Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A comparative measurement study of cross-layer 5G performance under different mobility scenarios☆

Jiahai Hu [a], Lin Wang [b], Jing Wu [a], Qiangyu Pei [a], Fangming Liu [a,c,*], Bo Li [d]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[b] Paderborn University, Paderborn, Germany
[c] Peng Cheng Laboratory, Shenzhen, China
[d] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

Keywords:
5G
Network measurement
Mobility
Quality of experience

ABSTRACT

The 5G technology is expected to revolutionize various applications with stringent latency and throughput requirements, such as augmented reality and cloud gaming. Despite the rapid 5G deployment, it is still a puzzle whether current commercial 5G networks can meet the strict requirements and deliver the expected quality of experience (QoE) of these applications. Especially in mobile scenarios, as user mobility (e.g., walking and driving) plays a critical role in both network performance and application QoE, it becomes more challenging to provide high performance stably and continuously. To solve this puzzle, in this paper, we present a comprehensive cross-layer measurement study of current commercial 5G networks under five mobility scenarios typically seen in our daily lives. Specifically, under these mobility scenarios, we cover (1) the impact of physical layer metrics on network performance, (2) general network performance at the network layer, (3) comparison of four congestion control algorithms at the transport layer, and (4) application QoE at the application layer. Our measurement results show that the achievable network performance and application QoE under current commercial 5G networks falls behind expectations. We further reveal some insights that could be leveraged to improve the QoE of these applications under mobility scenarios.

1. Introduction

With the evolution of mobile communication technology, the 5th generation (5G) mobile networks are emerging to connect everything from personal devices to industrial machines. The global 5G connections have reached 540 million by 2021, which doubles that in 2020 [1], showing the rapid deployment of 5G. In Germany, half of the territory has been covered by 5G by 2021 [2]. The promised properties of 5G, including low latency, high throughput, and high reliability, are expected to facilitate various modern applications [3] such as augmented reality (AR) [4] and cloud gaming [5]. These applications typically require extremely high network performance which the current 4G/LTE technology falls short of.

To support these typical applications, 5G is supposed to provide high network performance with *stability* and *continuity*. Stability refers to the ability to maintain consistent performance temporally (e.g., steady throughput, low jitter). This is critical to real-time applications like cloud gaming where instantaneous high delay can cause unexpected gaming actions that have a long-term effect on the game, leading to poor user experience. Continuity refers to the ability to avoid service disruption spatially (e.g., due to blockage or connection handovers). This is important for immersive applications like AR where connection hiccups can cause uncomfortable feelings like dizziness. Delivering stability and continuity besides high performance in 5G is fundamental to enabling these applications with high QoE, while 5G is more susceptible (than 4G/LTE) to blockage and attenuation by nature and suffers more frequent handovers due to the smaller coverage range of 5G base stations.

Recently, there have been several measurement studies on the performance of both mmWave 5G networks [6–10] and sub-6 GHz 5G networks [11–13]. However, these measurement studies are conducted either in stationary scenarios or in limited mobility scenarios (e.g., walking or driving), without covering the 5G network performance and

☆ This work was supported in part by the National Key Research & Development (R&D) Plan under Grant 2022YFB4501703, and the Major Key Project of PCL under Grant PCL2024A06 and PCL2022A05, and in part by the Shenzhen Science and Technology Program under Grant RCJC20231211085918010. Lin Wang was supported in part by DFG Collaborative Research Center 1053 MAKI B2. Bo Li was supported in part by a RGC RIF grant under the contract R6021-20, and RGC GRF grants under the contracts 16209120, 16200221 and 16207922.

* Corresponding author at: School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China.
E-mail address: fangminghk@gmail.com (F. Liu).

https://doi.org/10.1016/j.comnet.2024.110952
Received 7 May 2024; Received in revised form 15 November 2024; Accepted 21 November 2024

Table 1
Measurement dataset statistics.

Dataset Statistics	
Time span	10 months
Total travel distance	2808 km+
Total data usage	1153.6 GB
# of mobility scenarios	5
# of congestion control algorithms	4
# of applications	3

the QoE of typical 5G applications under diverse mobility scenarios (A more detailed comparison is provided in Section 2). Therefore, it remains a puzzle whether the current 5G networks can consistently provide high performance and meet the requirements of typical 5G applications under diverse mobility scenarios.

This paper fills this gap by presenting a comprehensive cross-layer measurement study of the current commercial 5G networks under various mobility scenarios. Our measurement study spanned 10 months and consumed 1153.6 GB traffic data, and the total travel distance of the measurement exceeds 2808 km, as shown in Table 1. Our focus is on the impact of the user mobility pattern on the network performance, stability, and continuity at different layers of the protocol stack. To capture user mobility, we define four metrics namely real-time moving speed, speed variation, connectivity density, and varying ambient environment along user movement. Based on these metrics, we consider five types of mobility patterns: walking, biking, riding a bus, riding a tram, and driving (or riding in a car). Under these mobility patterns, we (1) study the impact of *physical* layer metrics and mobility factors on network performance, including signal strength, obstruction, distance from the base station, and moving velocity, (2) investigate the general performance at the *network* layer, including round-trip time (RTT), up-/down-link throughput, and packet loss rate (PLR), (3) study the impact of transport protocols on network performance, where we compare four popular congestion control algorithms (CCA) at the *transport* layer, and (4) evaluate the QoE of three typical 5G applications at the *application* layer.

Based on our measurement results, we present three insights that could help both application developers and mobile service providers improve the QoE of these applications in the future: (1) Optimal route planning should consider not only real-time road traffic but also network conditions along the path when users are enjoying 5G applications. For example, a path with better network performance but with higher road traffic may be preferred by people streaming videos or playing cloud games if the road traffic is within an acceptable range. (2) In addition to real-time network performance, applications can leverage user mobility traces to predict network performance and further adapt to network variations in advance. (3) Considering the remaining high transmission latency of the backbone network, it is necessary to choose a suitable nearby edge server to provide services based on users' location for a lower overall end-to-end latency.

In summary, this paper makes the following contributions:

- We identify two additional challenges imposed by typical applications and user mobility for 5G. To meet the application QoE, 5G networks should provide temporal stability and spatial continuity besides high network performance.
- We define four key metrics to capture mobility and conduct extensive measurements to reveal their varying impacts on the general network performance. Based on these metrics, we consider five types of mobility patterns.
- We perform the first comprehensive cross-layer measurement study of 5G networks under diverse mobility patterns. Our cross-layer 5G measurements include 5G performance at the network layer, transport layer (four CCAs), and application layer (three typical 5G applications).
- Based on the measurement results, we reveal three insights on improving the application QoE, including network-aware route planning, handover-aware application adaptation, and mobility-aware edge resource adaptation.

The rest of the paper is organized as follows. In Section 3, we briefly introduce the emerging 5G technology as well as typical applications and outdoor mobility patterns for 5G. Next, we present our measurement methodology in Section 4. Then, we evaluate the basic 5G network performance (e.g., RTT, throughput, and PLR) and application performance indicated by QoE under different mobilities, in Sections 5 and 6, respectively. After that, we discuss some recent works related to 5G mobile network measurements and cellular network measurements under mobility in Section 2. Finally, we reveal three insights for improving QoE in Section 7 and conclude the paper in Section 9.

2. Related work

5G network performance under mobility. Narayanan et al. conducted the first measurement study of a commercial mmWave 5G network in [6], but they only present two throughput traces during walking and driving respectively. Lumos5G [9] identified key UE-side mobility factors that affect 5G performance and then utilized these factors to predict mmWave 5G throughput. Hassan et al. carried out a systematic analysis to uncover the handover mechanisms employed by 5G carriers [8]. In summary, these studies mainly focus on mmWave 5G, which is quite distinct from the sub-6 GHz 5G that we measure due to the extremely high frequency of mmWave. Xu et al. conducted measurements to characterize TCP performance in sub-6 Ghz NSA 5G under low mobility (walking/bicycling) [11]. In contrast, we measure the network performance of both NSA and SA 5G with four congestion control algorithms under five mobility patterns. Also, we explore the respective impacts of various mobility factors on network performance. Ghoshal et al. performed a comprehensive measurement study under driving scenario [14], while we focus on the impact of diverse mobility patterns and mobility factors on network performance. Pan et al. conducted the first measurement study on a high-speed railway to reveal 5G performance in extreme mobility (~300 km/h), while we focus on mobility patterns below 100 km/h.

5G application performance under mobility. Narayanan et al. explored the impact of mmWave 5G on mobile application QoE [7] for traditional applications such as web browsing and video streaming. Specifically, they collected throughput traces during walking/driving and used trace-driven emulation to investigate the QoE of adaptive video streaming, which only considers throughput regardless of latency and packet loss. In contrast, we measure the QoE of adaptive video streaming over commercial sub-6 GHz 5G networks in complex real-world deployments under *diverse mobility patterns*. Moreover, we explore the QoE of *typical 5G applications* like cloud-based AR and cloud gaming under various mobility patterns. Hassan et al. studied the impact of handover mechanisms on application QoE [8], while Ghoshal et al. performed a measurement study on 5G applications under driving scenario [14]. In contrast, we focus on the impact of diverse mobility patterns and factors on the QoE of different applications. Ghoshal et al. also studied AR QoE in mmWave 5G during walking and driving [15], while we study AR QoE in sub-6 GHz 5G under five mobility scenarios. Khan et al. focus on the delay of 1080p live video streaming on moving vehicles [16], while we study the QoE of 8K on-demand adaptive bitrate streaming under various mobility patterns.

We believe our measurement study can fill the gap of both network performance and application QoE in sub-6 GHz 5G network under various mobility scenarios and varying mobility factors.

3. Background

This section introduces the 5G technology and its current deployments, typical 5G applications, and typical mobility patterns.

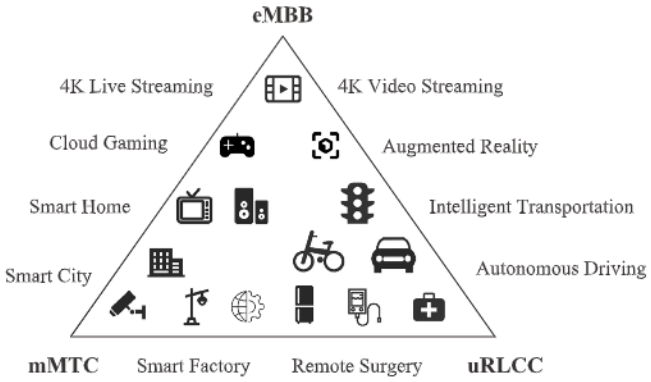


Fig. 1. Typical 5G application use cases.

3.1. 5G technology

The 5G technology promises to deliver high bandwidth, ultra-low latency, ultra-high reliability, and massive device connectivity [17]. In view of these outstanding advantages, 5G has been widely deployed all around the world lately. According to VIAVI's report, commercial 5G networks have been available in 1,947 cities at the beginning of 2022, an increase of about 50% from a year ago [18].

5G outperforms 4G by using higher frequency bands and leveraging emerging technologies such as massive multiple-input multiple-output (MIMO) [19]. Compared to 2 GHz used by 4G/LTE, there are two widely-used 5G frequency bands, i.e., sub-6 GHz and mmWave. Most countries deploy sub-6 GHz 5G at the beginning in large part due to its much higher coverage range than mmWave, while the US mainly focuses on mmWave [20]. Considering the capital expenditures and deployment cycles, early deployment of commercial 5G networks is non-standalone (NSA), which means 5G networks are still built atop existing 4G infrastructure, making 5G unable to give full play to its advantages. Under the NSA architecture, 5G is only utilized for the data plane, with the help of 4G for control plane operations [21]. In contrast, the standalone (SA) architecture is fully independent of the 4G infrastructure with potentially better performance [22]. Recently, operators are increasingly deploying and testing 5G SA networks. For example, the three biggest telecoms in China, namely China Mobile, China Telecom, and China Unicom, have all launched 5G SA networks [23].

Although 5G has developed rapidly recently, it is still a puzzle whether the actual perceived performance of current commercial 5G networks can consistently meet the requirements of typical 5G applications under different mobility patterns as we have expected.

3.2. Typical 5G applications

Now, we introduce typical 5G Applications evolving with 5G networks. Fig. 1 shows three sets of typical 5G use cases [24]. Specifically, enhanced mobile broadband (eMBB) is for high-bandwidth scenarios, supporting applications like AR, virtual reality (VR), cloud gaming, and 4K video streaming. Massive machine type communication (mMTC) provides connections to large numbers of devices, supporting applications like smart cities and smart factories. Ultra reliable low latency communication (uRLLC) aims to achieve ultra-low latency, supporting mission-critical applications like remote surgery and autonomous driving. We select the following three applications as representative examples.

