent strategy is required to determine when to trigger the HO.*

HO latency. A 5G-5G HO process starts with *LTE MAC RACH trigger*, and ends with the *NR MAC RACH Attempt (SUCCESS)* message, which can be captured by XCAL-Mobile. We compute the HO latency and plot the CDF of all the measured cases in Fig. 6. Surprisingly, the 5G-5G HO latency is 108.40 ms on average, while that of 4G-4G and 4G-5G is only 30.10 ms and 80.23 ms. We identify *the root cause to be the NSA architecture*, wherein 5G NR runs its own data plane, but relies on the control plane of the existing LTE network for control functions including HO management. In particular, the smartphone cannot directly switch to any 5G neighboring cells, but has to release its current 5G NR resource and roll back to the current 4G eNB. Then it performs a HO between the current 4G eNB and the target 4G eNB, and finally requests 5G NR resources on the target master 4G eNB. We confirm this complicated procedure by analyzing and extracting the compete HO signaling exchanges as given in Appendix A. It is expected that this long HO latency problem can be resolved in the future 5G SA architecture with independent data and control plane.

4 END-TO-END THROUGHPUT AND DELAY

4.1 Transport Layer Throughput

UDP throughput baseline. We use iperf3 to measure the maximum available bandwidth between the cloud server and the 5G smartphone. We gradually increase the UDP sending rate, and use the peak UDP throughput measured at the receiver side as the baseline. Each experiment is repeated 5 times for 60s during the daytime and late-night, respectively. From the results in Fig. 7, we see that the UDP baseline for 5G downlink (DL) is 880 Mbps on average during the day, in contrast to 130 Mbps for 4G. During the late-night, the UDP baseline of 5G DL increases slightly (*i.e.*, 900 Mbps), while that of 4G DL increases dramatically to 200 Mbps. The reason lies in the limited number of 5G users (small day-night variation), as 5G just entered an early commercialization stage. It is known that all users associated with the same base station need to share the same set of Physical Resource Blocks (PRBs), and for a given channel condition, a user's bit-rate is proportional to the number of PRBs. Using the XCAL-Mobile tool, we find that for 5G, almost all the PRBs (260~264 in a frame) are allocated to the smartphone under test regardless of time. In contrast, at nighttime, less user contention in the 4G network leads to more PRBs being allocated to the smartphone (*e.g.*, 95~100) than daytime (*e.g.*, only 40~85 PRBs).

Note that the maximum physical layer bit-rate is 1200.98 Mbps for 5G DL (time slot ratio is 3:1 for DL and UL in our ISP's configuration following Rel-15 TS 38.306 [13]) assuming all PRBs are allocated to one user and the highest Modulation and Coding Scheme (MCS) is selected. In particular, we often monitor the MCS index is 27, which corresponds to a maximum code rate of 0.925 for the highest spectral efficiency in 256 QAM. Thus the UDP baseline is 74.94% of the maximum physical bit-rate, which is reasonable considering the overhead in control channels and higher layer protocol operations.

In addition, the UL case is similar, for which the 4G/5G baselines are 50 Mbps/130 Mbps during daytime and 100 Mbps/130 Mbps at night. Unless otherwise specified, we use the daytime throughput as a baseline in the following experiments.

TCP throughput anomaly. We further examine the performance of three representative categories of TCP algorithms: Loss based Reno [28] and Cubic [39], delay based Vegas [22] and Veno [32], and the recently proposed capacity-probing based BBR [24]. For a fair comparison, we switch between different TCP algorithms by configuring the Linux kernel modules of the same pair of server/client, while keeping other settings intact. We use *bandwidth utilization* as the performance metric, defined as the throughput ratio between TCPs and the UDP baseline. From the results in Fig. 7, we observe that, for 4G, loss-based TCP and BBR perform reasonably well (utilization 52.9%, 64.4%, and 79.1% for Reno, Cubic and BBR, respectively), whereas delay-based TCP is known to perform poorly [91]. For 5G, BBR achieves reasonably high bandwidth utilization of 82.5%. However, *the traditional loss/delay based TCP algorithms suffer from extremely low bandwidth utilization—only 21.1%, 31.9%, 12.1%, 14.3%, for Reno, Cubic, Vegas, and Veno, respectively!*

To identify the root cause, we plot the evolution of congestion window (cwnd) of a typical 5G BBR and Cubic session, respectively, in Fig. 8. We find that BBR's cwnd remains high except for the slow-start phase (taking about 6 seconds as the gray line shows), while Cubic's cwnd never reaches its reasonable level due to frequent multiplicative decrease, which hints to severe packet losses. Thus, we proceed to examine the packet loss under different traffic load by sending UDP traffic at a certain fraction of the baseline bandwidth,

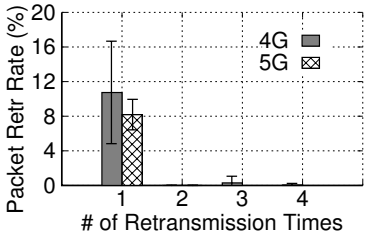

Figure 10: Retransmission stat.

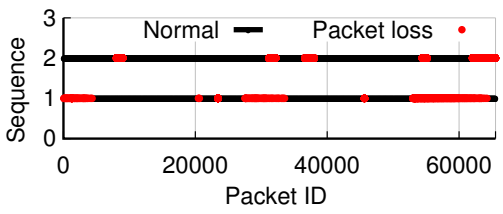

Figure 11: Bursty loss pattern of 5G.

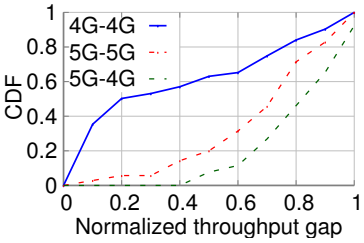

Figure 12: HO throughput drop.

Table 3: Buffer size on different network components.

Buffer Size	RAN	Wired Network	Whole Path
4G	468	10539	11007
5G	2586	26724	29310

i.e., $[\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1]$. *The results (Fig. 9) show that the packet loss of 5G sessions is multi-fold over the 4G sessions.* For instance, the loss already exceeds 3.1% (10× of 4G session) even at a mild $\frac{1}{2}$ of the baseline UDP bandwidth.

4.2 Locating the Performance Bottleneck

We proceed to locate where the packet loss anomaly takes place.

Packet loss in the radio access network (RAN). Due to the volatile wireless channel, packet losses are inevitable in the RAN, but the MAC/LLC layers usually adopt error checking/correction and retransmission mechanisms, such as ARQ and HARQ, to recover from losses and hide them from the upper layers. Although Rel-15 TS 38.321 [9] does not explicitly specify the retransmission threshold, we identify the value to be 32 based on the PDSCH configuration messages exposed by XCAL-Mobile. Even for an unusually lossy link with a loss rate of 50%, it is unlikely that it will experience 32 consecutive failed attempts (probability is only 2.3e-10). Our analysis of the XCAL-Mobile traces (Fig. 10) further verify that all retransmissions eventually succeed after up to 4 trials in 4G and 2 in 5G, which still falls far below the re-transmission threshold. In addition, we also find that almost all wireless resource has been allocated to the end device (observed from the PRB allocation statistics collected by XCAL-Mobile) when we measure the transport layer throughput. Therefore, the packet loss is irrelevant to MAC-layer resource allocation inside the gNBs. So, we can safely conclude that *the packet loss bottleneck is not on the 5G wireless link.*

In-network buffer estimation. Another potential reason for packet losses lies in buffer overflow along intermediate routers within the end-to-end path. We thus estimate the network buffer size, following the classical "max-min delay" method [25]. In a nutshell, the buffer size is the product of the longest packet queuing delay along the path (*i.e.*, the gap between RTT_{max} and RTT_{min}) and the estimated network capacity. Specially, we use traceroute to measure the RTTs of RAN and wired network, 30 times for each path and 60s for each measurement. Then we get the estimated buffer sizes (the maximum number of buffered packets) in Tab. 3. Note that the result is derived under the assumption of 1 Gbps path capacity and also 60 Bytes packet size. Though the absolute value of buffer sizes may deviate from the ground truth (due to inexact link capacity estimation), the ratios among them are accurate and support the following deduction. We observe that within the RAN, the 5G buffer size is 5× over 4G. But within the wired network (*i.e.*, from gNB to the cloud server), the difference is about 2.5×. As the wired network buffer takes a dominant role, the buffer size on the whole 5G path is roughly 2.5× compared with 4G.

In contrast, recall that the capacity of 5G DL is 5 × over 4G, *i.e.*, *the capacity growth is incommensurate with the buffer size expansion in the wireline network*, which is likely the reason for the high packet loss. The conjecture can be validated by the loss pattern. Specifically, we extract and plot two segments of packet sequence numbers in a 5G session, in Fig. 11. We find that the packet loss in 5G exhibits a clear bursty pattern, which should be caused by the intermittent buffer overflow.

