Case for 5G-Aware Video Streaming Applications

Eman Ramadan, Arvind Narayanan, Udhaya Kumar Dayalan, Rostand A. K. Fezeu,
Feng Qian, Zhi-Li Zhang
{eman,arvind,zhzhang}@cs.umn.edu,{dayal007,fezeu001,fengqian}@umn.edu
Department of Computer Science & Engineering, University of Minnesota – Twin Cities, USA

ABSTRACT

Recent measurement studies show that commercial mmWave 5G can indeed offer ultra-high bandwidth (up to 2 Gbps), capable of supporting bandwidth-intensive applications such as ultra-HD (UHD) 4K/8K and volumetric video streaming on mobile devices. However, mmWave 5G also exhibits highly variable throughput performance and incurs frequent handoffs (e.g., between 5G and 4G), due to its directional nature, signal blockage and other environmental factors, especially when the device is mobile. All these issues make it difficult for applications to achieve high Quality of Experience (QoE). In this paper, we advance several new mechanisms to tackle the challenges facing UHD video streaming applications over 5G networks, thereby making them *5G-aware*. We argue for the need to employ machine learning (ML) for effective throughput prediction to aid applications in intelligent bitrate adaptation. Furthermore, we advocate *adaptive content bursting*, and *dynamic radio (band) switching* to allow the 5G radio network to fully utilize the available radio resources under good channel/beam conditions, whereas dynamically switched radio channels/bands (e.g., from 5G high-band to low-band, or 5G to 4G) to maintain session connectivity and ensure a minimal bitrate. We conduct initial evaluation using real-world 5G throughput measurement traces. Our results show these mechanisms can help minimize, if not completely eliminate, video stalls, despite wildly varying 5G throughput.

CCS CONCEPTS

• **Networks → Mobile networks**; **Application layer protocols**;
• **Information systems → Multimedia streaming**;

KEYWORDS

5G, mmWave, 5G Throughput, Volumetric Video Streaming, 5G-Aware Applications, Adaptive Content Bursting, Dynamic Radio (Band) Switching

ACM Reference Format:
Eman Ramadan, Arvind Narayanan, Udhaya Kumar Dayalan, Rostand A. K. Fezeu,, Feng Qian, Zhi-Li Zhang. 2021. Case for 5G-Aware Video Streaming Applications. In *ACM SIGCOMM 2021 Workshop on 5G Measurements, Modeling, and Use Cases (5G-MeMU'21), August 23, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472771.3474036

5G-MeMU'21, August 23, 2021, Virtual Event, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8636-4/21/08…$15.00
https://doi.org/10.1145/3472771.3474036

1 INTRODUCTION

With its diverse new radio bands ranging from low-band and mid-band to high-band mmWave radio, 5G is touted as a key enabler for a variety of new applications that requires ultra-low latency and/or ultra-high bandwidth. These applications include 4K/8K video streaming, *interactive* 360° and volumetric video streaming, cloud gaming, Augmented Reality/Virtual Reality (AR/VR), among others. With a theoretical throughput up to 20 Gbps which is far beyond 4G [21], mmWave 5G is particularly suited to support these breeds of *bandwidth-intensive* video applications. On the other hand, based on theoretical modeling, simulation studies and limited field testing, it was widely believed that mmWave radio has limited ranges and requires line-of-sight (LoS) for good performance. This is because mmWave signals are highly directional and sensitive to various environmental factors.

We have conducted a first "in-the-wild" extensive measurement study [16] of commercial 5G services, focusing in particular on Verizon's mmWave 5G in several US cities. While confirming some of known or suspected issues associated with mmWave radio, our measurement study captures the "in-the-wild" performance of today's commercially deployed 5G services, and reveals new challenges and opportunities facing applications that are enabled by mmWave 5G. Through extensive and repeated experiments under various settings, we find that i) mmWave 5G can indeed offer ultra-high bandwidth, up to 2 Gbps under good channel conditions and clear LoS; and ii) even without direct LoS, mmWave 5G can often deliver throughput higher than 400 Mbps, due to reflections from surrounding buildings and other objects. This is in contrast to 4G LTE/LTE Advanced which has a *theoretical* peak bandwidth of 150/300 Mbps. On the other hand, iii) mmWave 5G throughput is highly variable over time and can fluctuate wildly from 100s Mbps to 1 or 2 Gbps with slight changes in orientations and locations or due to blockage from moving objects in the surroundings; and worse, iv) mmWave 5G throughput may at times drops to near zero (5G "dead zones") and incur frequent handoffs (e.g., between 5G and 4G), especially under mobility (see §2 and Fig. 1 for an example). Our findings not only demonstrate the exciting new opportunities offered by (mmWave) 5G for enabling new *bandwidth-intensive* applications, but also reveal new challenges for these applications.

The paper is centered around the following fundamental problem: *How can we endow bandwidth-intensive applications with the abilities to fully take advantage of the (potential) ultra-high bandwidth offered by (mmWave) 5G while at the same time overcome its highly variable throughput performance so as to deliver good and consistent quality-of-experience (QoE) to mobile users?* To address this fundamental challenge, we use mobile *volumetric video streaming* as a case study. Such application requires bandwidth as high as 750 Mbps. Using mmWave 5G throughput traces, we first conduct trace-driven simulations (see §3) to answer the following two basic questions: 1) are (volumetric) video streaming applications equipped with existing *adaptive bitrate* (ABR) algorithms ready to take advantage of 5G's high throughput? and 2) how does the wild throughput fluctuations affect the application performance from the perspective of QoE (measured in terms of video stall times)? Our investigation reveals that wild fluctuations in 5G throughput often lead to quick buffer depletion under poor channel conditions, especially when entering 5G "dead zones," thereby resulting in a large stall time that has a significant impact on user's QoE. Our findings illustrate that new mechanisms are needed to endow applications with the abilities to fully utilize the potential of 5G while overcoming its challenges. We refer to applications endowed with such capabilities as *being 5G-aware*.