Cloud gaming. As an emerging online gaming paradigm, cloud gaming collects user control actions from user devices (i.e., clients) to a cloud server in a timely manner, renders each frame sequentially on the server instead of the clients, and streams the encoded frames back to the clients via the network. Such a real-time streaming system needs sufficient bandwidth (especially for 4K/8K game streaming) as well as ultra-low latency to ensure timely control actions and a good experience for game players.

Cloud-based AR. Nowadays, most existing AR devices still lack the ability to detect and recognize complex objects in the real world [25]. It is prohibitive to execute large complex deep neural networks on AR devices due to the big mismatch between the excessive computations needed and the limited processing power and battery life of the AR device. To address this issue, cloud-based AR systems offload object detection tasks from AR devices to a cloud server [26]. However, guaranteeing low end-to-end latency without sacrificing detection accuracy is still a challenging problem due to the network overhead. 5G is expected to address this challenge by reducing the network latency significantly.

Video streaming. Video streaming is one of the most popular applications on user devices. With 5G's high bandwidth, users are able to watch 4K and even 8K streaming videos. Moreover, adaptive bitrate (ABR) algorithms are widely used to improve the QoE of video streaming applications. However, the performance of the ABR algorithm may be sabotaged by the high variation of 5G's throughput, especially in non-stationary scenarios.

The QoE of 5G applications depends on multiple factors including the 5G access network quality, backbone network quality, server performance, and device performance. According to our initial measurement results, the 5G networks (i.e., the access networks and backbone network) still contribute to a significant portion of the end-to-end latency of applications in most cases. Therefore, it is urgent to conduct a comprehensive measurement study to examine the application QoE under current 5G networks.

3.3. Outdoor mobility patterns

5G users may use applications in various scenarios (e.g., indoor/outdoor, static/mobile). Since 5G owns much higher frequency bands that are more sensitive to the environment, outdoor mobility could induce severe network performance degradation. Common mobility patterns of 5G users include walking, biking, riding a bus or a tram (light rail), and driving (or riding in a car).

These mobility patterns have different characteristics, which can mainly be captured by the following metrics: average moving speed, moving speed variation, connectivity density, and environment. These metrics play a key role in network performance. Firstly, the moving speed and its variation influence the network performance with respect to bit error rate (due to the fast signal fading) and handover frequency [27]. Handovers happen when a user moves away from the area covered by one base station (BS) and enters the area covered by another one, causing network connection hiccups. Secondly, high connectivity density may limit the maximum available bandwidth since users have to share the bandwidth with others while being connected to the same BS. Thirdly, the environment includes the surroundings (e.g., buildings, tunnels, trees, vehicle body shells), the distance between the user equipment (UE) and the BS, etc., which impact network performance by causing signal obstruction and attenuation. We will investigate such impacts in detail in Section 5.1.

For different mobility patterns, the typical values of these metrics vary as listed in Table 2. We use the GeoLife [28–30] and DR-Train [31] datasets to estimate the average moving speed and its variation under different mobility patterns by calculating the mean and standard deviation of moving speed. Additionally, buses and cars have a body shell of medium thickness, while trams usually have a thicker shell. Moreover, public transports usually have a high connectivity density, and in our city, there are more people taking the bus than the tram. The differences in these metrics under different mobility patterns lead to diverse effects on 5G network performance. For instance, a thicker body shell may present more obstacles to 5G signals, leading to potential signal loss or degradation. Also, higher moving speed may cause more

Table 2
Typical outdoor mobility patterns.

Mobility pattern	Speed (km/h)	Speed variation	Body shell	Connectivity density
Walking	4	2.6	none	low
Biking	23	11.1	none	low
Bus	20	17.6	medium	high
Tram	24	17.7	thick	moderate
Driving	36	29.7	medium	low

frequent handover. In short, the above metrics can impact network performance in different ways while users are moving.

In a nutshell, different mobility patterns have varying impacts on network performance and thus application QoE. In addition to providing high performance, it is critical for 5G networks to maintain both temporal stability and spatial continuity in the presence of mobility. The questions of whether 5G networks can already satisfy the requirements of typical applications under different mobility patterns, and how far we are from the expectation motivate us to perform the comprehensive measurement study on cross-layer performances of commercial 5G networks under five mobility patterns.

4. Measurement methodology

We now describe our measurement methodology and tools.

5G networks. Our measurement study is conducted in a densely populated city over the 5G networks of two large mobile service providers, denoted by P_M and P_U respectively.[1] Specifically, P_M offers 5G networks with the 4.8 GHz radio while P_U uses the 3.5 GHz radio. The mmWave 5G networks have not been deployed in the country, and hence our study only focuses on sub-6 GHz 5G networks. For the deployment architecture, P_M deploys 5G in NSA architecture while P_U in SA architecture. To know if any 5G is ready to support typical applications, we choose the more advanced provider P_U for the studies on application QoE because of its generally lower network latency and higher network throughput (Section 5.2). Additionally, for 4G networks, P_M offers the 1.9 GHz radio while P_U uses the 1.8 GHz radio.

User equipment (UE). We conduct measurements with a Redmi K30 Pro smartphone, a mid-end smartphone equipped with an octa-core CPU, 6 GB DRAM, and a Qualcomm Snapdragon 865 System-on-Chip (SoC). The SoC uses a separate 5G modem (i.e., X55 modem) to provide higher throughput, yet at the expense of more power consumption. The X55 modem supports both sub-6 GHz and mmWave frequency bands, and both NA and NSA modes. Before large-scale measurements in the wild, we conducted an initial measurement experiment to study whether different phone models would significantly affect network performance. Specifically, we compared the performance of Redmi K30 Pro with the other three phone models, i.e., Huawei Mate 30 Pro (Kirin 990 chip with Balong 5000 5G Modem), iPhone 13 (A15 chip with Qualcomm X60 5G modem), and Meizu 20 Infinity (Snapdragon 8 Gen 2 chip with Qualcomm X70 modem) using Speedtest [32]. The results indicate that the first two phone models present similar performance to that of Redmi K30 Pro, with a difference of within 5%. Additionally, Meizu 20 Infinity shows similar latency as well and a 9.6% higher average throughput. In view of their comparable performance, we choose Redmi K30 Pro for further measurements due to its convenient root privilege access and client deployment. For comparison between 4G and 5G networks, we conduct measurements with two smartphones of the same model (i.e., Redmi K30 Pro) simultaneously, one with 5G service enabled while the other with 4G service only.

Server selection. We rent the nearest available server (with 4 vCPUs, 8 GB DRAM, and 1Gbps bandwidth) from a major cloud provider to serve as the backend for our measurements, which is 16 hops away from our clients. The RTT for wired connections to the server is around 22 ms. By comparative experiments, we ensure that the cloud server used for our measurements is not the bandwidth bottleneck of an end-to-end path. Specifically, we first utilize Speedtest to measure the maximum achievable 5G throughput with their local edge server. The Speedtest results show that the throughput is comparable to what we measure with our server, which is far below 1Gbps. Therefore, we can confirm the cloud server is not the bottleneck. Compared to the servers provided by Speedtest, our cloud server can report diverse network performance metrics and capture packets so that we can perform a more detailed analysis. Furthermore, it is convenient to achieve fine-grained control of network measurement parameters by changing the server settings such as the congestion control algorithms adopted by the OS kernel. Note that for deploying cloud gaming and cloud-based AR applications, we also rent a GPU server with 8 vCPUs, 32 GB DRAM, and an NVIDIA T4 GPU with 16 GB memory at the same location from the same cloud provider, to ensure sufficient capability of graphics processing and parallel computing.

Measurement tools. We use `iperf3` [33] to measure the network throughput and save the result log reported by `iperf` on both the client (i.e., the smartphone) and server, including the total transferred bytes, throughput, number of retransmitted packets, and congestion window size. The RTT is measured by the `ping` tool. Noting that the ping tool is only used for measuring general 4G/5G latency at the network layer, which follows other state-of-the-art 5G measurement works [6,8,14]. We do not use the ping tool to measure the RTT of applications. Instead, we follow state-of-the-art measurement works on application QoE and use application-specific metrics to measure application latency perceived by users, e.g., response delay for cloud gaming. We use the `tcpdump` tool [34] to capture packets on both sides of the client and server for low-level root cause analysis.

Application deployment. To be able to collect server logs and control application settings for comparative experiments, we leverage custom-built tools to examine the QoE of the three typical 5G applications introduced in Section 3.2: (1) The *cloud gaming* application is developed with a popular open-source cloud gaming framework GamingAnywhere [35], which utilizes the real-time streaming protocol (RTSP) to stream the encoded frames from the server to the client. We build an Android application to simulate periodic touches on the smartphone screen, which is running in the background when cloud gaming is presented in the foreground. To obtain the QoE metric of cloud gaming, i.e., response delay, we also develop a script to record the screen and analyze each frame extracted from the recorded screen video with FFmpeg [36] and OpenCV [37]. (2) For *cloud-based AR*, we develop a server program and an Android client application by ourselves. The client continuously sends frames to the server at a rate of 10 FPS, and the server processes the frame with the YOLO object detection model [38]. We obtain the QoE metrics of AR based on the interval between starting sending a frame and receiving the detection result. (3) For *8K video streaming*, we build a web server with the open source DASH.js framework [39]. We obtain the QoE metrics under different adaptive bitrate (ABR) algorithms directly from the log file available from the framework.

Measurement under mobility. For a fair comparison of different mobility patterns, all measurement experiments are conducted on the same road (three BSs along the straight road) with a smartphone in hand, to ensure the same network environment across different mobility patterns. It is worth noting that finding such a road for all

[1] The names of the providers are not given for anonymization.

mobility patterns with 5G coverage is non-trivial, due to distinct fixed routes served by trams and buses, while other mobility patterns are more flexible. Therefore, we carefully identify a shared route of trams and buses that is as long as possible (about 2 km, compared to a 0.5 km × 0.92 km area in [11]), and then conduct the comparative measurement study on this route for all mobility patterns.

For all mobility patterns, we move back and forth to repeat the measurement experiments at least three times, and each measurement experiment lasts for five minutes. We also measure network performance when the UE remains stationary, which is considered our baseline for comparing results measured in different mobility patterns. To ensure a fair comparison with other mobility patterns, we randomly select five locations on the road in the stationary case and conduct measurement experiments twice on each location while standing still.

It is worth noting that the type of application may have an impact on user mobility [40]. In most mobility scenarios, users are typically passengers (i.e., on a bike, bus, tram, car). Therefore, the mobility patterns in these scenarios are only affected by road traffic and their inherent characteristics. For users who are walking, their mobility pattern may be influenced by the type of application. For example, when wearing a head-mounted display to experience immersive applications, users tend to have shorter stride length, greater stance time, and higher speed variability [41]. Instead, our measurement study focuses on applications for smartphones, which typically do not have significant impacts on user mobility. Therefore, our user mobility follows conventional behaviors (3.9 km/h of average speed and 0.75 m of stride length for walking in our study). We are interested in exploring the QoE of VR applications under diverse mobility patterns in the future.

5. Network performance

While user equipment (UE) is moving, its surrounding network environment keeps changing. In this section, we examine how different mobility patterns, including walking, biking, driving, bus, and tram, affect the 5G network performance. Specifically, we first explore how the elementary mobility features, including distance from the base station, signal strength, obstruction, and moving velocity, influence 5G network performance such as throughput and round-trip time (RTT). Then, we measure the overall network performance for both 4G and 5G under the above-mentioned mobility patterns. Finally, we study the impacts of congestion control algorithms on 5G network performance under various mobility patterns.

5.1. The impact of mobility

While a UE is moving, its surrounding network environment keeps changing: (1) The UE may be gradually moving away from or getting closer to the serving BS. (2) The propagation path of radio between the UE and the base station may be obstructed by an obstacle hard to penetrate. These will affect the strength/quality of the cellular signal or cause connection handovers. Besides the environment, the moving speed can also influence network performance by varying the bit error rate as well as the handover frequency. In this section, we examine how these factors affect the 5G performance. Note that the following measurement experiments are conducted in 4G/5G networks of P_U, since we find that only P_U provides 5G services at the specific locations where these experiments are conducted.

Distance from the base station. We first investigate the impact of the distance between a UE and the serving base station. We pick three spots whose distances from the base station are 100 m, 200 m, and 300 m, respectively. Each spot and the base station are in a straight line. At each spot, we measure the RTT with `ping` every second for 60 s and repeat this experiment 10 times for both 4G and 5G. It is worth noting that, to avoid unnecessary handovers when gradually increasing distance, we strive to conduct our experiment in a location where both 4G and 5G base stations are deployed as sparsely as possible.