An important question follows: *How much buffer is needed to eliminate the TCP anomaly in 5G?* Reasonable buffer size can be empirically determined by the *Stanford model* [16, 71, 85]: $B = (RTT \cdot C)/\sqrt{n}$, where C is the network capacity, and n is the number of concurrent flows. Here we assume the same flow number n, and similar RTT (validated in Sec. 4.4) for 4G and 5G networks, the total buffer size of 5G paths should be 5× of that of 4G paths. Considering the existing buffer statistics in Table 3, we suggest that *the buffer size in the wired network part should be increased 2× to accommodate 5G.* On the other hand, since buffer resizing may be costly and time consuming, *an easier solution is to adopt BBR-like algorithms that are less sensitive to packet loss/delay, at least for NSA – the transitioning phase of 5G.*

One concern is the bufferbloat issue [33, 38, 46], *i.e.*, deeper buffers may accommodate packets into long queues, thereby crashing delay-sensitive applications. In particular, 4G and 5G flows may share a common Internet path. While a larger buffer is helpful to reduce the packet loss rate of 5G data streams, it may hurts 4G flows. Therefore, the impact of large buffer on the trade-off between packet loss and delay requires in-depth research, which we leave for future exploration. For instance, there should be a more intelligent data distribution framework in the SA architecture, particularly for 5G and non-5G flows.

4.3 TCP Throughput During Hand-off

We now examine how 5G hand-off impacts TCP performance. We traverse the campus region and other areas in our city many times at a walking/bicycling speed (*i.e.*, 3~10 *km/h*) while continuously measuring the BBR throughput over 10ms windows. We use the same data set described in Sec. 3.4, consisting of 407 hand-off events. Fig. 12 plots the CDF of normalized throughput gap (*i.e.*, the percentage of TCP throughput drop immediately after the hand-off). We can observe that 5G-4G and 5G-5G hand-off suffer significant throughput degradation (83.04% and 73.15%, respectively), in contrast to only 20.10% for 4G-4G hand-off. The reason lies in the large hand-off latency (Sec. 3.4) which interrupts the normal TCP trans-

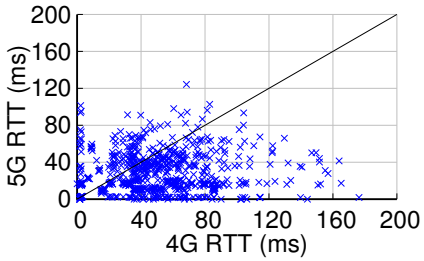

Figure 13: Delay statistic.

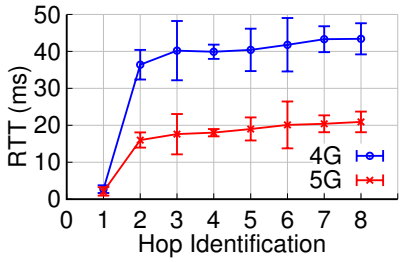

Figure 14: RTT along each path hop.

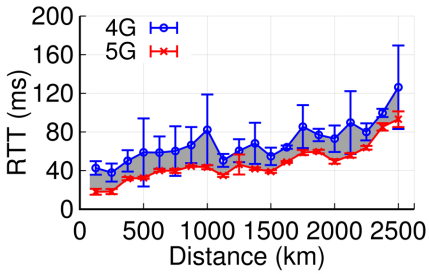

Figure 15: RTT vs. path length.

mission. The experiment again confirms the limitations of 5G NSA architecture.

4.4 End-to-end Latency

Overview. We measure the RTTs of 80 random paths crossing the 4G and 5G networks, respectively. Specifically, we select 4 5G gNBs (with co-sitting 4G eNBs) spatially spread across our city, and 20 other Internet servers nationwide. The location (latitude, longitude) of these servers can be found in Appendix C. For each pair gNB and server, we run traceroute on the 5G smartphone to measure the RTT. To ensure the traceroute probing packets not be fragmented on the router, we set the payload to be a minimum value of 1 Byte. In addition, we use UDP probing instead of the default ICMP to prevent the packets from being filtered out by some routers. We repeat the measurement 30 times for each path. The scatter plot in Fig. 13 shows the 4G vs. 5G RTT for each measurement. We have two observations: *(i)* 5G network paths achieve a network latency (*i.e.*, half of the RTT) of 21.8ms on average. In contrast, the Rel-8 TS 23.203 [6] mandates that for interactive real-time applications like VR, the transmission delay should be limited to 10ms. Clearly, the current end-to-end latency of 5G NSA is insufficient to meet such requirements. *(ii)* Nonetheless, the 5G paths still reduce RTT by 22.3ms (31.86%) on average, compared to 4G.

Delay breakdown. We then investigate where 5G's latency reduction comes from. We select one example network path consisting of 8 hops, and measure the RTT as hop count increases. From Fig. 14, we find that: *(i)* The RAN latency reduction (hop 1) is negligible: 2.19±0.36ms (5G) vs. 2.6±0.24ms (4G). Note that the Rel-15 38.913 [7] standardizes the 5G air interface delay 4ms for eMBB (enhanced Mobile Broadband, which is satisfied by the current NSA architecture), and 0.5ms for uRLLC (ultra Reliable Low Latency Communications, which remains to be met in the future SA architecture). *(ii) The delay reduction mainly comes from the second hop (i.e., from the gNB to cellular core network).* The reduction attributes to the flatten architecture of 5G (*i.e.*, part of cellular core network functions sinks to gNB so as to minimize processing latency [66]) and the specialized 25 Gbps fiber fronthual/backhual (according to our communication with ISP technicians).

Delay vs. path length. To put the above delay reduction in a big picture, we re-arrange the RTTs according to the geographical distance of each path, as shown in Fig. 15. We see that: *(i)* The RTTs of both 4G and 5G increase with path length. In particular, the RTT increase by 5× as distance increases from 100 km to 2500 km, and RTT reaches up to 82.35ms on average for 5G paths. *(ii)* The RTT gap between the 4G and 5G networks is 22±3.57ms on average, but the ratio between the gap (shown by the shade) and the absolute RTT value becomes smaller as path distance increases. The findings convey a message that *the untamed latency in the wireline paths, which is beyond mobile carriers' control, may neutralize 5G's latency advantage. To unleash the full potential of 5G applications, the legacy wireline networks also need to be retrofitted, so as to effectively reduce the end-to-end latency. Emerging architectures that shorten the path length, e.g., edge caching and computing, may also confine the latency.*

5 APPLICATION PERFORMANCE

We take the mobile web browsing and emerging UHD panoramic video telephony as two representative examples to examine the application QoE under 5G.

5.1 Web Browsing

We build an HTML5 website and deploy it in an Apache 2.0 cloud server. The website consists of multiple web pages, *i.e.*, web search, image, on-line shopping, map navigation and HTTP video streaming. For instance, there are some pages with different size/resolution images, which can be used to test the page loading time (PLT) of image browsing. PLT comprises two parts: The content downloading, and page rendering time. Both are measured using Google Chrome's developer tools [36] on a DELL G3 3779 laptop (Windows10 x64 | Intel Core i7 8th generation | 64 GB RAM | 512 GB SSD). We use HTTP/2.0 + BBR, and clear the web cache and cookie before each experiment to avoid their artifacts. Fig. 16 plots the mean and *std.* of PLT calculated across 10 websites in each category. Despite the 5× DL throughput gain (validated in Sec. 4), the 5G PLT shows minimum reduction (5% on average) compared with that in 4G.

A breakdown of the PLT latency in Fig. 17 reveals two causes: *(i)* The rendering time takes a dominant fraction in PLT (especially for large-size pages), which only depends on the computational capacity of end-devices rather than network throughput. *(ii)* Even when considering the downloading time alone, 5G only provides a marginal 20.68% reduction on average across the 5 categories. The reason lies in the transitioning behavior of TCP, *i.e.*, the slow-start phase. Our measurement shows that, even for the most 5G-friendly TCP (*i.e.*, BBR), the slow-start phase lasts about 6s before it converges to the high network bandwidth (Fig. 8). Unfortunately, most web pages are only a few MB and have already finished downloading well before TCP converges, which heavily underutilizes the 5G bandwidth.

To sum up, the web browsing performance is still handicapped by the computational efficiency of mobile devices, which cannot be resolved by 5G. Also, the transient behaviors of TCP severely hamper the 5G network efficiency, especially for short bursty-flows. It is unlikely

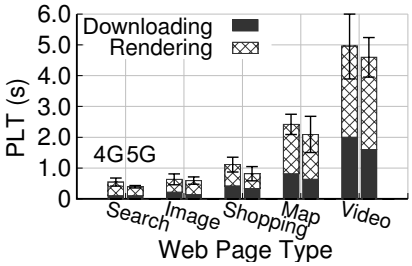

Figure 16: PLT of different websites.

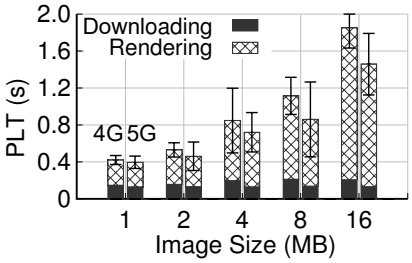

Figure 17: PLT of different images.