We advocate new mechanisms to make applications *5G-aware* (§4). We first note that ABR algorithms used in existing video streaming applications rely mostly on *in-situ* bandwidth "probing" for throughput estimation. The highly variable throughput performance of mmWave 5G, coupled with frequent handoffs, make such methods ineffectual [18]. We argue that a) more sophisticated *machine learning (ML) methods for effective throughput prediction*[1] that can account for diverse environmental factors and be able to forecast 5G throughput over a longer time horizon are needed to aid applications in intelligent bitrate adaptation. Furthermore, we advocate b) *adaptive content bursting* – namely, employing (significantly) larger buffers (both at the client side as well as within the 5G radio network) – to allow the 5G radio network to fully utilize the available radio resources under good channel/beam conditions to burst as much content as needed to the client so as to prepare for and bridge over the 5G bandwidth troughs and dead zones. In addition, c) employing *dynamic radio (band) switching* (e.g., between 5G and 4G or between 5G high, mid, and low bands) is crucial in maintaining session connectivity and ensuring minimal bitrates.

We conduct trace-driven experiments (§5) to evaluate the efficacy of these strategies in overcoming the wild fluctuations of 5G throughput performance. Our experimental results demonstrate that these strategies can consistently deliver high video quality (compared to the theoretical optimal performance), and in particular, minimize, and even completely eliminate video stall times, despite 5G dead zones.

In summary, we identify both the opportunities and challenges offered by emerging 5G services, and call for new mechanisms to make applications *5G-aware* – namely, enabling applications to take full advantages of opportunities offered by 5G while overcoming the new challenges it poses. Our study clearly constitutes only an initial step towards this direction – much more work needs to be done by the research community to make applications *5G-aware*.

2 CHARACTERISTICS OF 5G NETWORKS

5G-New Radio (5G-NR) supports a very wide range of frequency spectrum, right from the sub-6 GHz range (which includes both low- and mid- band 5G) to millimeter wave (mmWave) range. Due to the physical layer characteristics of wireless signal propagation, performance characteristics can dramatically vary across these different bands. For example, 5G services deployed at the mmWave-range provides ultra-high bandwidth capacity however posses poor signal propagation characteristics thus leading to poor coverage. On the other hand, low-band 5G provides superior coverage but has low-bandwidth capacity. Using Android APIs, our measurement tool [17] collects the following fields: timestamp, latitude, longitude, tower ID, mobility mode {i.e., walking, still, driving}, and active radio type {5G or 4G}. Experiments were conducted using a Samsung Galaxy S10 device with 5G capability. We refer to two of our recent measurement studies to understand the characteristics of 5G networks: (1) [16] looked at the network performance of several commercial 5G carriers in the US including the mmWave-based 5G networks, (2) Lumos5G[15] further conducts detailed impact factor analysis to understand how different UE-side factors (such as geolocation, mobility direction, speed, UE-Panel distance, etc.) impact mmWave 5G performance. In the context of this paper, we summarize the key findings of these studies.

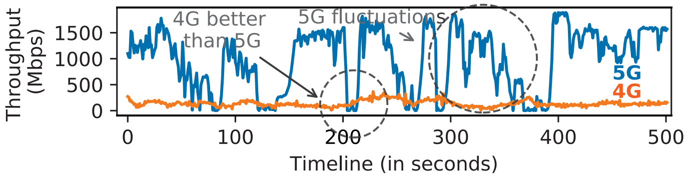


Figure 1: 4G and 5G Throughput Traces While Walking.

(1) With a peak downlink throughput of ~2 Gbps, *mmWave 5G clearly shows its ability to provide ultra-high bandwidth*. This is critical for bandwidth-hungry applications such as volumetric video streaming or real-time ultra-HD video streaming, which otherwise are not feasible using existing 4G/LTE services. **(2)** However, due to mmWave's signal propagation issues especially under mobility scenarios (e.g., driving or walking), mmWave 5G shows much higher throughput variation. For instance, comparing real-world throughput traces of 4G and 5G (see Fig. 1), 5G reports a standard deviation of 579 Mbps compared to 59 Mbps for 4G. Similarly, Due to the different 5G-NR bands that have implications on the coverage characteristics, 5G's performance characteristics can be tricky to map especially in the case of mmWave 5G. For instance, 5G throughput can suddenly drop to 0 Gbps where there is no mmWave coverage (referred to as 5G dead zones). In such spots, 4G/LTE might offer better performance than 5G (see Fig 1). In other words, *mmWave 5G shows wild and frequent fluctuations in performance which can potentially confuse network and application layer logic such as ABR video streaming potentially leading to under utilization of the channel bandwidth and resources provided by the carrier*. These issues are inherent characteristics of 5G mmWave due to its physical nature. Such performance characteristics of commercial 5G coupled with the different deployment strategies (e.g., NSA v/s SA[2]) have adverse implications on application performance that is not well explored or understood. We use volumetric video streaming application as a case study to first use real-world 5G traces to illustrate the new challenges posed by today's commercial 5G offerings. Secondly, we also propose new mechanisms that can help overcome them.

[1] In [15] we have demonstrated that it is feasible to predict (mmWave) 5G throughput using machine learning algorithms with weighted average F1 score of above 0.95. Such high accuracy is shown to be adequate for video ABR adaptation [28].

[2] In this paper, we address mmWave's signal propagation characteristics which will remain the same regardless of its deployment strategy (NSA or SA).