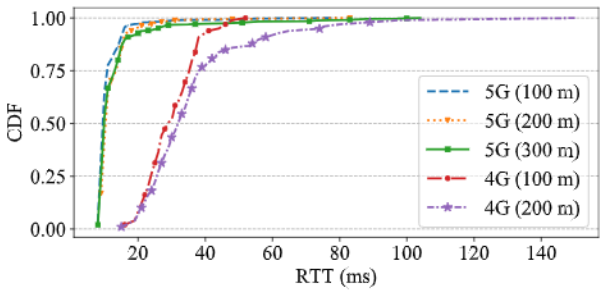


Fig. 2. Impact of distance on 5G/4G RTT.

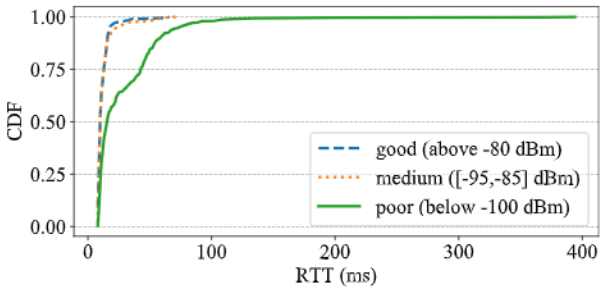


Fig. 3. Impact of signal strength on 5G RTT.

Fig. 2 plots the CDF of all measurement results. We do not plot the 4G RTTs with a 300 m distance in the figure because the connection always switches to another cell as we increase the distance to 300 m. This may seem counter-intuitive because typically 5G has a shorter coverage range and thus should have switched earlier than 4G. Theoretically, if there are only two adjacent 4G/5G base stations on the road, a shorter range of 5G networks may lead to earlier handover. However, in the real world, the current deployment of 5G base stations is not as widespread as 4G. Consequently, there are fewer potential neighbor cells for 5G to handover when compared to 4G. Furthermore, current measurement studies [8,13] reveal that even in a densely deployed area with full 5G coverage, 5G SA handover happens less frequently than 4G. This contributes to 5G's better robustness against poor signal quality and strength, leading to less likely sensed neighbor cells and reduced necessity of handovers [13].

As shown, 62.4% of the 5G RTTs are within 10 ms, and only a few results are greater than 20 ms. As the distance increases, higher RTTs appear more often. We attribute such a phenomenon to the change in signal strength. Specifically, the increase in the distance incurs signal attenuation and more interference. Thus, we later study the impact of signal strengths on network performance. 4G exhibits a similar pattern but has a 3× higher average latency. Moreover, 79.8% of the 4G RTTs range from 20 ms to 40 ms, which is unexpectedly more unstable in comparison with 5G.

Signal strength. To study the impact of signal strength on 5G RTT, we perform comparative measurements by standing near an open window. We observe that the signal strength rapidly decreases when we move away from the window. Thus, we conduct measurements at different distances away from the window (i.e., 0 m, 1 m, 3 m). The signal strengths at the three locations are good (above −80 dBm), medium ([−95,−85]dBm), and poor (below −100 dBm), respectively. We plot the 5G RTTs with different signal strengths in Fig. 3. As shown, better signal strength brings lower RTTs in general. While 5G RTTs perform similarly under good and medium signal strength, they fluctuate widely under poor signal strength. It is worth mentioning that high 5G RTTs (above 40 ms) in the case of poor signal strength perform similar distribution to 4G RTTs in Fig. 2. Also, poor signal strength leads to a long-tail latency distribution (up to 392 ms), which may hurt

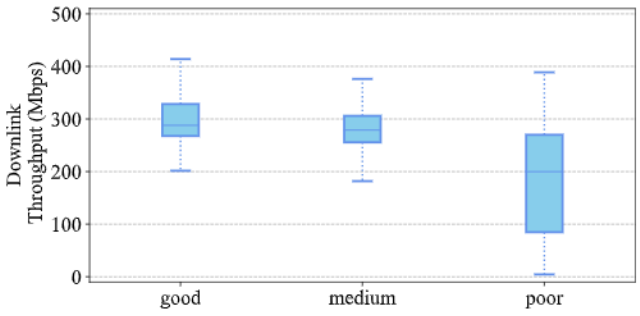


Fig. 4. Impact of signal strength on 5G throughput.

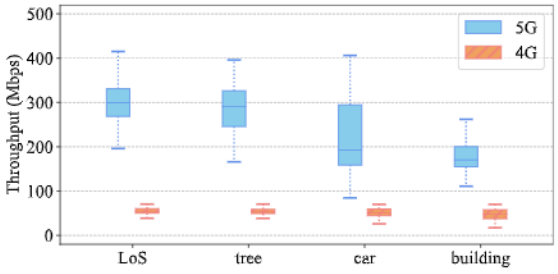


Fig. 5. Impact of obstruction on 5G/4G throughput.

the QoE of real-time applications like cloud gaming. We also study the impact of signal strength on 5G throughput at the same experiment location. The experiments are conducted in stationary state and no handovers occur during the measurement experiments. As shown in Fig. 4, we observe a generally decreasing trend in median throughput as the signal strength becomes worse. Specifically, compared to good signal strength (above −80 dBm), the median 5G throughput under medium strength ([−95,−85]dBm) only slightly decreases. However, it significantly drops under a poor signal strength (below −100 dBm).

Obstruction. We then study the impact of obstruction on 5G/4G throughput. We first select a spot with good signal quality and under a clear line-of-sight (LoS) to perform throughput measurements as a baseline. We park a car at the same spot and measure the throughput in the car (backseat). Then we measure the throughput beneath a nearby tree. As shown in Fig. 5, the 5G throughput beneath a tree is slightly lower than that under clear LoS. However, the median 5G throughput inside the car has a 34% drop compared to that under clear LoS.

We also study the impact of a building obstruction on throughput. Firstly, we conduct an extensive investigation into whether 5G service is available in a nearby building. We find that 5G service is unavailable, which indicates that 5G radio cannot penetrate buildings built with concrete. However, smartphones are able to connect to the 5G network near a transparent window or an open door. Thus, we conduct throughput measurements at the door inside the building. As shown in Fig. 5, the throughput inside the building is much smaller than that outside the building. We repeat the same throughput measurements for 4G. The results show that the impact of obstruction on 4G throughput is limited.

Moving velocity. We also study the impact of moving velocity on network performance. We measure the throughput and RTT at different velocities on a straight road. We conduct measurements while driving a car in different velocity levels, including slow (~3 km/h), medium (~20 km/h), and fast (~50 km/h). The UEs (i.e., smartphones) are placed on the car dashboard under the windshield to avoid obstruction.

We plot the CDF of the RTTs with all velocity levels. As shown in Fig. 6, in the 5G network, high RTTs appear more frequently at a higher velocity, especially at the fast level. While in 4G, we do not

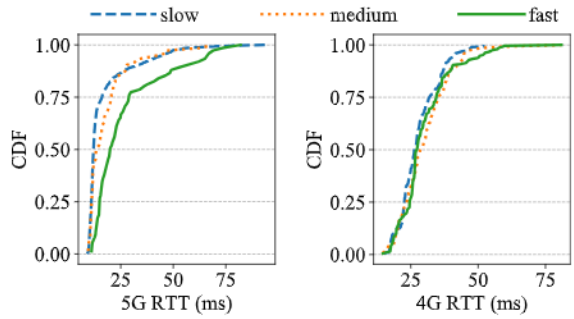


Fig. 6. Impact of moving velocity on 5G/4G RTT.

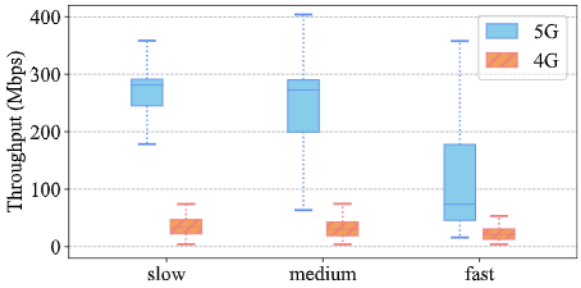


Fig. 7. Impact of moving velocity on 5G/4G throughput.

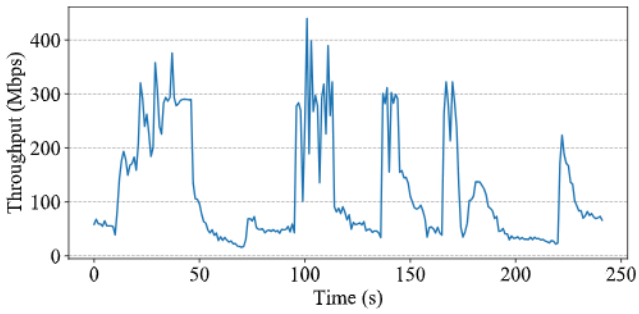


Fig. 8. 5G throughput trace snippet while driving.

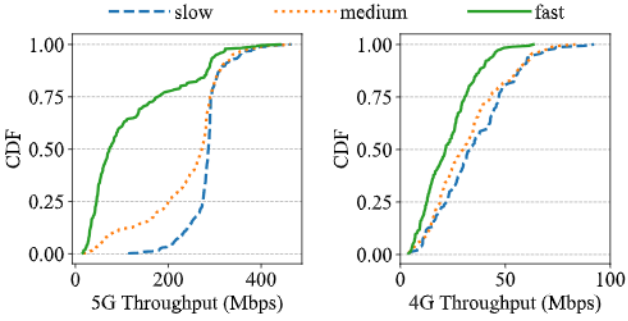


Fig. 9. CDF of 5G/4G throughput with different moving velocities.

observe such a phenomenon, which implies that 5G is more sensitive to moving velocity than 4G. Fig. 7 shows the 4G/5G throughput at different velocities. 5G throughput has a significant degradation when the moving velocity increases, while 4G throughput has little variation. Specifically, the median 5G throughput decreases by 72.7% when the user drives fast, compared with the throughput while driving at a medium speed.

We then take a further look into the throughput traces. Fig. 8 plots a four-minute trace of the downlink throughput while driving. We can

Table 3
Average throughput recovery after intra-NR handovers.

Time after handover (s)	0	1	2	5	10
Throughput Recovery (%)	59.4	64.0	91.6	101.2	102.5

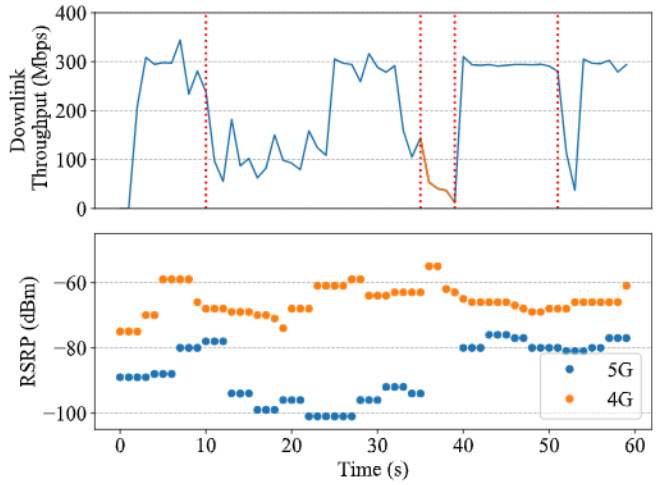


Fig. 10. A 5G throughput trace during mobility with handover events (denoted by red dotted vertical lines) and varying signal strength.

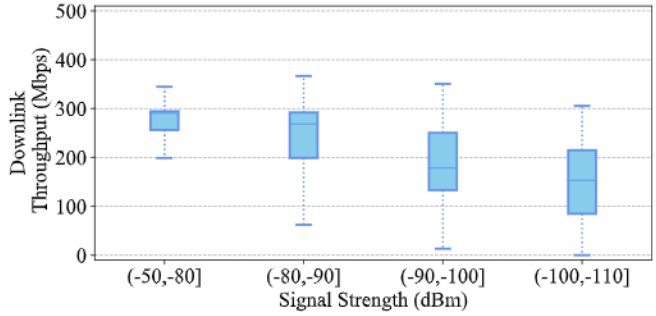


Fig. 11. Impact of signal strength on 5G throughput during mobility.

see that the 5G throughput is often less than 100Mbps, which is at the same level as the 4G throughput. Such performance degradation can be attributed to frequent handovers or sometimes even unsuccessful handovers (i.e., connection establishment failures). Fig. 9 plots the CDF of the throughput. It shows more clearly that 5G throughput is often less than 100Mbps when the user is driving at a high speed, which is even comparable to 4G's performance.