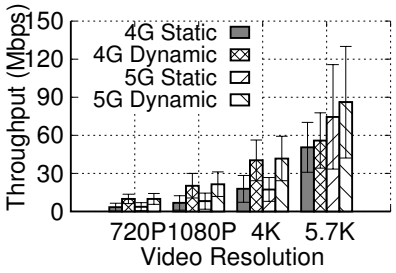

Figure 18: Video throughput.

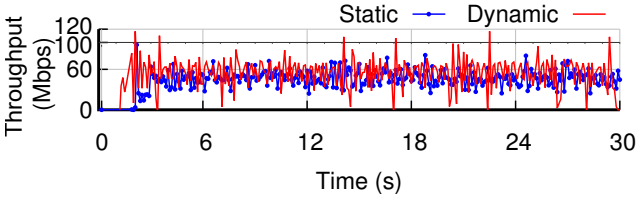

Figure 19: Received 5.7K video throughput fluctuation under 5G networks.

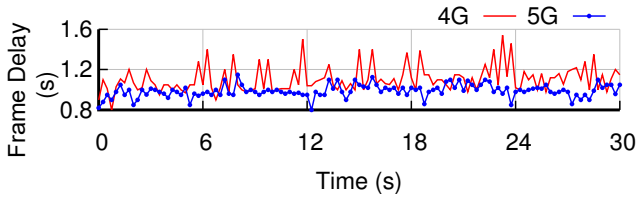

Figure 20: Frame delay of 4K video telephony.

that 5G will tailor itself for TCP, as it violates the end-to-end design principle of the Internet [27]. However, a minor upgrade of the TCP at the end hosts is still justifiable and may eventually unleash the 5G potential. For example, it has been shown that replacing TCP's slow-start probing with a deterministic bandwidth estimation [90] may substantially improve TCP efficiency over cellular networks.

5.2 UHD Panoramic Video Telephony

Mobile UHD panoramic video telephony poses a high demand on network capacity and stability, especially for the uplink (UL). The previous study [18] has shown that 4K telephony produces heavy traffic load with unpredictable fluctuations (35-68 Mbps [75]), making it unaffordable for 4G networks. It is much anticipated that 5G will be a niche technology to resolve this issue. We now validate the feasibility using the 360TEL system that we developed (Sec. 2).

Tolerance on video throughput fluctuation. 360TEL consists of a UL pushing flow (sender→server) and DL pulling flow (server→receiver). We first measure the UL video throughput at the server using Wireshark and present the results in Fig. 18, where *dynamic* represents the case of constantly changing the camera's view. We observe that the average received throughput of all HD resolution videos (720P, 1080P, 4K and 5.7K) does not exceed the 5G UL capacity (100 Mbps at daytime from Sec. 4.1). However, 4G networks cannot support a 5.7K video. The average throughput of 5.7K video under 4G is much smaller than that under 5G, because of the network congestion and the consequent significant video frame losses.

Despite enough bandwidth of 5G UL, the fluctuation of video streams [18] may still cause low QoE. In our experiments, we observe that 5G can well tolerate 4K's fluctuation, but is sometimes ineffective for 5.7K video. Fig. 19 plots the received video throughput over a 30s 5.7K video session, captured in static and dynamic scenarios, respectively. The large fluctuation is remarkable in the dynamic scenes, which escalates the video traffic (sometimes beyond the 5G UL bandwidth) and thus causes frame freezing. We find 6 frame freezing events within our real-world measurement.

Frame delay. The end-to-end video frame delay is critical to QoE of the real-time video telephony. We measure the frame delay using a "stopwatch timing" method. Specially, we use the Insta360 camera to shoot a stopwatch (t_1) and then record the time from the same stopwatch displayed on at the video receiver (t_2). $\Delta t = t_2 - t_1$ is the end-to-end video frame delay. Fig. 20 plots Δt for a 30s 4K video session described above. It is evident that 4G suffers from severe congestion and hence occasionally long latency due to its insufficient bandwidth. Surprisingly, even for 5G, the frame latency remains on the level of 950ms, which falls short of the 460ms requirements of real-time video telephony [88].

We breakdown the frame delay, by separately examining frame processing operations (camera capture, frame patch splice, codec and video rendering), RTMP streaming/receiving and network transmission. Following the same approach as in [18], we find that the encoding time of the H.264 hardware codec is about 160ms and the decoding time is 50ms. We then estimate the frame capture, patch splice and rending time to be about 440ms by calculating the latency gap between the stopwatch and the time of the preview video shown on local devices. Note that the preview video can roughly cover the whole sender's frame processing, because it is not sent to the transmission link by RTMP streaming. Overall, the frame processing latency is about 650ms in our measurement, which is 10× than the network transmission delay (66ms) for each frame! In other words, the frame processing latency conspicuously parallelizes the end-to-end delay in Sec. 4.4, even under 5G NSA pattern.

To sum up, the high bandwidth of 5G provides more redundancy to tolerate video traffic fluctuations, but the delay spent on smartphone's local processing remains as a prohibitive latency bottleneck, ruining the user experiences in real-time interactive. Thus, it is imperative to improve the smartphones' processing capacities in order to support 5G's niche applications, such as immersing interactive video telephony which demands both high bandwidth and low end-to-end latency.

6 5G SMARTPHONE ENERGY CONSUMPTION

In this section, we first profile a 5G smartphone's energy consumption when running mainstream applications. We then run a micro-

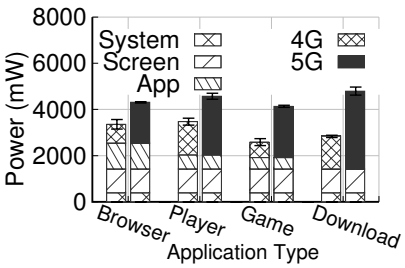

Figure 21: Energy consumption breakdown under daily app. usage.

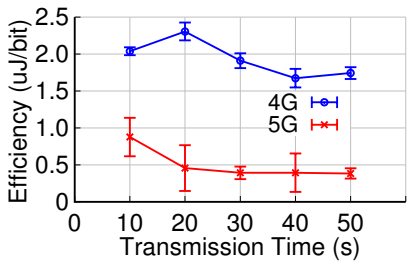

Figure 22: Energy efficiency under fully-saturated traffic.

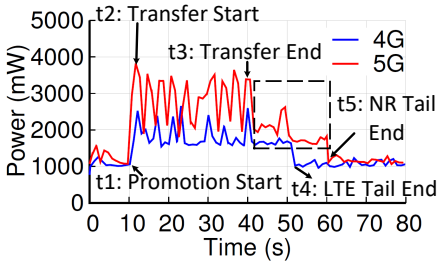

Figure 23: A showcase of 5G energy management.

scopic analysis of the power management scheme under the 5G NSA. Guided by the measurement and analysis, we propose and validate a power saving scheme. Our experimental results are derived from two ZTE Axon10 Pro 5G phones, but the power budget breakdown should be generalizable to other 5G phone models.

6.1 Profiling the Energy Consumption of 5G Applications

Energy consumption on 5G radios vs. other smartphone components. We use *pwrStrip* to measure the smartphone's energy consumption when running 4 typical applications: Google Chrome, Tencent video player, Arrow.io Cloud Game and File downloader. We breakdown the overall energy cost into 4 parts as follows: *(i)* To get the *Android system* consumption, we turn off the screen and turn on the "airplane" mode to kill all background applications. *(ii)* We then measure the screen element at the maximum brightness with other settings unchanged. *(iii)* To obtain the power consumption of the application alone, we load the application's contents in advance and run the applications off-line. *(iv)* We finally record the energy trace of the 4G/5G radio interface at normal operations. Notably, the wireless channel quality may affect energy consumption [29]. For instance, poor wireless channel degrades the bit-rate, HARQ efficiency and other MAC operations, which increases the energy per bit. To isolate the bias caused by wireless signal quality variation, we carefully carry out all the measurements in regions with a consistent RSRP level of [-80, -60] dBm.