3 VIDEO STREAMING PERFORMANCE UNDER 5G THROUGHPUT

Volumetric videos[3] differ from regular and 360° videos in that they are truly 3D, with each frame consisting of a 3D point cloud. During playback, users wearing a mixed reality (MR) headset can freely navigate themselves with six degrees of freedom (6 DoF) movement, gaining an immersive telepresence experience. A volumetric video can have 350K points per frame played at 30 frames per second (FPS). Each point takes 9 bytes (3 bytes for RGB color and 6 bytes for its 3D location). This yields a total of 350K×30×9×8 = 756 Mbps *when uncompressed*. While the 756 Mbps throughput requirement far exceeds the capacity of existing 4G LTE service, it is well within the ultra-high bandwidth offered by the commercial mmWave 5G service. Unfortunately, decoding (*compressed*) point cloud data requires heavy-weight algorithms such as *octree* [8, 13, 24] that cannot be effectively supported by today's mobile phones at the 30 FPS frame rate [19]. Thus, streaming uncompressed volumetric videos to mobile phones is the only practical solution at the moment.

To understand the impact of the large, wild fluctuations of 5G throughput on existing video streaming applications, we use the 5G trace from Fig. 1[4] as a representative trace to stream a volumetric video for 500 seconds played at a constant rate of 350K points per frame (see §5.1 for experiment settings). We measure the performance by total stall time; a stall (rebuffering) occurs for every missing frame at its playback time till the frame is downloaded from the server. This results in a total stall time of 90 seconds (18%).

Despite the very high throughput of 5G, this "non-smooth" QoE to users with frequent stalls is attributed to the sudden and quick drop in 5G throughput. Also, existing video streaming applications do not take full-advantage of the extra available throughput (that can reach as high as 2 Gbps) thus might end up being wasted. This is indicated in Fig. 2 which shows that the maximum number of frames at any point in the buffer corresponds to 4.2 secs (i.e., 126 frames) which are not enough to cover long 5G dead zones which can extend to 20 secs. Only, when the network throughput varies "smoothly", client-side buffering would work reasonably well and help further "smooth out" the effects of short-term throughput fluctuations, which clearly is not the case for mmWave 5G. This raises the questions of i) how long the buffer should be to cover 5G dead zones, and ii) which bitrate quality to request as it affects the time and bandwidth required to download each frame.

The bitrate is often determined by the estimated throughput. However, traditional bandwidth estimation approaches which rely on the short-term past history and use methods like harmonic mean or other methods (e.g., [11]) are not adequate for 5G throughput due to its wild and non-smooth variation. Moreover, 3G/4G networks can rely on location to predict the cellular performance [14, 25], however mmWave 5G throughput is more complicated as it is affected by multiple factors and is very sensitive to the surrounding environment. Hence, traditional location-based prediction models are insufficient.

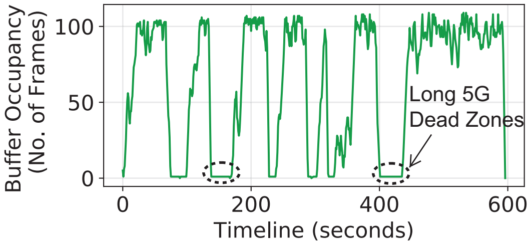


Figure 2: Buffer Occupancy During a 500 SEC Video Using 5G Throughput Shown in Fig. 1.

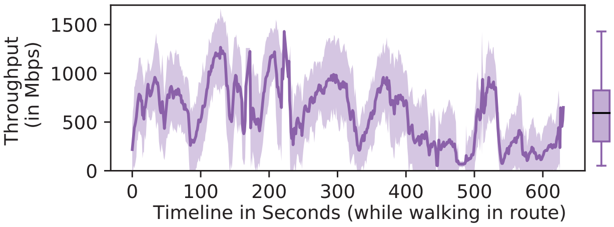


Figure 3: Variation in 5G Throughput.

This trace-driven simulation points out both the opportunities and challenges in mmWave 5G, and shows that existing video streaming applications do not work well over mmWave 5G. Hence, we need to rethink about the way these applications are built to become 5G-aware. There is a need to come up with novel mechanisms to effectively utilize the extra high bandwidth offered by 5G whenever available while at the same time coping with the wild fluctuations and occasional "dead zones" to improve the user's QoE.

4 5G-AWARE VIDEO STREAMING

We propose new mechanisms to make bandwidth-intensive applications *5G-aware* so as to take full advantage of 5G networks while overcoming their new challenges. First, we highlight the need for new ML throughput prediction mechanisms, then put forth several *cross-layer* mechanisms to effectively utilize the available radio resources and improve user's QoE despite 5G's high throughput variability and dead zones.

4.1 Need for ML 5G Throughput Prediction

Despite the wild variability of 5G throughput compared to 4G, our recent study [15] argues, through extensive experiments and statistical analysis, that by controlling the key user-side (UE) factors affecting 5G, the throughput can largely be characterized and can be predictable. These key factors include for example user's geolocation, mobility mode, mobility speed, and user's compass direction. Then, it proposes *Lumos5G* – a composable machine learning framework which considers different combinations of contextual and environmental factors, and applies the state-of-the-art machine learning algorithms for making context-aware 5G throughput predictions with a higher accuracy over existing traditional prediction methods. As an example, Fig. 3 shows the distribution (or spread) of variation seen in 5G throughput traces (aggregated using 40 runs collected over a span of 20 days) along a walking route: the dark center curve represents the average throughput and shaded areas represent the 25% to 75% percentile range. From this figure, we can notice that there are some patches when the throughput is consistently high, while others the throughput is consistently low.

[3] A sample of a high-quality volumetric video streaming can be found at https://www.youtube.com/watch?v=feGGKasvamg.

[4] Throughout this paper, we use mobility traces to study and overcome the impact of 5G dead zones, on top of 5G throughput variability which occurs for stationary users. These user mobility scenarios are likely to happen for Autonomous Vehicle applications and vehicle to everything (V2X) technology.

Although not shown, we also observe that the throughput characteristics and variation drastically vary when the user is walking in the opposite direction. This signifies the importance of compass direction as a key factor in characterizing 5G throughput.