Handover. User mobility can cause handovers between base stations, which may also affect network performance. We then study its impact on 5G throughput. Fig. 10 shows a throughput trace during driving with handover events (red dotted line), and varying signal strength, i.e., RSRP. We can observe from Fig. 10 that the throughput significantly decreases when handovers occur. For example, at 10s and 51s, the handover between 5G base stations (i.e., intra-NR handover) results in a 59.3% and 59.2% throughput decrease, respectively. While at 35s, we observe 5G falls back to 4G with an inter-RAT handover, thus the throughput drops to far below 100Mbps. By calculating the average throughput drop following handovers, we find that for intra-NR handovers (5G to 5G) lead to a 40.6% decrease in throughput. In contrast, for inter-RAT handovers (5G to 4G), the throughput experiences a more substantial drop of 80.1%. From Fig. 10, we also observe that the throughput does not immediately recover after an handover. Therefore, we further analyze how quickly the throughput can recover following an intra-NR handover, as shown in Table 3. The throughput shows a subsequent recovery of 64.0%/91.6%/101.2%/102.5% after 1/2/5/10 seconds post-handover when compared to the throughput before the handover. It is worth noting that the throughput does not immediately recover after an handover due to TCP's slow start mechanism. By contrast, we cannot observe an obvious causal relationship between the throughput and varying signal strength. For instance, the throughput of 5G during 25–35s is comparable to that of 40–50s, while the former's signal strength is much lower by up to 25dBm as shown in Fig. 10. Meanwhile, the throughput during 10–35s fluctuates widely while the signal strength generally remains the same level. Therefore, a poor signal strength does not always mean a low throughput. Fig. 11 presents the 5G throughput under varying signal strengths during mobility. As shown, the average 5G throughput gradually declines with deteriorating signal strength. However, when contrasted with stationary scenarios (Fig. 4), 5G throughput suffers greater fluctuations under medium signal strengths (i.e., [−80,−100]) during mobility. This can be attributed to additional mobility factors such as handovers and speed, indicating that the throughput in mobile scenarios is influenced by a combination of various mobility factors and their intricate interplay. This suggests that relying solely on signal strength may not be sufficient for network providers to make effective resource scheduling decisions. In comparison, as handovers have a more significant impact on network performance, it is promising to improve network performance and application QoE by handover-aware methods (detailed discussion in Section 7.2).

Summary. *5G networks generally achieve lower and more stable RTT compared to 4G, along with higher throughput. Nevertheless, the performance of 5G networks can be influenced by various mobility factors. For instance, poor signal strength can lead to a long-tail latency distribution, while fast-moving velocity can cause significant throughput degradation. In these network environments, 5G performance may even deteriorate to levels comparable to 4G. Consequently, such high fluctuations can result in inconsistent application QoE.*

5.2. Performance under mobility

Notably, the wireless channel quality may affect network performance. For instance, poor wireless channel quality can degrade the data rate and increase the bit error rate. To avoid the noise from wireless signal quality variation, all our comparative measurements under different mobility patterns are carried out on a road with a generally stable Reference Signal Received Power (RSRP) level of [−80, −60]dBm.

End-to-end network throughput. We first measure the end-to-end throughput under different mobility patterns. We utilize `iperf` to transfer bulk data between the client and server with TCP. To obtain measurement results more precisely and fully utilize the available bandwidth as possible, we use multiple concurrent TCP connections (i.e., 8 connections) for throughput measurement instead of a single connection which is more sensitive to packet loss and network congestion. To avoid the impact of TCP slow-start, we discard the measurement results for the first 10 s.

We plot the 4G/5G up-/down-link throughput of provider P_M and P_U in all considered mobility patterns. As shown in Fig. 12, the 5G downlink throughput of P_M can reach up to 400Mbps, but fluctuates violently and sometimes can have extremely poor performance (i.e., even close to zero), especially when driving or taking a tram. The 4G downlink throughput has a median of 60Mbps, which is much lower than that of 5G. Fig. 13 depicts the results for P_U, where the 5G downlink throughput can reach up to 500Mbps, but also exhibits significant fluctuations. Furthermore, the downlink throughput of P_U presents a more obvious dropping trend than that of P_M as the mobility pattern changes. Specifically, the median 5G downlink throughput decreases by approximately 30% when taking a bus/tram compared to slower mobility patterns such as walking (~4 km/h) or biking (~18 km/h),

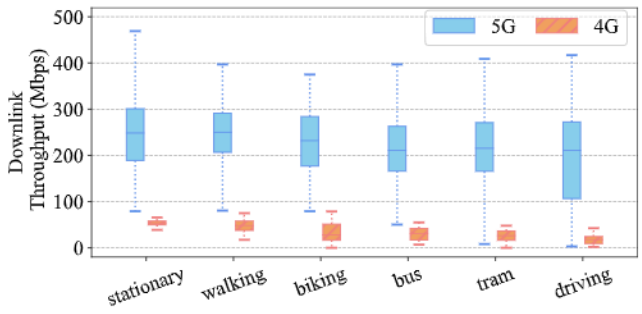


Fig. 12. Downlink throughput of provider P_M in different mobility patterns.

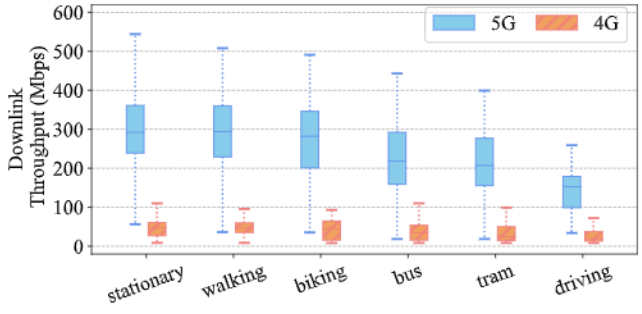


Fig. 13. Downlink throughput of provider P_U in different mobility patterns.

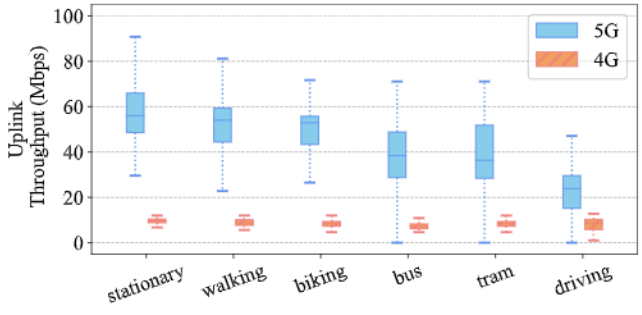


Fig. 14. Uplink throughput of provider P_M in different mobility patterns.

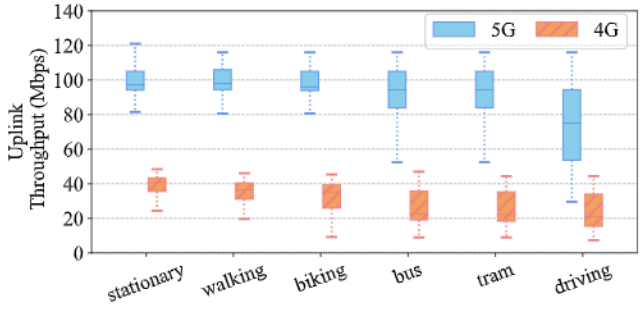


Fig. 15. Uplink throughput of provider P_U in different mobility patterns.

while the throughput drops by 45% when driving due to the higher moving velocity.

Figs. 14 and 15 show the 4G/5G uplink throughput under all the mobility patterns. From Fig. 14 we observe that the 5G uplink throughput of P_M can reach up to 90Mbps, but it fluctuates violently and sometimes becomes extremely poor, similar to the trend shown in the downlink throughput. The 4G uplink throughput has a median of 10Mbps, which is much lower than that of 5G. The results imply that 5G networks cannot provide sufficient uplink throughput for applications

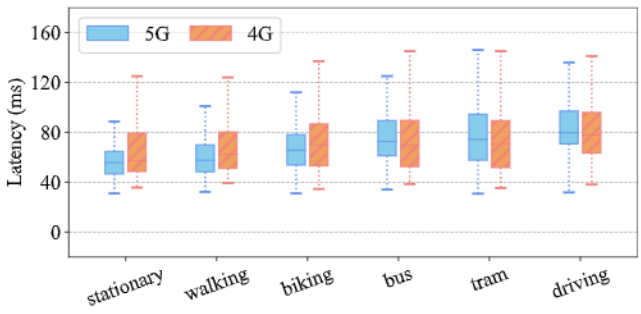


Fig. 16. End-to-end RTT of provider P_M in different mobility patterns.

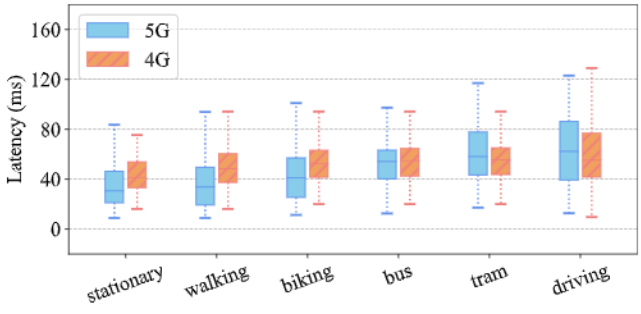


Fig. 17. First three-hop RTT of provider P_M in different mobility patterns.

that need to upload huge amounts of data, e.g., 360-degree 4K/8K live streaming as 120Mbps [42]. Also, the 5G uplink throughput tends to decrease as the mobility pattern changes. Specifically, the median drops from 55Mbps to 25Mbps when switching from the stationary scenario to the driving scenario. Consequently, this presents additional challenges for offloading machine learning (ML) models in driving scenarios, such as deep neural network (DNN) inference for autonomous driving tasks. Fig. 15 shows the results for P_U, where the uplink throughput of both 4G and 5G outperforms that for P_M. However, it still deteriorates when taking a bus/tram or driving a car. Specifically, the 5G uplink throughput presents significant fluctuations under these mobility patterns especially when driving, compared to slower ones like walking. Moreover, the median decreases by 23.5% when driving. We can also observe a slight downward trend for 4G uplink throughput.

End-to-end network latency. We measure the end-to-end latency by round-trip time (RTT) between the 5G smartphone and the cloud server using the `ping` tool. Fig. 16 shows the end-to-end RTT for P_M. We can observe significant fluctuations in both 4G and 5G RTTs, ranging from as low as 30 ms to sometimes exceeding 120 ms. 5G RTT is slightly lower than 4G RTT for the first three mobility patterns (i.e., stationary, walking, and biking), which can be attributed to 5G's lower radio access network (RAN) latency. Specifically, 5G is supposed to provide ultra-low RAN latency on the order of 1 ms, while it is on the order of 10 ms for 4G [43]. However, for the latter three mobility patterns (i.e., bus, tram, and driving), 5G RTT even grows to be slightly higher than 4G RTT. On the one hand, under these mobility patterns, radio signals are more likely to be interfered due to their high speed and vehicle body shell obstruction. Although the 5G radio frequency in our measurements is sub-6 GHz, it is still $\sim 2\times$ that of 4G, leading to potentially severer signal attenuation. On the other hand, 5G also suffers higher handover latency compared to 4G, due to additional control signaling overhead [11] and longer processing time [13]. Especially under high mobility, the handover process is more likely to fail due to unsuccessful Random Access Channel (RACH) procedures, which further increases the overhead and prolongs the handover latency.

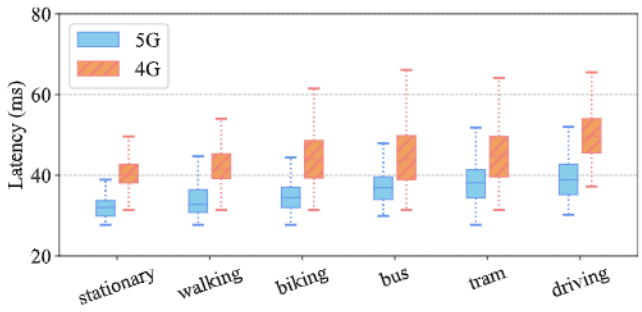


Fig. 18. End-to-end RTT of provider P_U in different mobility patterns.

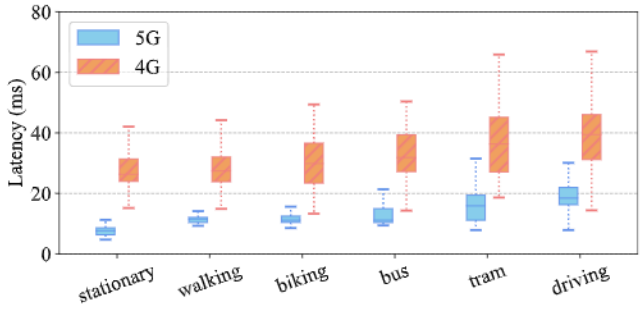


Fig. 19. First three-hop RTT of provider P_U in different mobility patterns.

To take a further look at RAN latency, we break down the end-to-end latency. For a better comparison of 4G/5G RAN latency, we use `traceroute` to find the nearest gateway and shorten the entire end-to-end path to the first few hops so that we can eliminate the latency of the wide area network between the gateway and the cloud server. The first two hops are not available due to the restriction of the mobile carrier. Thus, we use the IP address of the third hop from the UE for RTT measurements with `ping`.