The results in Fig. 21 show that: *(i) The 5G module dominates the energy cost (accounting for 55.18% on average of the total budget across 4 applications), far exceeding the screen (30.73%) which is known to be the most power-hungry component before 5G [42].* In comparison, 4G only accounts for 24.22% - 50.20%. The heavy energy consumption of 5G radio attributes to its more powerful baseband and RF hardware [86], *e.g.*, wide-band data converters (100 MHz vs. 20 MHz in 4G) and 4×4 MIMO [65]. In addition, we note that the mainstream 5G smartphones have not launched a SoC (System on Chip) solution [30, 56] with integrated CPU, GPU, DSP and 5G radio modems. Instead, they adopt separate but less energy efficient 5G modems as a plug-in to the legacy 4G SoC. An example is the Qualcomm Snapdragon TM855 [70] plus SDX 50M 5G modem combo used in our phone model. Immature packaging of 4G SoC and 5G modem, plus the interaction overhead between the processor and the modem, lead to much more power loss than an integrated solution. *(ii)* The total power consumption increases with application traffic intensity and the fraction of power spent in Data Transmission increases accordingly. We quantitatively run iperf3 UDP to download data for different lengths time with a saturated sending rate. We calculate and plot the energy-per-bit in Fig. 22. We find that the energy-per-bit of 5G is only $\frac{1}{4}$ of 4G. This implies that *5G can be much more energy efficient than 4G, but only when if upper layer protocols can fully utilize its available bit-rate, and a proper power management scheme is in position to activate the radio only when necessary.*

6.2 A Showcase of 5G Energy Management

The energy management of 5G radio follows a state machine (details in Appendix B). In general, 5G radio switches between *RRC_IDLE* (no Tx/Rx) and *RRC_CONNECTED* (on-going Tx/Rx) status, and adopts the discontinuous reception (DRX) mechanism for power saving. Fig. 23 plots the fine-grained energy traces for an example application session, where we trigger a web loading per 3s for 10 times, starting from time 10s (t_1) and ending at 40s (t_3). Comparing against the same experiment for 4G, we have the following observations: *(i)* Since 5G bandwidth is poorly utilized in short burst flows, it consumes 1.67× more energy (J) than 4G when running the same web loading sessions. *(ii)* The power consumption pattern of 5G shows very obvious jagged fluctuations, which is caused by the discontinuous page downloading operations. In particular, when triggering a webpage download, the energy consumption increases for entering the *RRC_CONNECTED*. After each downloading transmission (less than 3s interval), the UE uses the *DRX* mechanism that consumes less power. Similar fluctuation within a smaller range is also observed for 4G. *(iii)* 5G has an obvious long-tail stage when rolling back from *RRC_CONNECTED* to *RRC_IDLE*, which leads to an additional waste of energy. More specifically, after the transfer ending at t_3, we continue to monitor the Android kernel until the power value recovers to the *RRC_IDLE* level. From our analysis, 4G returns to *RRC_IDLE* after about 10s (at t_4), while 5G takes about 20s to finish the tail stage (at t_5). We can observe that the long tail stage, while existing in 4G [42], is exacerbated by the 5G NSA architecture. In particular, to finish the switch from NR *RRC_CONNECTED* to *RRC_IDLE*, the 5G module must first go through the 4G state machine via LTE RRC Reconfiguration (see Fig. 25 in Appendix B). The process is equivalent to activating an LTE tail period again (marked by the black box), which compounds the tail energy overhead.

6.3 Optimizing the 5G Power Management

The above experiments indicate a simple way to improve the power efficiency of the 5G state machine: We can adopt a dynamic mode

Table 4: Energy consumption (J) of different models.

Model	Web	Video	File
LTE	85.44±1.08	227.13±5.26	357.67±8.22
NR NSA	113.94±1.31	140.19±0.69	157.29±1.03
NR Oracle	95.69±1.18	123.03±0.57	139.72±1.03
Dyn. switch	85.41±1.07	133.66±0.71	150.80±1.03

selection scheme, which turns on the energy-hungry 5G module only when necessary. Specifically, if the instantaneous traffic intensity measured at the UE is approaching 4G's capacity, *i.e.*, 100 Mbps, we switch the radio into the 5G NR module; Otherwise, it should stay in 4G mode. To verify the effectiveness of this scheme, we use a trace-driven simulator because the smartphone UE does not expose an interface for 4G-5G switching. For comparison, we also implement the current power management approaches (with DRX configuration of 4G LTE and 5G NSA NR in Tab. 7 and an Oracle model (*i.e.*, with perfect sleep and awake transition) approach under 5G. We collect 3 types of real-wold traffic using Wireshark (*i.e.*, short web page browsing, frame-by-frame UHD video telephony and saturated file transfer, 10 flows for each type), and replay them in simulated state machines. Note that here each model finishes the whole data transfer of all flows, so that the completion time under 4G and 5G are diverse, which is different from that in Fig. 21 where all experiments last the same time.

Tab. 4 summarizes the results. We observe that: *(i)* Dynamic mode switching saves a remarkable amount of energy (24.8%) over the NR NSA for unsaturated web browsing flows. It is noteworthy that dynamic 4G-5G switching may also be a use case for MPTCP [53], which is an interesting topic particularly considering the long-term 4G/5G coexistence. We leave this for future exploration. *(ii)* The NR Oracle does not show a significant advantage over NR NSA, with an average gain of 13.2%, implying optimizing the 5G power management protocol alone provides marginal benefits, as the bottleneck may lie in the hardware itself.

7 RELATED WORK

5G measurements. In general, empirical studies of commercial 5G networks are quite limited as they were deployed only a few months ago. Qualcomm, as a major 5G radio manufacturer, released a white paper to profile 5G performance [69], but mainly focusing on signal quality and coverage. Narayanan *et al.* conducted a preliminary measurement of the 5G mmWave network [61], which differs drastically from the sub-6 GHz 5G NR due to the use of much higher frequencies and directional beams. In addition, the study mainly focused on upper layer performance due to lack of access to physical/MAC layer diagnostic information. In contrast, our measurement campaign represents the first to measure and analyze commercial sub 6 GHz 5G NR. Using a set of cross-layer tools, we were able to identify critical protocol level and computational bottlenecks for 5G (*e.g.*, the TCP anomaly and high energy cost under NSA), which has not been discussed before.

5G modeling and simulation. Prior to the deployment and commercialization of 5G NR, extensive theoretical modeling or simulation study have been performed, from various perspectives such as flatten radio access architecture [79], software-defined core network [26], ultra-dense picocell [40] to multi-radio (cloud) access [67], massive MIMO [57], interference management [62], spectrum sharing with cognitive radio [81],*etc.* More references can be found in survey studies [14, 35]. Our measurement study complements such works through a comprehensive profiling of commercial 5G networks in a realistic environment.

Measurement studies on legacy 4G. Substantial research effort has been devoted to 4G cellular networks since their commercial deployment. We categorize representative topics as follows: *(i)* The 4G cellular infrastructure, deployment and hand-off have been extensively studied in literatures, such as [54, 74] and references therein, which discover and propose optimization methods for the specific challenges as cells going smaller, including coverage holes and instability of mobility management, *etc. (ii)* In term of transport performance, many papers re-visited the classical congestion control algorithms under 4G LTE, and identified new problems such as bufferbloat[33, 38], cross-layer gap [43, 78, 89], and network bottleneck [19]. Besides the traditional urban or rural measurement fields, recent works also examined how 4G performs under emerging scenarios such as high-speed rails [53, 80]. *(iii)* Many popular mobile applications are known to suffer from low quality of experience in cellular networks. Approaches to improve application performance have been proposed for panoramic VR [37, 75, 88], web browsing [47, 93], *etc. (iv)* As for energy consumption, early work for 4G [84, 97] enhances smartphone's energy managements with adaptive DTX or DRX algorithms and optimized RRC state machine, while [41, 42] propose an empirical traffic-driven power model. We note that 5G exhibits substantially different behaviors compared against 4G, and poses new challenges and opportunities in multiple dimensions, such as infrastructure development, transport protocol design, and application QoE optimization. Our work represents the first to reveal such perspectives, which can be further explored by follow on research.

Our study relies on a set of custom-built and commercial tools to enable cross-layer cellular network analytics. Although such tools with similar capabilities have been developed for 4G LTE [51, 55, 63], they are heavily tailored for the 4G physical/MAC layer, and do not support 5G yet.

8 DISCUSSION AND FUTURE WORK

5G spectrum. According to the 3GPP standards, 5G's spectrum includes the sub-6 band (also called C-band, from 0.45 GHz to 6 GHz) and millimeter-wave (mmWave) band (from 24.25 GHz to 52.6 GHz). In practice, an operator's choice of 5G band depends on its spectrum license and technical/business considerations. For example, China Mobile builds its 5G network over 2.5 GHz - 2.6 GHz and 4.8 GHz - 4.9 GHz band, China Unicom uses 3.4 GHz - 3.5 GHz, while Verizon started their 5G deployment using the mmWave bands in Minneapolis and Chicago [60, 61]. Our measurement is conducted on the 3.5 GHz band, which is the most prevalent 5G deployment during our study. We believe our findings still hold true for other sub-6 bands. The mmWave bands may behave differently, especially due to the distinct channel characteristics at higher frequencies. For more details, one may refer to a recent measurement study of the 28 GHz 5G networks [61], and many experimental works on mmWave networks [82, 83, 95, 98].

In this work, we measure the 5G NR performance with a direct comparison with 4G/LTE. As a licensed wireless access technology,

5G has fundamental differences from the unlicensed WiFi, even though both use the sub-6 GHz band. For instance, the wireless resource allocation in 5G is performed in a central way, while WiFi client users contend resources in a distributed manner. More importantly, the 5G performance also depends on the cellular network core, which is not a problem with WiFi.

Measurement scale. The measurement of our work is mainly performed on a university campus ($0.5km \times 0.92km$), which is one of the first regions with 5G coverage worldwide (as of April 2019). A few findings (particularly the network coverage) may change if under another measurement area with different gNB density, but other findings, like end-to-end TCP performance, video streaming QoE and smartphone energy consumption, are unaffected by measurement scale. Even with respect to network coverage, since the ISP's deployment goals are consistent across areas, we do not expect fundamental differences between the sample area and a larger area.