Ideally these ML models can be deployed at 5G base stations, users can collect the UE key factors, and report them to the 5G base station to train the ML models. In return, the user receives a bandwidth prediction map containing 5G dead zones (with a start position and a length) as well as the current/future throughput prediction over a longer time horizon for different routes[5]. With the ability to predict the near future 5G performance in/around the current user's location, video streaming applications can then make intelligent decisions to download video frames as explained next to provide exceptional QoE while at the same time adapt smoothly to 5G's high variation and fluctuations. Additionally, these throughput prediction models can also be used by cellular networks themselves for adaptive beam forming, resource allocation, preemptive handoffs, and improving network coverage.

4.2 Adaptive Streaming Mechanisms

We put forth several mechanisms to enable applications to fully take advantage of ultra-high bandwidth afforded by (mmWave) 5G while also mitigate the impact of high throughput variability due to fast varying frequency radio bands.

• **Adaptive Content Bursting.** The goal of this mechanism is twofold: 1) to "burst" sufficient amount of application data to the 5G radio network so that the 5G radio resource control sub-layer can fully take advantage of available radio resources whenever possible, e.g., when a clear LoS path or good quality high-frequency channel is available; and 2) to bridge over 5G low-bandwidth troughs and "dead zones" by delivering as much data as needed to a user/UE when the channel conditions are good. Goal 1) requires provisioning larger buffer at the radio network, and is motivated by the fact that radio resource allocation and transmission scheduling are often based on the amount of per-user data in the radio network buffer. If a high-quality radio channel or LoS beam is available to a UE but there is little data in the per-user buffer, the 5G radio network cannot fully take advantage of the ultra-high bandwidth offered by 5G. Ensuring there is always sufficient data in the per-user buffer via adaptive content bursting will avoid such "lost opportunities". Goal 2) entails allocating larger buffer at the UE/client side. Clearly, for both to work effectively, the ability to predict channel conditions and (future) 5G throughput, e.g., based on the user orientation, mobility and environmental factors, with ML techniques, is crucial, so that the amount of burst data can be dynamically adapted to balance buffer requirement, QoE, and radio resource utilization.

• **Dynamic Radio Switching.** Through our extensive experiments, we find that in some patches while UE is connected to 5G (but with poor channel quality), 4G in fact yields a higher throughput (see Fig. 1). In other times, UE may enter a 5G dead zone while still under 4G coverage. Hence *proactively* switching between 5G and 4G based on estimated/predicted channel conditions or throughput performance will be crucial in maintaining connectivity and ensuring a minimal bitrate, especially during user mobility. Likewise, dynamically switching between diverse radio channels/bands is also essential in coping with diverse and fast varying channel characteristics (e.g., bandwidth, bit error rate).

In a nutshell, we believe that combining these new (cross-layer) mechanisms, coupled with effective ML-based throughput prediction, will be the key to enable a new class of bandwidth-intensive applications such as volumetric video streaming. Incorporating these new mechanisms entails re-designing the adaptive bitrate (ABR) and other algorithms used in existing video streaming applications so that they can fully utilize the ultra-high bandwidth and other capabilities afforded by (mmWave) 5G, while also help them mitigate various PHY-layer challenges posed by mmWave 5G radio – in other words, making them *5G-aware*.

5 EVALUATION

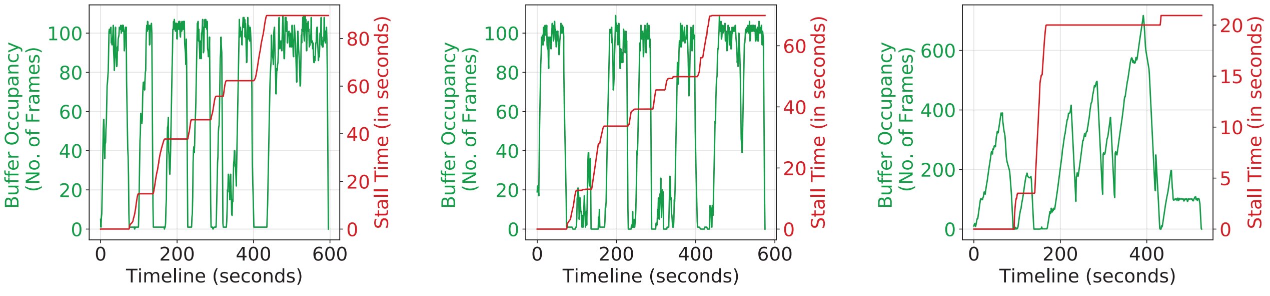
In this section, we conduct trace-driven experiments to demonstrate the benefits of these mechanisms. In particular, we investigate how effectively adaptive content bursting will allow the 5G network to fully take advantage of ultra-high bandwidth when available and help the application to bridge over 5G bandwidth troughs and dead zones. We will also use the real-world 5G/4G throughput traces we have collected to emulate *dynamic radio (band) switching* (between 5G and 4G) to examine its potential benefits in maintaining session connectivity and in further enhancing the user's QoE. These mechanisms will be aided by ML-based 5G throughput prediction [15]. We will in particular prioritize video stall times, and compare the results obtained with the theoretical bounds on the best video quality achieved without any stalls (see Appendix A.1).

5.1 Experimental Setup

Currently there is no way to do radio(band) switching, hence, we built our own emulated video player, using the TCP/IP protocol stack and C++, to fetch video frames from the server to show its effectiveness using real 5G commercial traces. The client player has a large playback buffer (virtually unlimited) to ensure our emulation's performance metrics are able to reflect the network's performance as opposed to the device's hardware specifications. Using our measurement tool, we have collected 4G and 5G traces simultaneously 3 times every day for more than 20 days using Samsung Galaxy S10 5G devices while walking in a dense 5G deployment area in downtown Minneapolis for Verizon's NSA 5G Service. These traces share a common behavior as shown in Fig. 3, hence we pick a representative 5G & 4G network traces shown in Fig. 1 captured during our study while the user is walking at a speed of ≈ 1.4 m/s, and replay it using `tc` [6] to throttle the bandwidth to match the 4G and 5G throughput. We use BBR as TCP congestion control algorithm developed by Google to reduce the impact of TCP slow start due to wild fluctuations. In these experiments, we request frames using constant bitrate[6] (i.e., all frames are requested with the same number of points per frame 350K), and we use the *stall time* (i.e., rebuffering duration) as a metric for user's QoE. We emulate watching the video using 3 modes: *1) 5G Only*: by only using the 5G throughput trace shown in Fig. 1. *2) Dynamic 5G/4G Switching*: with the bandwidth estimation knowledge, the player proactively switches between 4G and 5G networks depending on which one has the higher available bandwidth. *3) Content Bursting + Dynamic*

[5] See [15] for more details about the bandwidth prediction maps and ML deployment.