Fig. 17 shows that, surprisingly, the first three-hop RTT of both 4G and 5G remains high and unstable. The wired network between the UE and the cloud server, with possible network congestion along the path, does not contribute significantly to the high end-to-end latency. Specifically, the first three-hop RTT is lower than the end-to-end RTT by only around 20 ms, which should be the latency of the wide area network. For stationary mobility, while the first three-hop RTTs of 5G present a lower median than the ones of 4G, it sometimes becomes higher than 4G. This implies an unstable 5G RAN of P_M, but we are unable to confirm the cause due to the black box of the providers' network. Overall, the RTT of the first three hops is still high and far from the expectation for 5G RAN latency.

However, such an abnormal phenomenon does not occur in the networks of P_U. We repeat the latency measurement experiments the same as above. Fig. 18 shows the end-to-end RTT for P_U under different mobility patterns. The results of both 4G and 5G are much lower than that for P_M. Although 5G RTTs present an increasing trend, its median remains below 40 ms while the median of 4G RTTs is generally 10 ms higher than that of 5G in all mobility patterns. We also measure the first third-hop RTT of P_U. As shown in Fig. 19, the latency of the first three hops is much smaller (i.e., usually below 20 ms) and more stable. For 4G, this RTT ranges from 20 ms to 60 ms. The stable lower RAN latency of P_U is attributed to the flattened architecture of 5G (i.e., part of cellular core network functions sinks to gNB so as to minimize processing latency) [11]. As the mobility pattern changes, both the 4G and 5G RTTs increase. The median 5G RTT even increases by more than 3× while driving, compared to the stationary scenario. It is worth noting that this RTT exhibits significant fluctuations when taking a tram, compared to taking a bus with similar moving velocity.

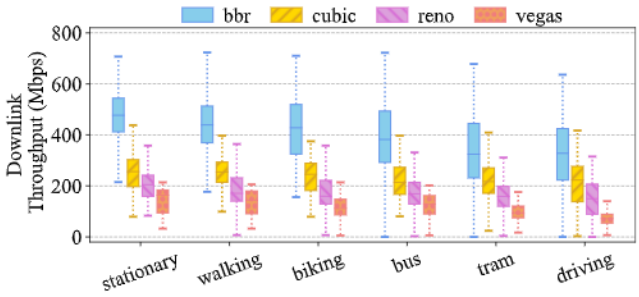


Fig. 20. 5G downlink throughput under different congestion control algorithms.

Table 4
Packet loss rate of 5G traces in different mobility patterns.

Mobility Pattern	Mean (%)	95th (%)
Stationary	0.4	1.2
Walking	0.9	1.7
Biking	1.1	1.9
Bus	1.3	4.5
Tram	1.5	2.9
Driving	1.7	3.4

We attribute this to the tram's thicker body shell, which may cause signal quality degradation and thus higher latency (Section 5.1).

Packet loss. To further study the cause of throughput differences between different mobility patterns, we examine the packet loss rate (PLR) of traces under all mobility patterns. We obtain the PLR by comparing the packet traces captured on the client and server. Table 4 shows the mean and 95th percentile PLR of all traces collected under different mobility patterns. Generally, the PLR increases as the mobility pattern changes. However, we observe the 95th percentile PLR is remarkably high when taking a bus, likely due to its high connection density. For instance, when the bus arrives at the station and a group of people get on the bus, they share the bandwidth resource. Consequently, the available bandwidth of users on the bus rapidly drops, leading to sudden network congestion and increased packet loss. Such a high packet loss rate can degrade the experience of various real-time applications that rely on low latency and reliable data transmission.

Summary. *5G offers substantially higher throughput than 4G, but it also exhibits significant fluctuations. Specifically, both downlink and uplink throughput experience considerable degradation under mobility patterns with high velocity. 5G RTT also increases in these scenarios due to not only high velocity but also obstruction of the vehicle body shell. Moreover, the high connection density experienced when taking a bus may cause abrupt network congestion and increased packet loss. Consequently, all these performance degradations can undermine the consistent and smooth experience of mobile applications.*

5.3. The impact of congestion control

We also study the 5G network performance when using different congestion control algorithms (CCA) as well as how mobility affects the performance of these CCAs. We consider four typical CCAs, namely capacity-probing based *BBR*, loss-based *Cubic* and *Reno*, and delay-based *Vegas*. We switch the congestion control algorithm by configuring the Linux kernel modules of the cloud server. We measure the downlink throughput of these CCAs under various mobility patterns. All measurements with different CCAs are carried out repeatedly with the same measurement method described in Section 4.1.

Fig. 20 plots the downlink throughput of all four CCAs under various mobility patterns. We can observe that BBR obviously outperforms the other three CCAs, as it can achieve a downlink throughput of up to 700Mbps while the others fall below 400Mbps. Cubic provides higher downlink throughput than Reno and Vegas. Vegas can only

Table 5
Packet loss rate of different congestion control algorithms in the driving scenario.

CCA	Vegas	BBR	Reno	Cubic
Mean (%)	~ 0	0.2	0.5	1.1

reach up to 200Mbps, which performs the worst among all CCAs. The two loss-based CCAs (i.e., Cubic, Reno) cannot sufficiently utilize the actual high bandwidth of the 5G networks due to falsely reducing the congestion window when packet loss happens even if it is not caused by congestion. As the mobility pattern changes, the above observations still hold. For example, BBR still outperforms all other CCAs, and Vegas is still the worst algorithm. However, the downlink throughput of all CCAs gradually decreases as the mobility pattern changes. For BBR, the median downlink throughput drops by 33% when driving in contrast to the stationary scenario. Moreover, the downlink throughput sometimes drops to zero, which means the connection breaks and the UE cannot receive data anymore. Such a phenomenon happens more often while users are driving or taking a bus/tram.

We also study the packet loss rate (PLR) of different CCAs under all mobility patterns. We obtain the PLR by comparing the packet traces captured on the client and server. Table 5 shows the mean PLR of different CCAs in the driving scenario, while the results of other mobility patterns present a similar trend. In general, packet loss appears the most while using Cubic in all mobility patterns, followed by Reno. The PLR of BBR is close to zero when users are stationary, walking, or biking. Although packet loss occurs slightly more frequently with BBR when driving, the PLR is still much lower than that of Cubic or Reno. This can be attributed to BBR's rapid adaptation to fluctuations in available bandwidth, facilitated by its efficient probing method. In contrast, loss-based CCAs fail to achieve such swift adaptation, as they only reduce the congestion window after packet loss has already occurred. The PLR of Vegas remains close to zero in all mobility patterns because it maintains a low throughput and seldom causes network congestion.

Summary. *BBR consistently outperforms the other three CCAs (i.e., Cubic, Reno, Vegas) across all mobility patterns. Specifically, BBR can achieve ~2 × 5G downlink throughput compared to Cubic, while also maintaining lower PLR at the same time thanks to its efficient probing method. However, the PLR tends to increase when users are driving.*

6. Application performance

In this section, we evaluate the QoE of the applications (i.e., cloud-based AR, cloud gaming, and video streaming) under 5G as well as 4G for comparison.

6.1. Cloud-based AR

We first focus on measuring the QoE of cloud-based AR, defined by the end-to-end latency. Specifically, we develop an application, where the client sends video frames to the cloud server continuously. The server applies deep learning to detect objects in the frames and returns the detection result to the client for rendering. We use OpenCV and the state-of-the-art object detection framework YOLO [38] to implement the cloud-side software components. The client sends video frames at 10FPS to the cloud server via a TCP connection. The received frames are put in a processing buffer at the server for detection and the results are returned to the client immediately with another TCP connection.

As shown in Fig. 21, the end-to-end latency of cloud-based AR over 5G is only a little smaller than that over 4G (e.g., 11.4% smaller median for the stationary scenario), which hints that 5G does not improve the QoE of AR as much as expected (e.g., below 20 ms for immersive experience). Moreover, the end-to-end latency slightly increases as the mobility pattern changes for 5G. In contrast, the latency increases more under 4G.

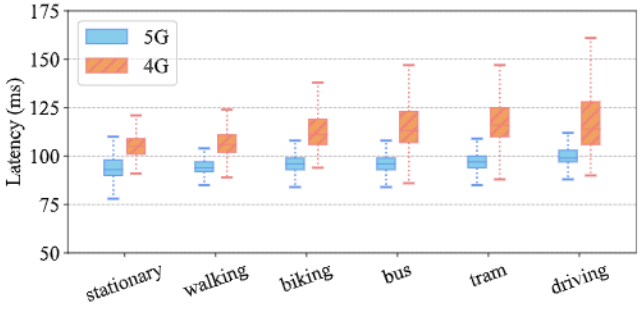


Fig. 21. End-to-end latency of cloud AR in different mobility patterns.

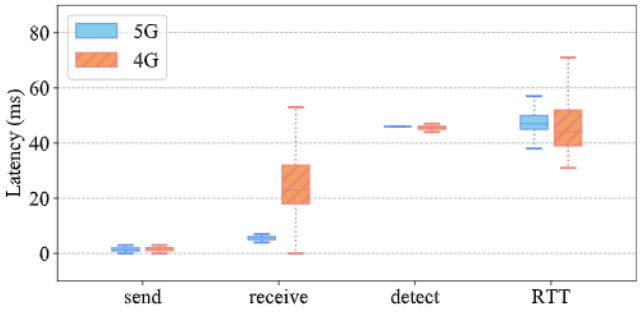


Fig. 22. End-to-end latency breakdown of driving traces.

To understand the contribution to the network latency, as shown in Fig. 22, we break down the end-to-end latency into frame sending latency (i.e., "send"), frame receiving latency (i.e., "receive"), detection latency (i.e., "detect"), and propagation latency (i.e., the latency between frame sending and receiving, plus the latency of returning the result), using the logs from both the client and server. The breakdown hints that the propagation latency (e.g., more than 40 ms) and the detection latency (e.g., around 46 ms) are both the bottlenecks in improving the QoE of the cloud-based AR system.

Because of the higher uplink throughput of 5G, we observe much lighter *network congestion* compared to 4G, which has a significant effect on application QoE. (1) Latency: the receiving latency (i.e., the time between the first packet and the last packet of a frame received by the server) over 5G networks is only a quarter of that over 4G, since the packets of a frame over 4G experience severe network congestion and take a long time to finally arrive at the server successively. (2) Throughput of frames: we observe from the client logs that the actual sending rate is slightly lower than 10FPS while transmitting frames over 4G. While the frames are sent at a fixed frame rate, poor network conditions can cause network congestion and make the sending socket buffer filled immediately, thus delaying the transmission of frames from the client to the server. By contrast, we observe that the FPS can reach above 30 over 5G based on our further measurement.

We also perform measurements with higher sending rates (i.e., 20 and 30FPS). Fig. 23 shows that as the sending rate increases, the end-to-end latency over 4G increases drastically since a higher sending rate makes the network more congested. However, the latency over 5G remains almost unchanged as a result of its sufficient uplink throughput.

Summary. *Currently, 5G does not improve the QoE as much as expected. The propagation latency in 5G is comparable to that in 4G, which indicates providers are supposed to achieve lower 5G RTT for immersive experience. Additionally, the detection latency on the server is also a bottleneck. Therefore, it is necessary to optimize the detection models and develop pipelining methods to take full advantage of 5G performance.*

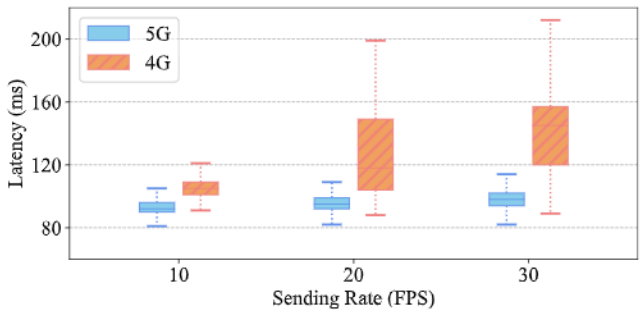


Fig. 23. End-to-end latency at different sending rates.

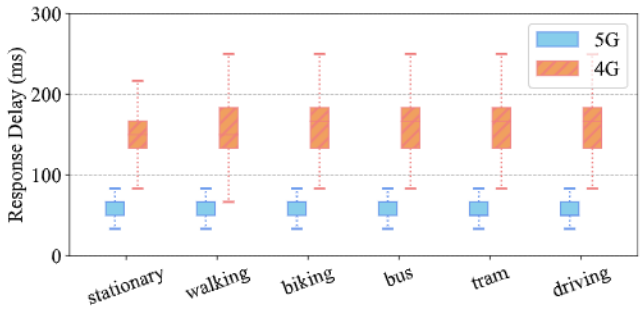


Fig. 24. Response delay of cloud gaming in different mobility patterns.