A larger-scale study may be important for high mobility scenarios involving frequent hand-off, *e.g.*, on moving cars or even high-speed trains traversing between cities [53, 80]. However, 5G coverage is far from continuous at intra-/inter-city scale for now, and cannot support such high mobility. We thus leave this investigation for future work.

Persistence of the measurement findings vs. network evolution. Most observations in this measurement study are intrinsic to 5G, although some issues could be resolved as the 5G RAN becomes mature and the core network evolves accordingly. In particular, the coverage holes can be eliminated as gNB density increases, and the packet loss ratio will decrease if the ISPs can enhance their Internet routers with larger buffer sizes, and the UEs' energy consumption may decrease as 5G-specific SoCs emerge. However, such an infrastructure update involves high monetary cost and may take a long time. The merit of this work lies in that it performs a rigorous investigation of the existing 5G from multiple dimensions, and the findings hint on directions for the coming 5G evolution.

Exploiting the coexistence of 4G and 5G. Our measurement shows that the NSA architecture has caused several issues, including large hand-off latency and low energy efficiency, which compromises the effectiveness of 5G. It is expected that 5G NSA will persist during the whole NSA→SA transition phase, lasting for 4~5 years [45]. A current pragmatic solution is to leverage on 4G/5G mode adaptation to facilitate mobility and energy efficiency. Though the idea has been primarily validated in this work, a systematic and rigorous treatment is left for future study.

Looking into far future, the 5G control plane in SA will also have a set of different challenges. For instance, there will be denser 5G gNBs due to smaller coverage, which shall entail higher coordination complexity. In addition, the control plane in SA is expected to be more efficient compared with the existing NSA architecture. To meet the objective, the incorporation of new technologies including MEC and SDN/NFV is needed [1, 21].

Can 5G replace DSL? A long-lasting debate is whether 5G can replace the existing digital subscriber line (DSL), particularly for home access networks. We examine such feasibility using trace-driven simulation. In particular, we measure the throughput from a gNB to a HUAWEI 5G CPE Pro [64] (*i.e.*, a 5G-WiFi gateway) placed in a residential building. The average throughput reaches 650 Mbps when the CPE stays at favorable locations (*e.g.*, near windows). Then, for a typical residential area (say 50 houses covered by a 5G gNB with 3 cells and 200m radius), each house can get around 39 Mbps throughput, which exceeds the average DSL bitrate of 24 Mbps in the US [23]. The result indicates that 5G can potentially replace DSL in such settings. In addition, 3GPPP is currently developing the ATSSS service [2], which supplements the WiFi/5G coexistence model as a trend for future indoor-outdoor cooperatively deeper network coverage.

The potential impact of mobile edge computing. In Sec. 4.1 and 4.2, we attribute the abnormal TCP behaviors of many congestion control algorithms to the dramatic packet loss occurred along the end-to-end Internet paths (*i.e.*, buffer overflow on the routers). One possible solution is mobile edge computing (MEC), *i.e.*, deploying edge caches and servers closer to users (*e.g.*, just behind base stations), and thus alleviating the problem of insufficient bandwidth along the end-to-end paths. For instance, bandwidth-hungry edge applications like UHD video playback or the more interactive AR/VR can use a MEC platform to offload massive data traffic, thereby reducing core network pressure and application-level latency. However, there are also end-to-end applications (*e.g.*, remote robotic surgery [92], video telephony [96]), for which edge computing is not very helpful, and their performance relies on sufficient bandwidth along the entire end-to-end Internet path. Therefore, to fully unleash the 5G potential, other components of the 5G ecosystem (*e.g.*, wireline paths, upper-layer protocols) need to co-evolve with the edge side of 5G, *i.e.*, radio access architecture and also MEC.

9 CONCLUSION

We perform a full-fledged, end-to-end measurement study spanning multiple interacting networking layers, on one of the first commercial 5G networks, focusing on network coverage, end-to-end throughput and delay, applications performance as well as energy consumption. Our measurement reveals imperative issues that hamper the performance of 5G. Some of these problems (*e.g.*, surprisingly low bandwidth utilization) can be solved through proper network resource provisioning or more intelligent protocol adaptation, but others (*e.g.*, long latency and high power consumption) entail long-term co-evolution of 5G with the legacy Internet infrastructure and radio/computing hardware. Our measurement insights point to feasible directions to optimize 5G as an ecosystem to meet the demanding application requirements.

ACKNOWLEDGMENTS

We appreciate the insightful feedback from the anonymous reviewers and our shepherd Prof. Ellen Zegura who helped improve this work. The work is supported by the Innovation Research Group Project of NSFC (61921003), NSFC (61772084, 61720106007,61722201, 61832010), the 111 Project (B18008), and the Fundamental Research Funds for the Central Universities (2019XD-A13). Anfu Zhou and Huadong Ma are corresponding authors. We claim that this work does not raise any ethical issues.