[6] See Appendix A.2 for variable bitrate quality.



(a) 5G Only (b) Dynamic Switching (c) Content Bursting + Dyn. Switching

Figure 4: Buffer Occupancy and Stall Time During a 500 SEC Video Streamed With Quality 350K.

Switching: in addition to the dynamic switching, the video player also proactively bursts future content as much as possible when extra high bandwidth is available as estimated by the bandwidth estimation module to handle the 5G dead zones shown in Fig. 1. We emulated a 500 seconds video requested at 350K points per frame for these modes, each experiment was repeated at least 3 times with minimal differences among runs, hence a representative run from each mode is shown in Fig. 4 for buffer occupancy and stall time.

5.2 Experimental Results

• **Buffer Occupancy and Stall Time.** *1) 5G Only* mode: Fig. 4a shows the user experiences a large stall time of around 90 secs (out of 8-min walk) with 17.92% of the video frames experiencing stalls. This is due to having a maximum throughput of 200 Mbps in 5G dead zones which is not enough to receive and play frames of 350K points which require a total of 350K×30×9×8 = 756 Mbps. Thus, the user has to wait till they pass these dead zones and get back 5G connectivity to resume fetching frames. Also, the buffer occupancy never exceeds 126 (i.e., a playback length of 4.2 secs) which is clearly not enough to cover 5G dead zones which have longer duration. *2) Dynamic Switching* mode: Fig. 4b shows that with the bandwidth estimation knowledge, switching to 4G shields 5G dead zones reducing the stall time to 70 secs experienced by 14.04% of the video frames. This is attributed to 4G's omnidirectional radio which helps maintain the basic data connectivity during mobility. *3) Content Bursting + Dynamic Switching* mode: Fig. 4c shows when the client player utilizes the ultra-high bandwidth of 5G to proactively request additional frames from the server, the stall time is reduced to 21 secs but was not completely eliminated. However, we can notice that the maximum buffer occupancy increased to 724 frames which helped overcome some 5G dead zones but not all.

Table 1: Stall Time for Video Playback.

Points/Frame	300K	350K	400K
Required Throughput	648 Mbps	756 Mbps	864 Mbps
5G Only	82 sec.	90 sec.	106 sec.
Dynamic 5G/4G Switching	52 sec.	70 sec.	79 sec.
Content Bursting + Dynamic Switching	0 sec.	21 sec.	68 sec.

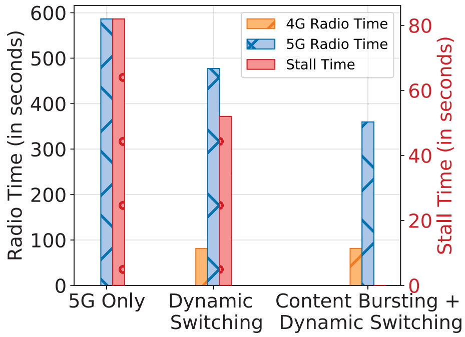


Figure 5: Radio Time for 4G and 5G During a 500 SEC Video Streamed With Quality 300K.

• **Selecting Appropriate Bitrate.** Applying Theorem 1, listed in Appendix A.1, to the given trace in Fig. 1, we found that requesting frames using the video quality at 300K points eliminates any stalls, while other higher video qualities always result in a stall time. We repeated the same experiments by streaming the video using a quality of 300K points per frame with *Content Bursting + Dynamic Switching* mode. The stall time was completely eliminated while maintaining the full frame quality overcoming the throughput fluctuation and dead zones in the 5G throughput trace. We noticed that when the video quality increases, the buffer takes more time to build and consequently gets depleted quickly before/at the dead zones increasing the stall time even when *Content Bursting + Dynamic Switching* mode is employed. The reason for this behavior is that requesting a bitrate higher than what can be supported by the available bandwidth prevents the buffer from building up as it requires more time to download each frame. Table 1 summarizes the stall time for the different modes and video qualities.

• **Radio Time for 4G & 5G.** We use the time spent using each radio (4G/5G) shown in Fig. 5 as a simplified representation for the consumed energy during streaming the video using 300K points per frame. When *Dynamic 5G/4G Switching* mode is used, 4G is enabled for a limited time when its throughput is higher than 5G, and the stall time is minimized to 52 secs and hence 5G radio time decreased. Using *Content Bursting + Dynamic 5G/4G Switching* leads to completely eliminating the stall time, and both radios were ON for the shortest time.

6 RELATED WORK

Several studies have been conducted on mmWave deployments from theoretical point of view [7, 9, 22, 23, 29, 31], however, [16] is the first measurement study on the performance of commercial 5G services by different US carriers. Using 5G traces, the authors in [18] illustrate why current video streaming ABR algorithms do not work well with 5G mmWave. One of the main reasons is attributed to the inaccurate 5G throughput estimation, as was also shown by Zou *et al.* in [32] that better throughput prediction can indeed improve the video performance in cellular networks. Lumos5G [15] was the first ML model to predict 5G throughput with high accuracy illustrating the inefficiency of existing 3G/4G throughput prediction ML-based and data models which can only rely on user location [14, 25]. These studies further support our argument for the need to build robust 5G ML throughput prediction models in video streaming apps as well as the need for new mechanisms to make them 5G-aware.