6.2. Cloud gaming

We measure the QoE of cloud gaming using the same metric as the response delay mentioned in [35]. This metric captures the time interval between a user submitting a command and the corresponding in-game action appearing on the screen. Specifically, in our experiments, the client starts the command by clicking a checkbox in the game (pingus [44]), and the server responds by rendering and returning a frame with the previous checkbox as selected. Moreover, we simulate users' clicks with the Android command-line tool `input`, which periodically triggers the click events in the game. To make it easy to evaluate the response delay, we enable the touch pointer position to set clicks visible while recording the screen with the built-in recording feature. When a click event is finished, the touch pointer on the screen disappears. After capturing the screen, we use FFmpeg [36] to extract all frames from the recorded video, and then utilize OpenCV [37] to figure out the frames with touch pointers and selected checkboxes. The first frame after the touch pointer disappears is marked as $F_{clicked}$, and the first frame with the selected checkbox as $F_{checked}$. Finally, we calculate the time between $F_{clicked}$ and $F_{checked}$ to obtain the response delay.

Fig. 24 shows that the response delay of cloud gaming is much lower over 5G compared with that over 4G under all mobility patterns. Specifically, the results over 4G are usually higher than 140 ms, which fails to support real-time games. Although 5G has reduced the delay by around 50%, it is still far from the latency requirement of cloud gaming (i.e., 20–30 ms [45]). We also observe extremely high response delay (e.g., 250 ms) sometimes, which is 10× the latency requirement of cloud gaming. Note that the game logic also has an impact on the response delay, i.e., complex game logic can further increase this delay.

Summary. *5G reduces the response delay by around 50% compared to 4G. However, the improved result is still far from the latency requirement of cloud gaming.*

6.3. 8K video streaming

Internet traffic is increasing very rapidly, and video streaming accounts for the dominant part [42]. Nowadays, more and more streaming service providers adopt adaptive bitrate (ABR) streaming to deliver video content to users. However, since the network performance of 5G fluctuates violently, it is unclear whether current ABR algorithms can provide expected QoE for 4K/8K video streaming over 5G. To solve this puzzle, we measure the QoE of 8K adaptive bitrate streaming over 5G under various mobility patterns. We use Dash.js [39], a reference client implementation of the MPEG-DASH standard, to stream video from our cloud server over 4G or 5G to the client. Dash.js also provides useful APIs for examining the QoE of adaptive streaming.

We study several ABR algorithms, including buffer-based BB [46] and BOLA [47], rate-based RB and FESTIVE [48], control-theoretic MPC, and robustMPC [49]. We evaluate the QoE of adaptive video streaming with various metrics, such as bitrate, buffer occupancy as well as QoE reward used in FastMPC [49]. We prepare a four-minute 8K video and encode it into 6 tracks with different bitrates using `FFmpeg` and `libx264`. The bitrate of the highest video quality is 160Mbps. We then measure the QoE of adaptive video streaming over 4G and 5G networks. Each experiment of different ABR algorithms is repeated for 5× in all mobility patterns including stationary (S), walking (W), biking (B), driving (D), bus (BS), and tram (T). We present the measurement results in Figs. 25 and 26 (S4 means that the measurement is conducted over 4G networks in the stationary state).

We observe higher QoE rewards under 5G than 4G as expected, due to the higher throughput of 5G to support higher bitrate. The QoE reward of 4G is often below zero (e.g., fastMPC). This is because video streaming over 4G may suffer a long rebuffering time (0.5-3s, as shown in Fig. 27), while that over 5G below 0.5s. In fact, the buffer length grows faster under 5G due to its sufficient throughput, which benefits users with a good QoE for a long time, even when they enter areas with bad network conditions, without stuttering or resolution degradation. Also, BOLA outperforms other ABR algorithms when streaming over 5G in all mobility patterns. All of the ABR algorithms rarely select the best video quality (i.e., 160Mbps) over 5G, which hints that 5G cannot well support 8K video streaming currently.

Summary. *5G can support higher video bitrates compared to 4G but cannot support 8K yet, which may necessitate the higher throughput provided by mmWave 5G. In terms of ABR algorithms, BOLA outperforms other algorithms in all mobility patterns when streaming over 5G.*

7. Improving 5G application QoE under mobility

This section provides some insights on improving 5G applications QoE under mobility scenarios, to bridge the gap between wishes and reality of 5G applications.

7.1. Network-aware route planning

Current navigation applications mostly rely on the length and road traffic of roadways to recommend the route, without considering the availability and performance of 5G networks. However, recommending routes with good network conditions is an urgent demand to improve the QoE of 5G applications under mobility, e.g., in-car video streaming and cloud gaming.

Here we implement a prototype of a network-aware route planning system. The system consists of three components, i.e., network map construction, application QoE estimation, and network-aware route planning. We next introduce the role of each component and the implementation respectively in the following.

Network map construction. To make the system network aware, we first need to integrate 5G network information (e.g., RSRP, bandwidth, latency) into the geographic map. In our system, we construct

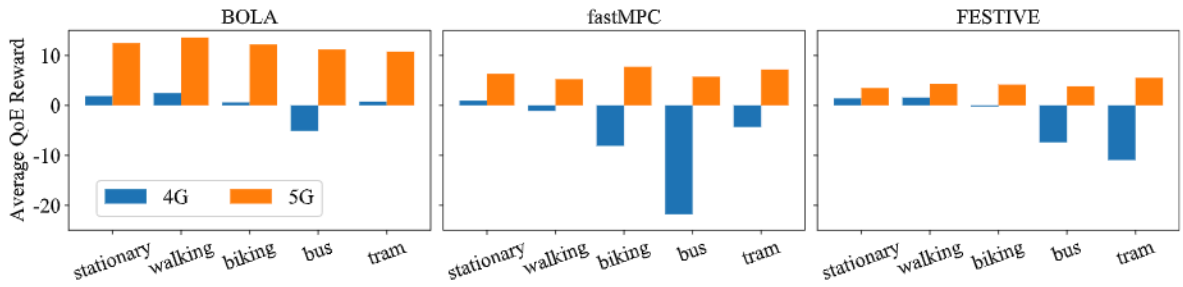


Fig. 25. Average QoE reward of video streaming in different mobility patterns.

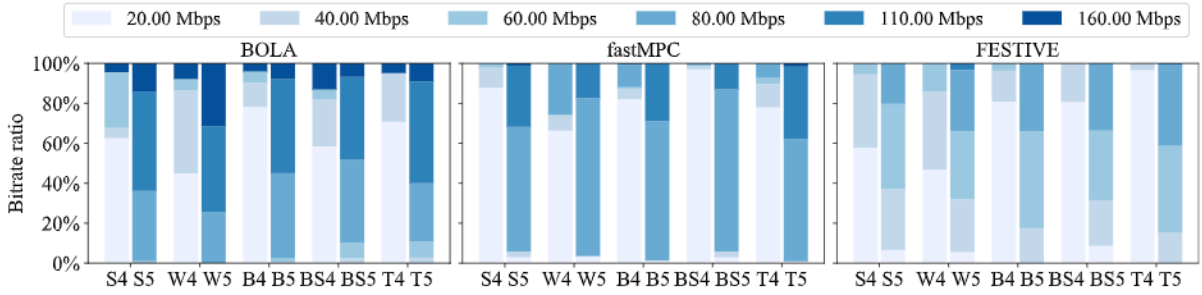


Fig. 26. Bitrate selection of video streaming in different mobility patterns.

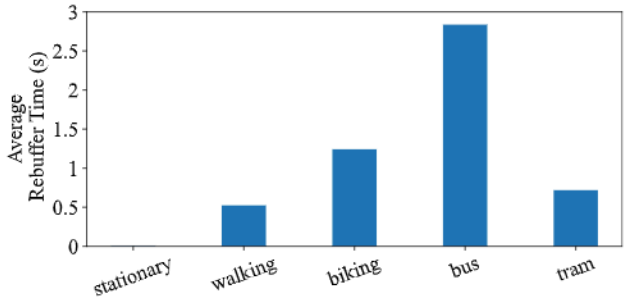


Fig. 27. Rebuffering time of fastMPC over 4G in different mobility patterns.

Table 6
QoE penalty of routes.

Coverage	RSRP (dBm)	Penalty
Good	[−80, −40)	0
Fair	[−90, −80)	1
Bad	[−100, −90)	2
No coverage	[−140, −100)	3

the network map based on an open-source geographic map OSM [50] and a 5G dataset [11]. Specifically, we incorporate the 5G network dataset into the geographic map by mapping the measured network points to the nearest road segment according to their GPS coordinates.

Application QoE estimation. Based on the network map, we then estimate the application QoE on each road segment for future route planning. We take bandwidth-intensive applications as an example, whose QoE typically depends on throughput. As poorer signal strength correlates with lower average throughput statistically (Fig. 11), leading to QoE degradation, here we estimate the QoE penalty based on signal strength for simplicity in the prototype design. Moreover, signal strength is much easier to measure and collect than throughput in the wild. For a certain road segment, the total penalty is calculated according to the following formula: $P = \frac{\sum_{i=1}^{n} p_i}{n} \times d$, where P is the total penalty of the road and p_i is the penalty of a sample point on the road. The penalty of a sample point is defined by its measured RSRP value, as shown in Table 6. And d is the distance of the road. Finally,

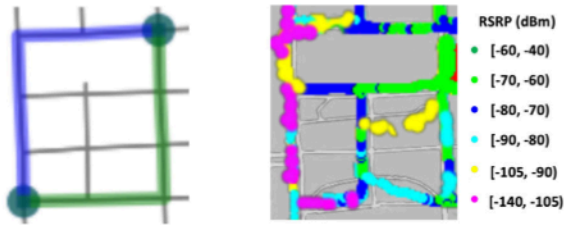
Table 7
Characteristics of routes.

Characteristics	Blue route	Green route
Distance (m)	384	386
Duration (s)	276.5	277.9
Total penalty	789.8	364.1

we recommend routes with the least QoE penalty using the routing engine. It is worth mentioning that the above QoE penalty formula is just a basic model and does not hinder the effectiveness evaluation of our network-aware route planning system, given the substantial difference in network performance across different routes. For accurate QoE estimation in real-world systems, more sophisticated estimation models for application QoE are essential, which necessitate additional network metrics and in-depth application-specific knowledge. We will further discuss the challenges of QoE estimation later.

Network-aware route planning. Given the estimation of application QoE on each road segment, network-aware route planning can be achieved by utilizing common shortest path algorithms. Compared to traditional route planning problems that focus on travel distance and duration, our system regards application QoE as the weight of each road. Specifically, we utilize an open-source routing engine OSRM [51] for route planning, which employs the Contraction Hierarchies (CH) algorithm [52] to find the shortest path. We modify the configuration of the routing engine (via a Lua script [53]) to use customized weight (i.e., the estimated QoE penalty) instead of distance for each road segment. When receiving a request with the starting and target location, the routing engine first needs to compute the weight of each road segment. To estimate the QoE on each road segment, the routing engine queries network information from a PostgreSQL database (i.e., our network map with 5G data) and then computes the QoE penalty by the aforementioned formula. Finally, it finds the shortest path using the CH algorithm.

Performance evaluation. We then conduct a case study to evaluate the effectiveness of our network-aware route planning system. We choose a pair of starting and target locations and then use the current navigation algorithm and our network-aware route planning system to recommend the best route respectively.



(a) The shortest path (blue) and the network-aware path (green)

(b) 5G coverage (i.e., RSRP) in the area from the 5G dataset [11]

Fig. 28. Recommended routes of the case study and its 5G coverage.

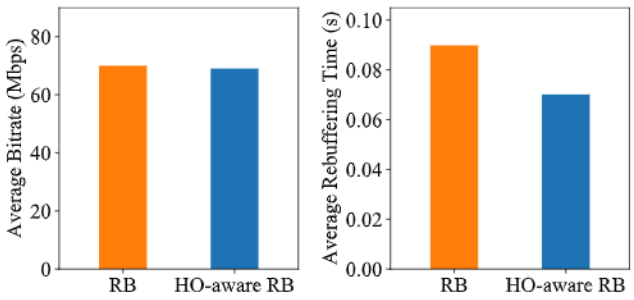


Fig. 29. Average bitrate and rebuffering time of RB and HO-aware RB.

Fig. 28(a) shows two alternative recommended routes for the same starting and target location, with the 5G network coverage on the road as depicted in Fig. 28(b). The blue route is recommended for the shortest distance by the current navigation algorithm, while the green route is recommended for the best application QoE by our network-aware route planning system. As shown in Table 7, although the distance and duration of the green route are slightly longer than the blue route, it offers much better network conditions than the blue route (~2× penalty). Obviously, the green route can provide much better application QoE and should be recommended for users. Therefore, network-aware route planning can improve the QoE of 5G applications effectively.