REFERENCES

[1] 2018. MEC in 5G networks. https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf. (2018).
[2] 2019. 5G ATSSS. https://www.mpirical.com/blog/5g-atsss. (2019).
[3] 2019. Wireshark. https://www.wireshark.org/. (2019).
[4] 2019. XCAL Mobile. http://accuver.com/acv_products/xcal-mobile/. (2019).
[5] 2020. SPEEDTEST. https://www.speedtest.net/. (2020).
[6] 3GPP. 2014. Base Station (BS) radio transmission and reception. https://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/. (2014).
[7] 3GPP. 2018. Study on Scenarios and Requirements for Next Generation Access Technologies. https://www.3gpp.org/ftp/specs/archive/38_series/38.913/. (2018).
[8] 3GPP. 2019. Base Station (BS) radio transmission and reception. https://www.3gpp.org/ftp/Specs/archive/38_series/38.104/. (2019).
[9] 3GPP. 2019. Medium Access Control (MAC) protocol specification. https://www.3gpp.org/ftp/specs/archive/38_series/38.321/. (2019).
[10] 3GPP. 2019. Physical channels and modulation. https://www.3gpp.org/ftp/specs/archive/36_series/36.211/. (2019).
[11] 3GPP. 2019. Radio Resource Control (RRC) protocol specification. https://www.3gpp.org/ftp/Specs/archive/36_series/36.331/. (2019).
[12] 3GPP. 2019. Radio Resource Control (RRC) protocol specification. https://www.3gpp.org/ftp/specs/archive/38_series/38.331/. (2019).
[13] 3GPP. 2019. User Equipment (UE) radio access capabilities. https://www.3gpp.org/ftp/specs/archive/38_series/38.306/. (2019).
[14] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. 2016. Next generation 5G wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 18, 3 (2016), 1617–1655.
[15] amatsuda. 2018. A Rake task gem that helps you find the unused routes and controller actions for your Rails 3+ app. https://github.com/amatsuda/traceroute. (2018).
[16] Guido Appenzeller, Isaac Keslassy, and Nick McKeown. 2004. Sizing router buffers. *ACM SIGCOMM Computer Communication Review* 34, 4 (2004), 281–292.
[17] Babosa. 2019. An open RTMP server. https://github.com/EasyDSS/EasyDSS. (2019).
[18] Ghufran Baig, Jian He, Mubashir Adnan Qureshi, Lili Qiu, Guohai Chen, Peng Chen, and Yinliang Hu. 2019. Jigsaw: Robust live 4k video streaming. In *ACM MoibiCom*.
[19] Arjun Balasingam, Manu Bansal, Rakesh Misra, Kanthi Nagaraj, Rahul Tandra, Sachin Katti, and Aaron Schulman. 2019. Detecting if LTE is the Bottleneck with BurstTracker. In *ACM MobiCom*.
[20] New Beijign. 2019. 5G base station deploy in Beijing. http://mobile.ctocio.com.cn/mobile/2019/1226/5890.html. (2019).
[21] C. Bouras, A. Kollia, and A. Papazois. 2017. SDN NFV in 5G: Advancements and challenges. In *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*. 107–111.
[22] Lawrence S Brakmo, Sean W O'Malley, and Larry L Peterson. 1994. TCP Vegas: New techniques for congestion detection and avoidance. In *ACM SIGCOMM*. 24–35.
[23] BroadbandNow. 2020. DIGITAL SUBSCRIBER LINE INTERNET IN THE UNITED STATES. https://broadbandnow.com/DSL. (2020).
[24] Neal Cardwell, Yuchung Cheng, C Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2016. BBR: Congestion-Based Congestion Control. *Queue* 14, 5 (2016), 20–53.
[25] Stanley CF Chan, KM Chan, Ke Liu, and Jack YB Lee. 2013. On queue length and link buffer size estimation in 3G/4G mobile data networks. *IEEE Transactions on Mobile Computing* 13, 6 (2013), 1298–1311.
[26] Hsin-Hung Cho, Chin-Feng Lai, Timothy K Shih, and Han-Chieh Chao. 2014. Integration of SDR and SDN for 5G. *IEEE Access* 2 (2014), 1196–1204.
[27] D. Clark. 1988. The Design Philosophy of the DARPA Internet Protocols. *SIGCOMM Comput. Commun. Rev.* 18, 4 (1988).
[28] Ben Cox, Jan G Laufer, Simon R Arridge, Paul C Beard, A Jan G Laufer, A Simon R Arridge, et al. 1984. Long range dependence: A review. In *Iowa State University*. Citeseer.
[29] Ning Ding, Daniel Wagner, Xiaomeng Chen, Abhinav Pathak, Y Charlie Hu, and Andrew Rice. 2013. Characterizing and modeling the impact of wireless signal strength on smartphone battery drain. *ACM SIGMETRICS Performance Evaluation Review* 41, 1 (2013), 29–40.
[30] Wikipedia enciclopedia libera. 2019. System-on-a-Chip. https://ro.wikipedia.org/wiki/System-on-a-Chip. (2019).
[31] FierceWireless. 2013. AT&T's HARP Tool Tackling Smart Cell Propagation Challenge. https://www.fiercewireless.com/tech/at-t-s-harp-tool-tackling-smart-cell-propagation-challenge. (2013).
[32] Cheng Peng Fu and Soung C Liew. 2003. TCP Veno: TCP enhancement for transmission over wireless access networks. *IEEE Journal on selected areas in communications* 21, 2 (2003), 216–228.
[33] Jim Gettys and Kathleen Nichols. 2011. Bufferbloat: Dark buffers in the internet. *Queue* 9, 11 (2011), 40–54.
[34] Arun Ghosh. 2019. Using High Tech ANTS to Unlock 5G's Potential. https://about.att.com/innovationblog/2019/06/ants_and_5g.html. (2019).
[35] Asvin Gohil, Hardik Modi, and Shobhit K Patel. 2013. 5G technology of mobile communication: A survey. In *ISSP*. IEEE.
[36] Google. 2019. Tools for Web Developers. https://developers.google.com/web/tools/chrome-devtools?hl=zh-cn. (2019).
[37] Yu Guan, Chengyuan Zheng, Xinggong Zhang, Zongming Guo, and Junchen Jiang. 2019. Pano: Optimizing 360 video streaming with a better understanding of quality perception. In *ACM SIGCOMM*.
[38] Yihua Guo, Feng Qian, Qi Alfred Chen, Zhuoqing Morley Mao, and Subhabrata Sen. 2016. Understanding on-device bufferbloat for cellular upload. In *ACM IMC*.
[39] Sangtae Ha, Injong Rhee, and Lisong Xu. 2008. CUBIC: a new TCP-friendly high-speed TCP variant. *ACM SIGOPS operating systems review* 42, 5 (2008), 64–74.
[40] Fourat Haider, Cheng-Xiang Wang, Harald Haas, Dongfeng Yuan, Haiming Wang, Xiqi Gao, Xiao-Hu You, and Erol Hepsaydir. 2011. Spectral efficiency analysis of mobile femtocell based cellular systems. In *IEEE ICCT*.
[41] Pan Hu, Pengyu Zhang, Mohammad Rostami, and Deepak Ganesan. 2016. Braidio: An integrated active-passive radio for mobile devices with asymmetric energy budgets. In *ACM SIGCOMM*.
[42] Junxian Huang, Feng Qian, Alexandre Gerber, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck. 2012. A close examination of performance and power characteristics of 4G LTE networks. In *ACM MobiSys*.
[43] Junxian Huang, Feng Qian, Yihua Guo, Yuanyuan Zhou, Qiang Xu, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck. 2013. An in-depth study of LTE: effect of network protocol and application behavior on performance. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 363–374.
[44] Ehab Ahmed Ibrahim, MRM Rizk, and Ehab F Badran. 2015. Study of lte-r x2 handover based on a3 event algorithm using matlab. In *2015 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 1155–1159.
[45] ifeng net. 2019. From NSA to SA: 5G network evolution roadmap. https://tech.ifeng.com/c/7onp0cB69lw. (2019).
[46] Haiqing Jiang, Zeyu Liu, Yaogong Wang, Kyunghan Lee, and Injong Rhee. 2012. Understanding bufferbloat in cellular networks. In *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*. 1–6.
[47] Yurong Jiang, Lenin Ravindranath Sivalingam, Suman Nath, and Ramesh Govindan. 2016. WebPerf: Evaluating what-if scenarios for cloud-hosted web applications. In *ACM SIGCOMM*.
[48] jocelyndang. 2017. Battery Historian is a tool to analyze battery consumers using Android "bugreport" files. https://github.com/google/battery-historian. (2017).
[49] kevinkong. 2019. Android performance test tool-CPU,memory,network traffic,starting time,battery current and status. https://github.com/NetEase/Emmagee. (2019).
[50] Tarmo Koppel, Andrei Shishkin, Heldur Haldre, Nikolajs Toropovs, Inese Vilcane, and Piia Tint. 2017. Reflection and Transmission Properties of Common Construction Materials at 2.4 GHz Frequency. *Energy Procedia* 113 (2017), 158 – 165. https://doi.org/10.1016/j.egypro.2017.04.045 International Scientific Conference "Environmental and Climate Technologies", CONECT 2016, 12-14 October 2016, Riga, Latvia.
[51] Swarun Kumar, Ezzeldin Hamed, Dina Katabi, and Li Erran Li. 2014. LTE Radio Analytics Made Easy and Accessible. In *Proceedings of ACM SIGCOMM*.
[52] Juho Lee and Yongjun Kwak. 2016. 5G standard development: technology and roadmap. *Signal Processing for 5G* (2016).
[53] Li Li, Ke Xu, Tong Li, Kai Zheng, Chunyi Peng, Dan Wang, Xiangxiang Wang, Meng Shen, and Rashid Mijumbi. 2018. A measurement study on multi-path tcp with multiple cellular carriers on high speed rails. In *ACM SIGCOMM*.
[54] 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 *ACM SIGMETRICS*.
[55] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Haotian Deng, Jiayao Li, and Tao Wang. 2017. Mobileinsight: Analyzing cellular network information on smartphones. *GetMobile* 21, 1 (2017), 39–42.
[56] Ming Liu, Tianyi Cui, Henry Schuh, Arvind Krishnamurthy, Simon Peter, and Karan Gupta. 2019. Offloading distributed applications onto smartNICs using iPipe. In *ACM SIGCOMM*.
[57] Lu Lu, Geoffrey Ye Li, A Lee Swindlehurst, Alexei Ashikhmin, and Rui Zhang. 2014. An overview of massive MIMO: Benefits and challenges. *IEEE journal of selected topics in signal processing* 8, 5 (2014), 742–758.
[58] Bruce A. Mah. 2019. iperf 3.7. https://github.com/esnet/iperf/releases/tag/3.7. (2019).
[59] monsoon company. 2019. A battery used to measure a user's electric quantity. https://www.msoon.com/online-store. (2019).
[60] Arvind Narayanan, Jason Carpenter, Eman Ramadan, Qingxu Liu, Yu Liu, Feng Qian, and Zhi-Li Zhang. 2019. A First Measurement Study of Commercial mmWave 5G Performance on Smartphones. *arXiv preprint arXiv:1909.07532* (2019).
[61] 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 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 894–905. https://doi.org/10.1145/3366423.3380169

[62] Navid Nikaein, Mahesh K Marina, Saravana Manickam, Alex Dawson, Raymond Knopp, and Christian Bonnet. 2014. OpenAirInterface: A flexible platform for 5G research. *ACM SIGCOMM Computer Communication Review* 44, 5 (2014), 33–38.

[63] Ashkan Nikravesh, Hongyi Yao, Shichang Xu, David Choffnes, and Z Morley Mao. 2015. Mobilyzer: An open platform for controllable mobile network measurements. In *ACM MobiSys*.

[64] HUAWEI office. 2019. HUAWEI 5G CPE Pro. https://consumer.huawei.com/cn/routers/5g-cpe-pro/. (2019).

[65] ZTE office. 2019. ZTE Axon10 pro. https://www.ztedevices.com/cn/product/smart-phone/axon_10s_pro. (2019). This is the first commercial 5G device in China.

[66] 3GPP Org. 2019. 3GPP Release 15. https://www.3gpp.org/release-15. (2019). Update of April 26, 2019.

[67] Mugen Peng, Yong Li, Zhongyuan Zhao, and Chonggang Wang. 2015. System architecture and key technologies for 5G heterogeneous cloud radio access networks. *IEEE network* 29, 2 (2015), 6–14.

[68] piaobozaizai. 2020. A series of 5G measurement tools and dataset. https://github.com/piaobozaizai/5G_measurement. (2020).

[69] Qualcomm. 2019. A Global Perspective of 5G Network Performance. https://www.qualcomm.com/media/documents/files/signals-research-group-s-5g-benchmark-study.pdf. (2019).

[70] Qualcomm.org. 2020. Snapdragon 855 Mobile Platform. https://www.qualcomm.com/products/snapdragon-855-mobile-platform. (2020).