Volumetric video streaming is a hot topic which has been recently investigated. For example [12] proposes a manifest file format for volumetric video streaming following the DASH standard. Nebula [19] utilizes edge servers to decode the 3D data and generates a 2D video instead. ViVo [10] applies visibility-aware optimizations to enable real-time streaming. These techniques are complementary to our work and can be integrated with our proposed strategies. Other research studies focus on evaluating the QoE performance for video streaming using simulated 5G traces such as [20, 27]. To the best of our knowledge, our paper is the first to study the issues in using commercial mmWave 5G for volumetric video streaming using real-world 5G throughput traces, and propose new mechanisms to build 5G-aware applications.

7 DISCUSSION & FUTURE WORK

In this section, we elaborate on future directions for video streaming applications to further enhance their performance.

• **Scalable Video Coding (SVC).** Most video players use advanced video coding (H.264/MPEG-4 AVC) standardized in 2003 [1] which encodes a video frame into different bitrate versions independently of each other leading to redundant information. A major drawback in AVC encoding is that it cannot adapt to the high fluctuations of 5G bandwidth. Thus, another alternative encoding Scalable Video Coding (SVC) was developed which is an extension to H.264 standardized in 2007 [26]. In SVC, a frame is encoded in a base layer (lowest quality), and multiple enhancement layers which can be used to improve the quality in an incremental way. For each frame, if the base layer is missing at the playback time, a stall will occur; if the higher-quality enhancement layers are missing but not the base layer, the frame will be played at a low quality to avoid stalls; if all layers are present, the frame will be played at the original (highest) quality. This resolves the wasted bandwidth problem of AVC by using layering technique and hence can just download the additional layers up to the specified quality level. SVC comes at the cost of decoding overheads at the client, however nowadays hardware decoders using GPU are available in smart phones.

• **Adaptive Bitrate Algorithms (ABR).** When the available bandwidth changes, instead of prefetching frames with a constant bitrate, a more judicious decision can be made to decide which quality to use based on the predicted future bandwidth, its variability, and the buffer occupancy. Thus, avoid requesting frames with the highest quality which yields only few frames in the buffer that will be depleted quickly. The goal is to develop an adaptive algorithm which can avoid stalls while at the same time deliver the highest possible quality with smooth quality variation instead of frequent changes from the highest quality to the lowest quality which degrade user's QoE (see Appendix A.2 for more details).

• **Multi-Band Aggregation.** 5G supports a broad and diverse range of frequency spectrum. The low-band frequency provides maximum coverage but limited bandwidth, while high-band provides very high bandwidth but its signals are highly sensitive and vulnerable to obstacles thus limiting its coverage. Between both these extremes lies the mid-band range, which provides higher bandwidth capacity than low-band & better coverage than high-band. Since the debut of commercial 5G deployments, carriers supported a single class of frequency range. While high-band (mmWave) range can provide very high bandwidth capacity, its suffers from limited coverage. Hence, several carriers now consider deploying multiple classes to leverage multiple frequency bands which is known as *multi-band 5G*, enabling carriers to aggregate multiple channels to achieve higher data rates. In such situations, low-band and mid-band 5G will allow carriers to provide stable 5G service with wider coverage, while offering mmWave 5G to support bandwidth-heavy applications [2, 4, 5]. Multi-band 5G is now also supported by 5G chip manufacturers who have developed a single-chip which supports multi-band, e.g., Qualcomm's Snapdragon X55 5G modem-RF supports both mmWave and sub-6 GHz 5G new radio [3]. Streaming *uncompressed* volumetric videos makes it easier to adopt a flexible, *layered* approach for multi-band 5G deployment and video bitrate adaptation. Low-band and reliable radio channels with good conditions can be used to stream the base layer with the minimum video quality & bandwidth requirement, while *simultaneously* mid-band/high-band 5G are used to stream higher quality enhancement layers by dynamically adapting to the available network bandwidth through adjusting the resolution (i.e., increasing or decreasing the number of points) of an entire (or portions of) 3D video frame.

• **Cross-layer Design.** Due to the new challenges posed by 5G, we believe cross-layer mechanisms are required to improve user's QoE such as e.g., dynamic radio resource allocation (see [30] for discussion), PHY-layer/MAC-Layer/RRC-Layer info passed to the transport layer so that congestion control (CC) algorithms can work well. E.g., due to frequent handoffs in mmWave 5G, packet loss might affect the congestion window (cwnd). If signal strength improves and if we know it is going to be stable, then we might want to increase the cwnd sooner than following the CC algorithm approach which might under-utilize the available bandwidth.

8 CONCLUSION

This paper points out both the opportunities for UHD video streaming applications as well as the challenges they face in mmWave 5G affecting their performance. We argued for the need to shift the way we develop applications for 5G to utilize ML throughput prediction, adaptive content bursting, dynamic radio(band) switching to make video streaming applications 5G-aware. Using real-world 5G traces, our results show these mechanisms can improve user's QoE, despite wildly varying 5G throughput.

ACKNOWLEDGMENTS

We thank our shepherd Özgü Alay and the anonymous reviewers for their insightful suggestions and feedback. This research was in part supported by NSF under Grants CNS-1903880, CNS-1915122, CNS-1618339, CNS-1617729, CNS-1814322, CNS-1831140, CNS-1836772, and CNS-1901103.

REFERENCES

[1] 2003. *H.264/MPEG-4 AVC.* Retrieved July 2021 from http://handle.itu.int/11.1002/1000/6312

[2] 2019. *5G Low Latency Requirements.* Retrieved July 2021 from https://broadbandlibrary.com/5g-low-latency-requirements/

[3] 2019. *Snapdragon X55 5G modem-RF system.* Retrieved July 2021 from https://www.qualcomm.com/products/snapdragon-x55-5g-modem

[4] 2019. *The 5G Status Quo is Clearly Not Good Enough.* Retrieved July 2021 from https://www.t-mobile.com/news/the-5g-status-quo-is-clearly-not-good-enough

[5] 2020. *5G spectrum: strategies to maximize all bands.* Retrieved July 2021 from https://www.ericsson.com/en/networks/trending/hot-topics/5g-spectrum-strategies-to-maximize-all-bands

[6] 2021. Traffic Control in the Linux kernel. https://linux.die.net/man/8/tc/. Last accessed July 2021.