Potential challenges and future research directions. While our proposed network-aware route planning system can significantly enhance the QoE of 5G applications, it also presents several challenges and requires further exploration.

First, network congestion may occur when multiple users are recommended to the same path. An intuitive approach to avoid this congestion is adapting recommended routes according to real-time network conditions (e.g., available bandwidth, latency) on the road, as current navigation applications do when traffic jams occur. However, it is challenging to construct a large-scale real-time network map, which requires joint efforts among network providers, service providers, and users. Specifically, it needs collaborative crowdsourcing measurements with both users and network providers to collect network information [9]. While users contribute UE-side data, network providers provide their knowledge about 5G network as well as user data usage. Furthermore, application developers or service providers become able to access network information from network providers via the 5G *network exposure function* (NEF) [54]. However, while the NEF provides network information that can be leveraged for developers, it is important to note the exposure of network information should avoid any data privacy concerns.

Second, it is non-trivial to estimate application QoE based on network QoS, according to varying specific requirements for different applications. For example, the QoE of bandwidth-intensive applications (e.g., video streaming) typically depends on network throughput, while that of real-time applications (e.g., cloud gaming) puts more emphasis on network latency. Furthermore, better network QoS does not always translate to a proportional improvement in application QoE. For instance, the marginal improvement in perceived video quality may decrease at higher throughput. Additionally, the complex interplay between application behavior and network protocol makes QoE estimation even harder. Thus, developers should carefully utilize their application-specific knowledge to ensure an accurate QoE estimation.

Furthermore, network-aware route planning also poses new challenges for route planning algorithms. While considering application QoE as the weight of each road, it might make traditional shortest path algorithms ineffective sometimes. For example, the QoE of adaptive video streaming is typically affected by video bitrate and stall time. And the video stall time experienced on the current road is significantly influenced by the total duration of video chunks in the buffer, which is buffered on the previously traveled road. Therefore, the QoE on the current road is affected by not only the network conditions on the current road but also the ones on the previous road. These dependencies between roads can invalidate the optimal substructure property, which is a fundamental requirement for traditional route planning or shortest path algorithms.

7.2. Handover-aware application adaptation

User mobility can incur handovers, which have a significant impact on network performance and application QoE. When a handover occurs, the connection between a UE and the serving base station breaks. Therefore, applications' data transmission is disrupted until a new connection to the target base station is established, leading to significant QoE degradation. We take adaptive video streaming as an example. If the client requests a high bitrate video chunk before a handover, it may suffer a prolonged downloading time due to subsequent throughput reduction caused by the handover, leading to potential video stalls and degraded QoE.

Currently, applications typically adapt to varying network conditions according to end-to-end throughput/latency or application-specific metrics (e.g., buffer occupancy in adaptive video streaming). However, these adaptation methods are reactive, as they adjust application behavior only after performance degradation is observed, compromising the time efficiency and effectiveness. To address the above issue, we propose a handover-aware adaptation method based on handover prediction. We take adaptive video streaming as a case study. First, our method predicts whether handover will occur in the next prediction window according to measured historical signal strengths [8,55]. Then, it estimates the subsequent throughput following the handover. Based on the estimated throughput in the future, our method select a maximum bitrate that can avoid video stalls to ensure a smooth streaming experience. Compared to classic ABR algorithms, our method can achieve early adaptation prior to potential throughput degradation, and thus reap the opportunity for avoiding unnecessary stalls and maintaining higher application QoE.

We then evaluate the performance of our method using throughput trace-driven emulation as in [56]. We modify the rate-based (RB) algorithm to integrate our handover-aware method and compare their performance. Fig. 29 shows the average bitrate and rebuffering time of RB algorithm and our method (i.e., HO-aware RB). As shown, our method effectively reduces the average rebuffering time by 21.9% while maintaining comparable bitrate levels.

7.3. Mobility-aware edge resource adaptation

Our measurement results indicate that the QoE of 5G applications can be bottlenecked by the backbone network when communicating with remote cloud servers (e.g., high AR latency presented in Fig. 22). While edge computing is expected to improve application QoE due to its proximity to users, the QoE can be still affected by varying network

Table 8
Resources and pricing of different functions.

Specification	Function 1	Function 2
Memory Size (MB)	128	512
Processing Latency (ms)	48	12
Pricing ($\times 10^{-9}$\$/ms)	2.1	8.3

Table 9
SLO violation rate and cost of different resource employment methods.

Metric	Fixed-Low	Fixed-High	Our method
SLO Violation Rate (%)	11.7	0.3	0.3
Total Cost ($\times 10^{-6}$\$)	3.78	14.94	5.04

performance during mobility. Specifically, the end-to-end application latency consists of the network latency across the network path and the processing latency at edge servers. While network latency can be unstable during mobility as depicted in Fig. 6, it leaves less time for edge servers to process requests, given a specific end-to-end latency requirement. This poses a challenge in effectively utilizing edge computing resource. On one hand, limited computing resource may struggle to meet latency requirements when network latency unexpectedly spikes during mobility. On the other hand, deploying excessive computing resource may result in wastage of both resources and costs. Therefore, we propose an adaptive mobility-aware edge resource adaptation method to guarantee satisfactory end-to-end latency by mitigating the above latency variability with appropriate resources. Our proposed method dynamically employs increased computing resources when the current resources cannot meet the latency requirement during mobility.

We evaluate the effectiveness of the proposed method with our collected network traces. The network latency between the mobile device and the edge servers is emulated based on our collected network traces. We use an event processing application as in [57], which is implemented by an AWS Lambda function where you only pay for the computing resources and time you actually use [58]. In AWS Lambda, the allocated resources scales linearly with the selected memory size of functions. We assume that there are two available functions with different memory sizes in the edge servers. The detailed specifications of the functions are listed in Table 8. We compare the performance of our method with two baselines, consistently selecting a function with either low or high memory size, denoted as Fixed-Low and Fixed-High, respectively. Table 9 shows the SLO violation rate and total cost of different selection methods. Given the SLO of 100 ms, Fixed-Low results in a high violation rate of 11.7%. On the other hand, Fixed-High brings the violation rate down to 0.3%, yet leading to a fourfold total cost. In contrast, our adaptive method can maintain the same SLO guarantee as Fixed-High, while significantly reducing the cost by 66.3%. The evaluation results indicate that our method is promising to improve application performance during mobility with adaptive edge resources.

Determining the most suitable edge resource involves accurately formulating the relationship between the mobility and the network performance, as well as profiling processing latency. Yet, this task is challenging due to the substantial searching space constituted by a variety of mobility patterns and highly heterogeneous edge servers [59], thus hindering effective edge resource adaptation significantly.

8. Future work

We discuss the limitations and future work of our measurement work in this section.

Sub-6 GHz vs mmWave. While our measurement study focuses on sub-6 GHz 5G networks, we are also interested in exploring the performance of mmWave 5G and its impact on application QoE under various mobility patterns in future work. Here, we briefly discuss the potential implications of mmWave 5G on our findings. Specifically, compared to sub-6 GHz 5G, mmWave 5G is expected to (1) offer significantly higher throughput, (2) exhibit similar network latency, (3) experience more pronounced fluctuations in network throughput and latency due to its greater fragility to obstructions, and (4) have shorter coverage per base station. As a result, while bandwidth-hungry applications in the eMBB scenario may benefit from improved QoE, the increased network fluctuations could lead to highly inconsistent QoE. For real-time applications in the URLLC scenario, mmWave may not significantly enhance their QoE, underscoring the need to reduce latency in both the RAN and core networks. Additionally, in mobility scenarios, as studied in this work, the fragility to obstructions and shorter coverage of mmWave may present new challenges in maintaining a reliable connection. We believe these implications present intriguing avenues for future work.

5G Advanced and 3CC. To the best of our knowledge, while new network features like 3CC provided by 5G-Advanced may improve network performance, they are still in initial deployment and have not been widely commercially deployed yet [60]. Moreover, the majority of off-the-shelf smartphones still cannot support those advanced network features including 3CC. Therefore, we believe our measurement study can better reflect the perceived 5G performance of most existing smartphones. We are very interested in studying the performance of 5G-Advanced under various mobility patterns in the future.

Impact of physical layer metrics on application QoE. While we measured the impact of physical layer metrics on network performance, it is also interesting to study their respective impacts on applications QoE. For example, it is unclear whether "poor average channel quality" and "good but unstable channel quality" have the same impact on the application layer. We believe that the effects of these two types of channel quality can vary depending on the specific requirements of different applications. For instance, "good but unstable channel quality" often leads to frequent data retransmissions due to a high bit error rate when channel quality deteriorates. This significantly affects the QoE for latency-sensitive applications by increasing tail latency, as illustrated in Fig. 3 in Section 5.1. However, for bandwidth-hungry applications, this type of channel quality may have a lesser impact, as these applications can mitigate temporary quality degradation through content prefetching (e.g., buffering in video streaming). On the other hand, "poor average channel quality" tends to have a more significant impact on bandwidth-hungry applications compared to latency-sensitive ones. This is because bandwidth-hungry applications demand more network capacity, while poor channel quality fails to provide consistently sufficient throughput as demonstrated in Fig. 4 in Section 5 We are very interested in further exploring the impact of physical layer metrics on applications QoE in the future.

9. Conclusions

In this paper, we conduct a comprehensive cross-layer measurement study of current commercial 5G networks with typical 5G applications under various mobility patterns. Considering the requirements of stability and continuity of applications, current 5G networks still cannot sustain high application QoE. Our measurement results show that the achievable application QoE under current commercial 5G networks falls behind the requirements imposed by these applications. We further provide three insights on improving the QoE of these 5G applications: network-aware route planning, mobility-aware application adaptation, and locality-based edge server selection.

CRediT authorship contribution statement

Jiahai Hu: Writing – original draft. **Lin Wang:** Writing – review & editing. **Jing Wu:** Writing – review & editing. **Qiangyu Pei:** Writing – review & editing. **Fangming Liu:** Supervision. **Bo Li:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] Bnamericas, 5G on pace to exceed 500 million connections by 2021, 2021, https://www.bnamericas.com/en/news/5g-on-pace-to-exceed-500-million-connections-by-2021.

[2] J.P. Tomás, 5G already covers 53% of Germany's territory: Regulator, 2021, https://www.rcrwireless.com/20211209/5g/5g-already-covers-53-germany-territory-regulator.

[3] Qualcomm, ABI Research, Augmented and virtual reality: The first wave of 5G killer apps, 2017, https://www.qualcomm.com/media/documents/files/augmented-and-virtual-reality-the-first-wave-of-5g-killer-apps.pdf.

[4] Y. Siriwardhana, P. Porambage, M. Liyanage, M. Ylianttila, A survey on mobile augmented reality with 5G mobile edge computing: Architectures, applications, and technical aspects, IEEE Commun. Surv. Tutor. 23 (2) (2021) 1160–1192, http://dx.doi.org/10.1109/COMST.2021.3061981.

[5] X. Zhang, H. Chen, Y. Zhao, Z. Ma, Y. Xu, H. Huang, H. Yin, D.O. Wu, Improving cloud gaming experience through mobile edge computing, IEEE Wirel. Commun. 26 (4) (2019) 178–183, http://dx.doi.org/10.1109/MWC.2019.1800440.

[6] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, Z.-L. Zhang, A first look at commercial 5G performance on smartphones, in: Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 894–905.

[7] A. Narayanan, X. Zhang, R. Zhu, A. Hassan, S. Jin, X. Zhu, X. Zhang, D. Rybkin, Z. Yang, Z.M. Mao, F. Qian, Z.-L. Zhang, A variegated look at 5G in the wild: Performance, power, and qoe implications, in: Proceedings of the ACM SIGCOMM 2021 Conference, Association for Computing Machinery, New York, NY, USA, 2021, pp. 610–625.

[8] A. Hassan, A. Narayanan, A. Zhang, W. Ye, R. Zhu, S. Jin, J. Carpenter, Z.M. Mao, F. Qian, Z.-L. Zhang, Vivisecting mobility management in 5G cellular networks, in: Proceedings of the ACM SIGCOMM 2022 Conference, Association for Computing Machinery, New York, NY, USA, 2022, pp. 86–100.

[9] A. Narayanan, E. Ramadan, R. Mehta, X. Hu, Q. Liu, R.A.K. Fezeu, U.K. Dayalan, S. Verma, P. Ji, T. Li, F. Qian, Z.-L. Zhang, Lumos5G: Mapping and predicting commercial MmWave 5G throughput, in: Proceedings of the ACM Internet Measurement Conference, Association for Computing Machinery, New York, NY, USA, 2020, pp. 176–193.

[10] H. Lim, J. Lee, J. Lee, S.D. Sathyanarayana, J. Kim, A. Nguyen, K.T. Kim, Y. Im, M. Chiang, D. Grunwald, et al., An empirical study of 5g: Effect of edge on transport protocol and application performance, IEEE Trans. Mob. Comput. (2023).