[71] Gaurav Raina, Don Towsley, and Damon Wischik. 2005. Part II: Control theory for buffer sizing. *ACM SIGCOMM CCR* 35, 3 (2005), 79–82.

[72] Ignacio Rodriguez, Huan C Nguyen, Niels TK Jørgensen, Troels B Sørensen, Jan Elling, Morten B Gentsch, and Preben Mogensen. 2013. Path loss validation for urban micro cell scenarios at 3.5 GHz compared to 1.9 GHz. In *IEEE GLOBECOM*.

[73] Ghoul Smail and Jia Weijia. 2017. Techno-economic analysis and prediction for the deployment of 5G mobile network. In *2017 20th Conference on innovations in clouds, internet and networks (ICIN)*. IEEE, 9–16.

[74] Marco Sousa, André Martins, and Pedro Vieira. 2016. Tracking Down High Interference and Low Coverage in 3G/4G Radio Networks Using Automatic RF Measurement Extraction. In *International Conference on E-Business and Telecommunications*. Springer, 401–425.

[75] Zhaowei Tan, Yuanjie Li, Qianru Li, Zhehui Zhang, Zhehan Li, and Songwu Lu. 2018. Supporting mobile VR in LTE networks: How close are we? *ACM POMACS* 2, 1 (2018), 1–31.

[76] Insta360 team. 2019. Insta360 ONEX API. https://www.insta360.com/download/insta360-onex. (2019).

[77] Juan Pedro Tomas. 2019. China Mobile already operates 50,000 5G base stations: report. https://www.rcrwireless.com/20191119/5g/china-mobile-alread-operates-50000-5g-base-stations-report. (2019).

[78] Guan-Hua Tu, Yuanjie Li, Chunyi Peng, Chi-Yu Li, Hongyi Wang, and Songwu Lu. 2014. Control-plane protocol interactions in cellular networks. *ACM SIGCOMM Computer Communication Review* 44, 4 (2014), 223–234.

[79] Cheng-Xiang Wang, Fourat Haider, Xiqi Gao, Xiao-Hu You, Yang Yang, Dongfeng Yuan, Hadi M Aggoune, Harald Haas, Simon Fletcher, and Erol Hepsaydir. 2014. Cellular architecture and key technologies for 5G wireless communication networks. *IEEE communications magazine* 52, 2 (2014), 122–130.

[80] 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 *ACM MobiCom*.

[81] Li Wang and Huaqing Wu. 2014. Fast pairing of device-to-device link underlay for spectrum sharing with cellular users. *IEEE Communications Letters* 18, 10 (2014), 1803–1806.

[82] Song Wang, Jingqi Huang, Xinyu Zhang, Hyoil Kim, and Sujit Dey. 2020. X-Array: Approximating Omnidirectional Millimeter-Wave Coverage Using an Array of Phased Arrays. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom)*. Article 5, 14 pages.

[83] Teng Wei, Anfu Zhou, and Xinyu Zhang. 2017. Facilitating Robust 60 GHz Network Deployment by Sensing Ambient Reflectors. In *Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation (NSDI'17)*. USENIX Association, USA, 213–226.

[84] Jeroen Wigard, Troels Kolding, Lars Dalsgaard, and Claudio Coletti. 2009. On the user performance of LTE UE power savings schemes with discontinuous reception in LTE. In *2009 IEEE International Conference on Communications Workshops*. IEEE, 1–5.

[85] Damon Wischik and Nick McKeown. 2005. Part I: Buffer sizes for core routers. *ACM SIGCOMM CCR* 35, 3 (2005), 75–78.

[86] Gang Wu, Chenyang Yang, Shaoqian Li, and Geoffrey Ye Li. 2015. Recent advances in energy-efficient networks and their application in 5G systems. *IEEE Wireless Communications* 22, 2 (2015), 145–151.

[87] Feng Xia, Laurence T Yang, Lizhe Wang, and Alexey Vinel. 2012. Internet of things. *International journal of communication systems* 25, 9 (2012), 1101.

[88] Xiufeng Xie and Xinyu Zhang. 2017. Poi360: Panoramic mobile video telephony over lte cellular networks. In *ACM CoNext*. ACM, 336–349.

[89] Xiufeng Xie, Xinyu Zhang, Swarun Kumar, and Li Erran Li. 2015. pistream: Physical layer informed adaptive video streaming over lte. In *ACM MobiCom*.

[90] Xiufeng Xie, Xinyu Zhang, and Shilin Zhu. 2017. Accelerating mobile web loading using cellular link information. In *ACM MobiSys*. 427–439.

[91] Yasir Zaki, Thomas Pötsch, Jay Chen, Lakshminarayanan Subramanian, and Carmelita Görg. 2015. Adaptive Congestion Control for Unpredictable Cellular Networks. In *Proc. of ACM SIGCOMM*.

[92] Qi Zhang, Jianhui Liu, and Guodong Zhao. 2018. Towards 5G enabled tactile robotic telesurgery. *arXiv preprint arXiv:1803.03586* (2018).

[93] Xu Zhang, Siddhartha Sen, Daniar Kurniawan, Haryadi Gunawi, and Junchen Jiang. 2019. E2E: embracing user heterogeneity to improve quality of experience on the web. In *ACM SIGCOMM*.

[94] Peng zhao. 2020. Last year, 13.76 million 5G mobile phones were shipped, and this year's sales may exceed 170 million. http://www.bjd.com.cn/a/202001/16/WS5e1fca44e4b0e6e583938e4d.html. (2020).

[95] Anfu Zhou, Shaoqing Xu, Song Wang, Jingqi Huang, Shaoyuan Yang, Teng Wei, Xinyu Zhang, and Huadong Ma. 2019. Robot Navigation in Radio Beam Space: Leveraging Robotic Intelligence for Seamless MmWave Network Coverage. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobihoc '19)*. Association for Computing Machinery, New York, NY, USA, 161–170. https://doi.org/10.1145/3323679.3326514

[96] Anfu Zhou, Huanhuan Zhang, Guangyuan Su, Leilei Wu, Ruoxuan Ma, Zhen Meng, Xinyu Zhang, Xiufeng Xie, Huadong Ma, and Xiaojiang Chen. 2019. Learning to Coordinate Video Codec with Transport Protocol for Mobile Video Telephony. In *The 25th Annual International Conference on Mobile Computing and Networking, MobiCom 2019, Los Cabos, Mexico, October 21-25, 2019*. ACM, 29:1–29:16. https://doi.org/10.1145/3300061.3345430

[97] Lei Zhou, Haibo Xu, Hui Tian, Youjun Gao, Lei Du, and Lan Chen. 2008. Performance analysis of power saving mechanism with adjustable DRX cycles in 3GPP LTE. In *IEE VTC*.

[98] Yibo Zhu, Zengbin Zhang, Zhinus Marzi, Chris Nelson, Upamanyu Madhow, Ben Y. Zhao, and Haitao Zheng. 2014. Demystifying 60GHz outdoor picocells. In *The 20th Annual International Conference on Mobile Computing and Networking, MobiCom'14, Maui, HI, USA, September 7-11, 2014*, Sung-Ju Lee, Ashutosh Sabharwal, and Prasun Sinha (Eds.). ACM, 5–16. https://doi.org/10.1145/2639108.2639121

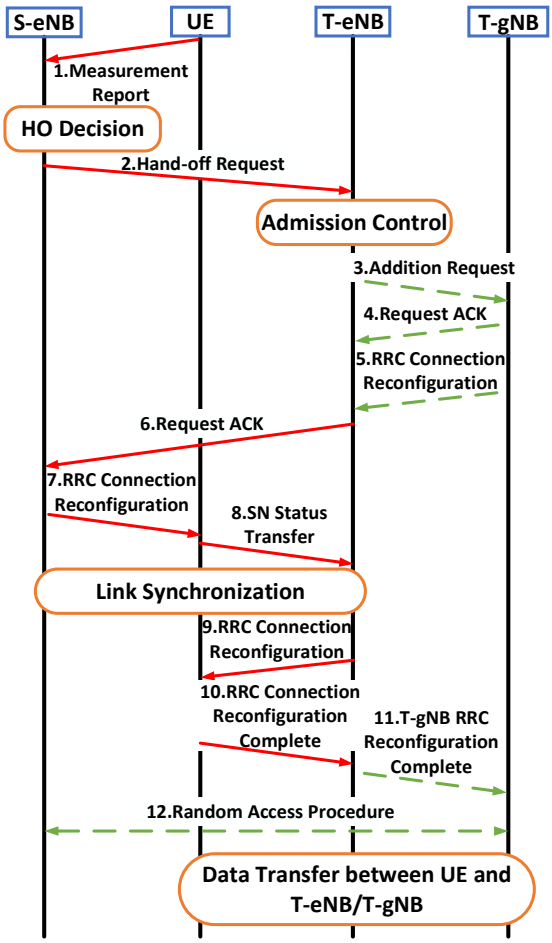

Figure 24: Hand-off procedure in the 5G NSA pattern.