[7] Sylvain Collonge, Gheorghe Zaharia, and G EL Zein. 2004. Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel. *IEEE Transactions on Wireless Communications* 3, 6 (2004).

[8] T. Golla and R. Klein. 2015. Real-time point cloud compression. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5087–5092. https://doi.org/10.1109/IROS.2015.7354093

[9] Muhammad Kumail Haider, Yasaman Ghasempour, Dimitrios Koutsonikolas, and Edward W Knightly. 2018. Listeer: mmwave beam acquisition and steering by tracking indicator leds on wireless aps. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM.

[10] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-Aware Mobile Volumetric Video Streaming. In *ACM MobiCom.*

[11] Qi He, Constantine Dovrolis, and Mostafa Ammar. 2005. On the predictability of large transfer TCP throughput. In *ACM SIGCOMM Computer Communication Review*, Vol. 35. ACM, 145–156.

[12] Mohammad Hosseini and Christian Timmerer. 2018. Dynamic Adaptive Point Cloud Streaming. In *Proceedings of the 23rd Packet Video Workshop (PV)*. 6 pages.

[13] Yan Huang, Jingliang Peng, C. C. Jay Kuo, and M. Gopi. 2008. A Generic Scheme for Progressive Point Cloud Coding. *IEEE Transactions on Visualization and Computer Graphics* 14, 2 (March 2008), 440–453. https://doi.org/10.1109/TVCG.2007.70441

[14] Robert Margolies, Ashwin Sridharan, et al. 2016. Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms. *IEEE/ACM Transactions on Networking (TON)* 24, 1 (2016), 355–367.

[15] Arvind Narayanan, Eman Ramadan, et al. 2020. Lumos5G: Mapping and Predicting Commercial MmWave 5G Throughput. In *ACM IMC'20.*

[16] Arvind Narayanan, Eman Ramadan, Jason Carpenter, Qingxu Liu, Yu Liu, Feng Qian, and Zhi-Li Zhang. 2020. A First Look at Commercial 5G Performance on Smartphones. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 894–905. https://doi.org/10.1145/3366423.3380169

[17] Arvind Narayanan, Eman Ramadan, Jacob Quant, Peiqi Ji, Feng Qian, and Zhi-Li Zhang. 2020. 5G Tracker – A Crowdsourced Platform to Enable Research Using Commercial 5G Services. In *Proceedings of the ACM SIGCOMM 2020 Conference Posters and Demos* (Virtual Event, USA) *(SIGCOMM Posters and Demos '20)*. Association for Computing Machinery, Virtual Event, USA. https://doi.org/10.1145/3405837.3411394

[18] Arvind Narayanan, Xumiao Zhang, Ruiyang Zhu, Ahmad Hassan, Shuowei Jin, Xiao Zhu, Xiaoxuan Zhang, Denis Rybkin, Zhengxuan Yang, Z. Morley Mao, Feng Qian, and Zhi-Li Zhang. 2021. A Variegated Look at 5G in the Wild: Performance, Power, and QoE Implications. *ACM SIGCOMM'21* (2021).

[19] Feng Qian, Bo Han, et al. 2019. Toward Practical Volumetric Video Streaming on Commodity Smartphones. In *HotMobile*. https://doi.org/10.1145/3301293.3302358

[20] J. Qiao, Y. He, and X. S. Shen. 2016. Proactive Caching for Mobile Video Streaming in Millimeter Wave 5G Networks. *IEEE Transactions on Wireless Communications* 15, 10 (2016), 7187–7198. https://doi.org/10.1109/TWC.2016.2598748

[21] Qualcomm. 2021. Everything You Need to Know About 5G. https://www.qualcomm.com/invention/5g/what-is-5g. Last accessed July 2021.

[22] Theodore S Rappaport, Felix Gutierrez, et al. 2013. Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications. *IEEE transactions on antennas and propagation* 61, 4 (2013), 1850–1859.

[23] Theodore S Rappaport, Shu Sun, et al. 2013. Millimeter wave mobile communications for 5G cellular: It will work! *IEEE access* 1 (2013), 335–349.

[24] Ruwen Schnabel and Reinhard Klein. 2006. Octree-Based Point-Cloud Compression. In *Proceedings of the 3rd Eurographics / IEEE VGTC Conference on Point-Based Graphics* (Boston, Massachusetts) *(SPBG'06)*. Eurographics Association, Goslar, DEU, 111–121.

[25] Aaron Schulman, Vishnu Navda, Ramachandran Ramjee, Neil Spring, Pralhad Deshpande, Calvin Grunewald, Kamal Jain, and Venkata N Padmanabhan. 2010. Bartendr: a practical approach to energy-aware cellular data scheduling. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 85–96.

[26] Heiko Schwarz, Detlev Marpe, et al. 2007. Overview of the scalable video coding extension of the H. 264/AVC standard. *IEEE Transactions on circuits and systems for video technology* 17, 9 (2007), 1103–1120.

[27] Susanna Schwarzmann, Clarissa Cassales Marquezan, et al. 2019. Estimating Video Streaming QoE in the 5G Architecture Using Machine Learning. In *Internet-QoE.*

[28] Yi Sun, Xiaoqi Yin, Junchen Jiang, Vyas Sekar, Fuyuan Lin, Nanshu Wang, Tao Liu, and Bruno Sinopoli. 2016. CS2P: Improving video bitrate selection and adaptation with data-driven throughput prediction. In *Proceedings of the 2016 ACM SIGCOMM Conference*. 272–285.

[29] Sanjib Sur, Vignesh Venkateswaran, et al. 2015. 60 GHz indoor networking through flexible beams: A link-level profiling. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 43. ACM, 71–84.