[11] D. Xu, A. Zhou, X. Zhang, G. Wang, X. Liu, C. An, Y. Shi, L. Liu, H. Ma, Understanding operational 5G: A first measurement study on its coverage, performance and energy consumption, in: Proceedings of the ACM SIGCOMM 2020 Conference, Association for Computing Machinery, New York, NY, USA, 2020, pp. 479–494.

[12] M. Xu, Z. Fu, X. Ma, L. Zhang, Y. Li, F. Qian, S. Wang, K. Li, J. Yang, X. Liu, From cloud to edge: A first look at public edge platforms, in: Proceedings of the 21st ACM Internet Measurement Conference, Association for Computing Machinery, New York, NY, USA, 2021, pp. 37–53.

[13] Y. Pan, R. Li, C. Xu, The first 5G-LTE comparative study in extreme mobility, in: Proceedings of the ACM on Measurement and Analysis of Computing Systems, Vol. 6, (1) ACM New York, NY, USA, 2022, pp. 1–22.

[14] M. Ghoshal, I. Khan, Z.J. Kong, P. Dinh, J. Meng, Y.C. Hu, D. Koutsonikolas, Performance of cellular networks on the wheels, in: Proceedings of the 2023 ACM on Internet Measurement Conference, 2023, pp. 678–695.

[15] M. Ghoshal, Z.J. Kong, Q. Xu, Z. Lu, S. Aggarwal, I. Khan, J. Meng, Y. Li, Y.C. Hu, D. Koutsonikolas, Can 5G mmwave enable edge-assisted real-time object detection for augmented reality? in: 2023 31st International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS, IEEE, 2023, pp. 1–8.

[16] I. Khan, T.X. Tran, M. Hiltunen, T. Karagioules, D. Koutsonikolas, An experimental study of low-latency video streaming over 5G, 2024, arXiv preprint arXiv:2403.00752.

[17] Qualcomm, Everything you need to know about 5G, 2022, https://www.qualcomm.com/5g/what-is-5g.

[18] VIAVI, The state of 5G, 2022, https://www.viavisolutions.com/en-us/literature/state-5g-may-2022-posters-en.pdf.

[19] A. Gupta, R.K. Jha, A survey of 5G network: Architecture and emerging technologies, IEEE Access 3 (2015) 1206–1232.

[20] Qualcomm, Deploying 5G NR mmwave to unleash the full 5G potential, 2020, https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/deploying_mmwave_to_unleash_the_full_5g_potential_web.pdf.

[21] B. Cai, H. Zhang, H. Guo, G. Zhang, W. Xie, 5G network evolution and dual-mode 5G base station, in: 2020 IEEE 6th International Conference on Computer and Communications, ICCC, 2020, pp. 283–287.

[22] J. Lee, Y. Kwak, 5G standard development: Technology and roadmap, in: Signal Processing for 5G, John Wiley & Sons, Ltd, 2016, pp. 561–576, http://dx.doi.org/10.1002/9781119116493.ch23.

[23] GSA, 5G Standalone Update: Member Report, 2021, https://gsacom.com/paper/5g-standalone-update-member-report-june-2021/.

[24] Huawei, Huawei 5G Wireless Network Planning Solution White Paper, 2018, https://www-file.huawei.com/-/media/corporate/pdf/white%20paper/2018/5g_wireless_network_planing_solution_en_v2.pdf.

[25] L. Liu, H. Li, M. Gruteser, Edge assisted real-time object detection for mobile augmented reality, in: The 25th Annual International Conference on Mobile Computing and Networking, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–16.

[26] R. Shea, A. Sun, S. Fu, J. Liu, Towards fully offloaded cloud-based AR: Design, implementation and experience, in: Proceedings of the 8th ACM on Multimedia Systems Conference, Association for Computing Machinery, New York, NY, USA, 2017, pp. 321–330.

[27] L. Li, K. Xu, D. Wang, C. Peng, Q. Xiao, R. Mijumbi, A measurement study on TCP behaviors in HSPA+ networks on high-speed rails, in: 2015 IEEE Conference on Computer Communications, INFOCOM, 2015, pp. 2731–2739.

[28] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from GPS trajectories, in: Proceedings of the 18th International Conference on World Wide Web, Association for Computing Machinery, New York, NY, USA, 2009, pp. 791–800.

[29] Y. Zheng, Q. Li, Y. Chen, X. Xie, W.-Y. Ma, Understanding mobility based on GPS data, in: Proceedings of the 10th International Conference on Ubiquitous Computing, Association for Computing Machinery, New York, NY, USA, 2008, pp. 312–321.

[30] Y. Zheng, X. Xie, W. Ma, GeoLife: A collaborative social networking service among user, location and trajectory, IEEE Data Eng. Bull. 33 (2) (2010) 32–39.

[31] J. Liu, S. Chen, G. Lederman, D.B. Kramer, H.Y. Noh, J. Bielak, J. Garrett, J. Kovacevic, M. Berges, Dynamic responses, GPS positions and environmental conditions of two light rail vehicles in Pittsburgh, Sci. Data 6 (146) (2019).

[32] Ookla, Speedtest, 2021, https://www.speedtest.net/.

[33] ESnet, iperf3: A TCP, UDP, and SCTP network bandwidth measurement tool, 2021, https://github.com/esnet/iperf.

[34] tcpdump: a powerful command-line packet analyzer, 2021, https://www.tcpdump.org.

[35] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, K.-T. Chen, GamingAnywhere: An open cloud gaming system, in: Proceedings of the 4th ACM Multimedia Systems Conference, Association for Computing Machinery, New York, NY, USA, 2013, pp. 36–47.

[36] FFmpeg: A complete, cross-platform solution to record, convert and stream audio and video, 2021, https://www.ffmpeg.org/.

[37] OpenCV: Open source computer vision library, 2021, https://opencv.org/.

[38] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 779–788.

[39] dash.js: A reference client implementation for the playback of MPEG DASH via Javascript and compliant browsers, 2021, https://github.com/Dash-Industry-Forum/dash.js.

[40] O. Chukhno, O. Galinina, S. Andreev, A. Molinaro, A. Iera, Interplay of user behavior, communication, and computing in immersive reality 6G applications, IEEE Commun. Mag. 60 (12) (2022) 28–34.

[41] F. Menegoni, G. Albani, M. Bigoni, L. Priano, C. Trotti, M. Galli, A. Mauro, Walking in an immersive virtual reality, in: Annual Review of Cybertherapy and Telemedicine 2009, IOS Press, 2009, pp. 72–76.

[42] Producing live 8k, 360-degree streaming media events: An owner's blueprint, 2020, https://builders.intel.com/docs/networkbuilders/producing-live-8k-360-degree-streaming-media-events.pdf.

[43] I. Parvez, A. Rahmati, I. Guvenc, A.I. Sarwat, H. Dai, A survey on low latency towards 5G: RAN, core network and caching solutions, IEEE Commun. Surv. Tutor. 20 (4) (2018) 3098–3130, http://dx.doi.org/10.1109/COMST.2018.2841349.

[44] Pingus, 2021, https://pingus.seul.org.

[45] Mobile cloud gaming – an evolving business opportunity, 2020, https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/mobile-cloud-gaming.

[46] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, M. Watson, A buffer-based approach to rate adaptation: Evidence from a large video streaming service, in: Proceedings of the ACM SIGCOMM 2014 Conference, Association for Computing Machinery, New York, NY, USA, 2014, pp. 187–198.

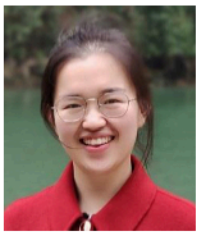
[47] K. Spiteri, R. Urgaonkar, R.K. Sitaraman, BOLA: Near-optimal bitrate adaptation for online videos, IEEE/ACM Trans. Netw. 28 (4) (2020) 1698–1711.

[48] J. Jiang, V. Sekar, H. Zhang, Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE, in: Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies, Association for Computing Machinery, New York, NY, USA, 2012, pp. 97–108.

[49] X. Yin, A. Jindal, V. Sekar, B. Sinopoli, A control-theoretic approach for dynamic adaptive video streaming over HTTP, in: Proceedings of the ACM SIGCOMM 2015 Conference, Association for Computing Machinery, New York, NY, USA, 2015, pp. 325–338.
[50] OpenStreetMap contributors, 2017, Planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org.
[51] D. Luxen, C. Vetter, Real-time routing with OpenStreetMap data, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '11, ACM, New York, NY, USA, 2011, pp. 513–516, http://dx.doi.org/10.1145/2093973.2094062.
[52] R. Geisberger, P. Sanders, D. Schultes, D. Delling, Contraction hierarchies: Faster and simpler hierarchical routing in road networks, in: Experimental Algorithms: 7th International Workshop, WEA 2008 Provincetown, MA, USA, May 30-June 1, 2008 Proceedings 7, Springer, 2008, pp. 319–333.
[53] OSRM profiles, 2024, https://github.com/Project-OSRM/osrm-backend/blob/master/docs/profiles.md.
[54] ETSI white paper, MEC in 5G networks, 2018, https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf.
[55] L. Mei, J. Gou, Y. Cai, H. Cao, Y. Liu, Realtime mobile bandwidth and handoff predictions in 4G/5G networks, Comput. Netw. 204 (2022) 108736.
[56] H. Mao, R. Netravali, M. Alizadeh, Neural adaptive video streaming with pensieve, in: Proceedings of the Conference of the ACM Special Interest Group on Data Communication, 2017, pp. 197–210.
[57] S. Eismann, L. Bui, J. Grohmann, C. Abad, N. Herbst, S. Kounev, Sizeless: Predicting the optimal size of serverless functions, in: Proceedings of the 22nd International Middleware Conference, 2021, pp. 248–259.
[58] Amazon Web Services, AWS Lambda Pricing, 2024, https://aws.amazon.com/lambda/pricing.
[59] H. Ning, Y. Li, F. Shi, L.T. Yang, Heterogeneous edge computing open platforms and tools for internet of things, Future Gener. Comput. Syst. 106 (2020) 67–76.
[60] Telecom Review, China mobile pioneers the future with commercial 5G-a deployment, 2024, https://www.telecomreview.com/articles/telecom-operators/8048-china-mobile-pioneers-the-future-with-commercial-5g-a-deployment.

Jiahai Hu received the B.Eng. degree from Huazhong University of Science and Technology, China, in 2020. He is currently a Ph.D. student in the School of Computer Science and Technology, Huazhong University of Science and Technology, China. His research interests include edge computing, 5G network, and mobile systems.

Lin Wang is currently a Full Professor and Head of the Chair of Computer Networks in the Department of Computer Science at Paderborn University. Previously, he held positions at Vrije Universiteit Amsterdam, TU Darmstadt, SnT Luxembourg, and IMDEA Networks Institute. He is broadly interested in networked systems, with a focus on in-network computing, machine learning systems, and intermittently-powered IoT systems. He has received a Google Research Scholar Award, an Outstanding Paper Award of RTSS 2022, Best Paper Awards of IPCCC 2023 and HotPNS 2016, and an Athene Young Investigator Award of TU Darmstadt.



Jing Wu received the M.S. degree in the School of Computer Science and Engineering, Northeastern University, Shenyang, China, in 2018. She is currently a Ph.D. student in the School of Computer Science and Technology, Huazhong University of Science and Technology, China. Her research interests include edge computing, augmented reality, and deep learning.

Qiangyu Pei received his B.S. degree in physics from Huazhong University of Science and Technology, China, in 2019. He is currently a Ph.D. student in the School of Computer Science and Technology, Huazhong University of Science and Technology, China. His research interests include edge computing, green computing, and deep learning.

Fangming Liu is currently a Full Professor with the Huazhong University of Science and Technology, Wuhan, China. His research interests include cloud/edge computing, datacenter and green computing, SDN/NFV/5G and applied ML/AI. He received the National Natural Science Fund (NSFC) for Excellent Young Scholars, and the National Program Special Support for Top-Notch Young Professionals. He is a recipient of the Best Paper Award of IEEE/ACM IWQoS 2019, ACM e-Energy 2018 and IEEE GLOBECOM 2011, the First Class Prize of Natural Science of Ministry of Education in China, as well as the Second Class Prize of National Natural Science Award in China.

Bo Li is a Chair Professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, which he has been affiliated with since 1996. He held the Cheung Kong Chair Professor in Shanghai Jiao Tong University between 2010 and 2015. His research interests cover broad areas in networking and distributed systems, with recent focuses on big data and machine learning systems, cloud and edge computing. He was a co-recipient of seven Best Paper Awards from IEEE including the Test-of-Time Paper Award from IEEE INFOCOM (2015) and the Best Paper Award from IEEE INFOCOM (2021).