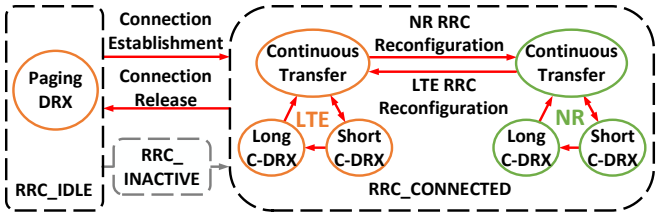

Figure 25: State machine of 5G NSA power management.

Appendices

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

A 5G HAND-OFF PROCEDURE

Here we provide some background about the types and trigger conditions of hand-off events. Once a mobile phone leaves the coverage area of one cell (serving cell), it should connect to another cell (neighboring auxiliary cell) to avoid link outage. The procedure is realized through the measurement report (MR) signaling. Specifically, a mobile device actively uploads periodic MR through RRC signaling. The MR contains the signal quality (RSRP, RSRQ, *etc*) of the serving cell and the neighboring auxiliary cell. Then, the serving cell decides whether to allow/trigger the mobile phone to hand-off. In 4G LTE or 5G NR, there are seven hand-off related events, as shown in Tab. 5. In our analysis in Sec. 3.4 and 4.3. From our measurement, A3 is the dominant event, which indicates that the cellular operator mainly uses this event to trigger hand-offs.

We proceed to detail the handoff-procedure in 5G NSA. Note that 5G radio resource is controlled by the 4G master eNB under the NSA infrastructure, so a hand-off from a 5G source gNB (referred to as S-gNB) and a target gNB (referred to as T-gNB) will undergo a more complicated signaling procedure. Fig. 24 plots the procedure that we reverse engineered based on the signaling message traces from XCAL-Mobile. To focus on the hand-off, we only present the signal exchange of the control plane, and omit the data transfer on the user plane. In particular, the whole hand-off process can be divided into four phases: UE (user equipment, *i.e.*, a smartphone in our setting) measurement, results reporting, HO decision and HO execution: *(i)* The UE periodically monitors the physical signal from the serving gNB and neighboring gNBs, and uses key performance indicators (*i.e.*, RSRP, RSRQ) to quantify the channel quality. *(ii)* Then, the UE feedbacks the measurement report to the S-eNB, which lets the S-eNB be aware of the channel quality variation in real-time, and make the HO decision correspondingly. *(iii)* After comparing itself with other neighboring cells in terms of channel quality, S-eNB will select the T-eNB among them. In particular, the selecting criterion triggered by the A3 event is that the channel quality gap $>3dB$ for a duration longer than 324ms. *(iv)* S-eNB proposes the hand-off request (by an LTE MAC RACH trigger message), and the T-eNB will enter an admission control cycle to respond the hand-off. If the request is permitted, the T-eNB will feedback an ACK to the S-eNB. Afterwards, the S-eNB initializes the hand-off operations, including releasing the source 5G radio connection, transmitting an RRC connection reconfiguration and performing link synchronization with T-eNB, *etc*. Note that for 5G-5G hand-off, the T-gNB must board a 4G master eNB, which is accomplished by the signal exchanges as marked by green dotted lines.

B 5G POWER MANAGEMENT PROCEDURE

5G NSA has two states for Radio Resource Control (RRC), *i.e.*, *RRC_IDLE* and *RRC_CONNECTED* in Rel-15 TS 36.331 and 38.331[11, 12], which operates following the state machine in Fig. 25: *(i)* A 5G radio on UEs initializes from the *RRC_IDLE* state. Once the UE has a packet to transmit, it will send a RRC connection request to the eNB and enter the *RRC_CONNECTED* state for T_{LTE_pro}. *(ii)* At *RRC_CONNECTED* state, the UE can be in one of the two modes - the LTE connection or the NR connection. Specific to the NSA architecture, a 5G radio must go through the 4G LTE mode (T_{4r_5r}) before reaching the 5G NSA NR mode (T_{NR_pro}). The switch from LTE to NR is natural, as long as the phone supports 5G. Inside the RRC state machine, the UE usually adopts the discontinuous reception (DRX) mechanism for power saving: The UE remains in sleep mode by default and only wakes up and listens to the channel for a short T_{on}, at the beginning of each periodic interval T_{cycle}. There are three DRX cycles in

Table 5: Hand-off related event description.

Type	Hand-off Event Description
A1	When the signal quality of the serving cell is higher than a fixed threshold, it will tell the mobile phone not to detect other auxiliary cells' service quality, so that the mobile phone can save energy consumption.
A2	When the signal quality of the serving cell is lower than a fixed threshold, it will tell the mobile phone to start detecting other auxiliary cells' service quality.
A3	The signal quality of the neighboring cell is continuously higher than that of the serving cell within an offset for a certain period. This is the main hand-off event.
A4	The signal quality of one neighboring cell is higher than a fixed threshold.
A5	The signal quality of the serving cell is lower than a predefined threshold (threshold1), while that of the neighboring cell is higher than the other threshold (threshold2).
B1	The signal quality of the heterogeneous radio access technology (RAT, like 4G and 5G) cell is better than a fixed threshold.
B2	The signal quality of the serving cell is lower than threshold1, while the heterogeneous RAT cell is higher than threshold2.

Table 6: Remote SPEEDTEST server details in the end-to-end delay analysis.

Server ID	Server Name	Server IP	Local City	Latitude	Longitude	Distance (km)
5145	Beijing Unicom	61.135.202.2	Beijing	39.9289	116.3883	1.67
27154	China Unicom 5G	61.181.174.254	Tianjin	39.1422	117.1767	111.65
5039	China Unicom Jinan Branch	119.164.254.58	Jinan	36.6683	116.9972	366.42
25728	China Mobile Liaoning Branch Dalian	221.180.176.102	Dalian	38.9128	121.4989	462.77
27100	Shandong CMCC 5G	120.221.94.86	Qingdao	36.1748	120.4284	553.80
5396	China Telecom Jiangsu 5G	115.169.22.130	Suzhou	31.3566	120.4682	638.00
16375	China Mobile Jilin	111.26.139.78	Changchun	43.7914	125.4784	859.32
5724	China Unicom	112.122.10.26	Hefei	31.8639	117.2808	900.06
5485	China Unicom Hubei Branch	113.57.249.2	Wuhan	30.5801	114.2734	1056.52
4690	China Unicom Lanzhou Branch Co.Ltd	180.95.155.86	Lanzhou	36.0564	103.7922	1183.99
6715	China Mobile Zhejiang 5G	112.15.227.66	Ningbo	29.8573	121.6323	1213.23
4870	Changsha Hunan Unicom Server1	220.202.152.178	Changsha	28.1792	113.1136	1341.73
5530	CCN	117.59.115.2	Chongqing	29.5628	106.5528	1459.16
4884	China Unicom Fujian	36.250.1.90	Fuzhou	26.0614	119.3061	1563.93
16398	China Mobile Guizhou	117.187.8.178	Guiyang	26.6639	106.6779	1730.12
26678	Guangzhou Unicom 5G	58.248.20.98	Guangzhou	23.1167	113.25	1890.52
5674	GX Unicom	121.31.15.130	Nanning	22.8167	108.3167	2048.98
16503	China Mobile Hainan	221.182.240.218	Haikou	19.9111	110.3301	2285.12
27575	Xinjiang Telecom Cloud	202.100.171.140	Urumqi	43.801	87.6005	2404.00
17245	China Mobile Group Xinjiang	117.190.149.118	Kashi	39.4694	76.0739	3426.37

common DRX scheme [42], *i.e.*, paging DRX, short C-DRX and long C-DRX. In our measurement, we only identify the paging DRX and long C-DRX, but no short C-DRX in the ISP's configuration. *(iii)* If no more data is received for a period of time (T_{inac}) during the *RRC_CONNECTED* state, the radio will return to the *RRC_IDLE* state after T_{tail} waiting time. Notably, Rel-15 38.331 [12] has add a new state (*RRC_INACTIVE*) to the forthcoming 5G SA architecture to save the context information of the last switch from *RRC_CONNECTED* to *RRC_IDLE*, which will enable a rapid bridge from its reversion, so as to trade off the data transfer response and more energy saving. We enumerate a list of NR radio energy related parameters, observed in XCAL-Mobile, in Tab. 7.

Table 7: Parameter of 5G NSA power management.

Parameter	Description	Value (ms)
T_{idle}	Paging DRX cycle	1280
T_{on}	On-duration timer	10
T_{LTE_pro}	LTE promotion delay	623
T_{4r_5r}	LTE to NR activity delay	1238
T_{NR_pro}	NR promotion delay	1681
T_{inac}	DRX inactivity timer	80 / 100
T_{long}	Long C-DRX cycle	320
T_{tail}	4G/5G traffic pattern tail cycle	10720 / 21440

C INFORMATION OF SERVERS USED IN THE END-TO-END DELAY MEASUREMENT

We select 20 Internet servers nationwide for the end-to-end delay analysis in Sec. 4.4. The servers (Tab. 6) belong to Ookla.Speedtest [5], but allow ICMP/UDP probing from our end-devices.