[30] Zhi-Li Zhang, Udhaya K. Dayalan, Eman Ramadan, and Timothy J. Salo. 2021. Towards a Software-Defined, Fine-Grained QoS Framework for 5G and Beyond Networks. In *Proceedings of the ACM SIGCOMM Workshop on Network Meets AI & ML (NetAI'21)*.

[31] Hang Zhao, Rimma Mayzus, Shu Sun, et al. 2013. 28 GHz millimeter wave cellular communication measurements for reflection and penetration loss in and around buildings in New York city. In *ICC*. 5163–5167. https://doi.org/10.1109/ICC.2013.6655403

[32] Xuan Kelvin Zou, Jeffrey Erman, Vijay Gopalakrishnan, Emir Halepovic, Rittwik Jana, Xin Jin, Jennifer Rexford, and Rakesh K. Sinha. 2015. Can Accurate Predictions Improve Video Streaming in Cellular Networks?. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications* (Santa Fe, New Mexico, USA) *(HotMobile '15)*. Association for Computing Machinery, New York, NY, USA, 57–62. https://doi.org/10.1145/2699343.2699359

A APPENDIX

A.1 Theoretical Bounds for Choosing Video Quality for 5G Throughput

Increasing the video quality (number of points for each frame) leads to increasing the stall time if the current network conditions can not support the requested quality as the client's buffer would not be able to maintain a threshold number of frames. Thus, selecting the appropriate quality given the network conditions is crucial as it impacts the user's QoE. We attempt to answer this question by considering an *ideal* case where we have *perfect knowledge* of the available 5G network throughput over a period of time, and derive theoretical bounds on the best video quality we can achieve without any stalls.

Suppose we start streaming a video of length T seconds at time t_{start}. With a start delay of d seconds, the playback begins at $t_1 = t_{start} + d$, and ends at $t_{end} = t_1 + T$. Let F be the frame rate (e.g., $F = 30$); $n = T * F$ is the total number of frames to be played, with a rate of one frame played every $1/F$ seconds. (We will use τ_k, $k = 1, ..., n$, to denote the playback time of the kth frame, where $\tau_1 = t_1$ and $\tau_n = t_{end}$.) Given a trace of available 5G bandwidth from t_{start} to t_{end} (see Fig. 1 for example), we are interested in finding out what is the *best achievable video quality* Q defined as the *highest constant* (thus the "smoothest") bitrate *without any stalls*. We obtain the following theorem for the upper- and lower-bound of Q using content bursting to fully utilize the available bandwidth.

Theorem 1. *Given a trace of (instantaneous) network throughput rate* $b(t)$, $t_{start} \leq t \leq t_{end} (= t_{start} + d + T)$, *let* $B(t) = \int_{t_{start}}^{t} b(t)dt$. *Then the highest achievable constant bitrate without any stall is given by* $Q_* \leq Q \leq Q^*$, *where* $Q_* = \min_{1 \leq k \leq n} B(\tau_k)/k$ *and* $Q^* = B(t_{end})/n$, *where* $n = T * F$.

We remark that in the statement of the theorem, we are ignoring the network latency (and round trip delays) between a mobile client and a video streaming server. We are essentially assuming that this

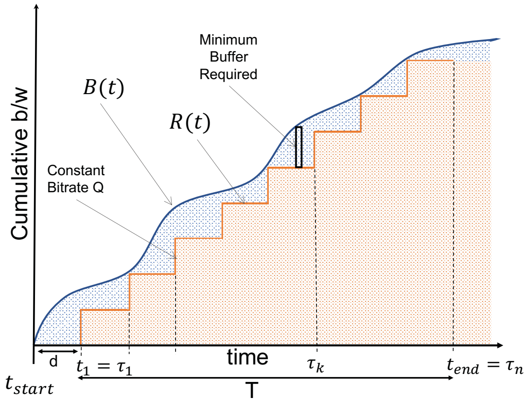


Figure 6: Illustration of the Proof

latency is negligible, e.g., when the video streaming server is located in a mobile edge cloud (e.g., co-located with the cell towers) very close to the mobile user. With a non-negligible network latency λ, we need to subtract λ and use, e.g., $t'_{end} = t_{end} - \lambda$, in the statement of the theorem so as to ensure the last bit of a kth video frame has arrived at the mobile client side before its scheduled playback time τ_k.

Proof of Theorem 1. The proof is illustrated in Fig. 6, where we have plotted the cumulative throughput $B(t)$, $t_{start} \leq t \leq t_{end}$ $(= t_{start} + d + T)$ as a function of time. Note that given a constant bitrate video of quality Q, namely, each frame contains Q bits, the total amount of bandwidth required for the video delivery at this level is $Q * n$, where n is the total number of frames in the video. Since $B(t_{end})$ is the maximum cumulative network bandwidth available between times t_{start} and $t_{end} = \tau_n$, the maximal video quality achievable is at most $Q^* = B(t_{end})/n = B(\tau_n)/n$. More generally, we note that by τ_k (the playback time at the kth frame, at least $Q*k$ amount of data must have been delivered to the client in order for the client player not to stall. In other words, we must have $B(\tau_k) \geq Q * k$. It is not hard to see the minimal video quality level we can achieve *with no stalls* is given by $Q_* = \min_{1 \leq k \leq n} B(\tau_k)/k$. The minimum buffer size required to avoid stalls while serving frames with quality Q is $buf = \max_{t_{start} \leq t \leq t_{end}} (B(t) - R(t))$ which represents the maximum difference between the two curves $B(t), R(t)$. □

When dynamic 5G/4G switching is employed along with content bursting, this is equivalent to using a modified network throughput trace $\bar{b}(t)$, $t_{start} \leq t \leq t_{end}$ which uses the maximum value of the 5G throughput and the 4G throughput. The theoretical bounds can then be obtained via Theorem 1 with $\{\bar{b}(t)\}$.

A.2 Streaming Variable Quality Levels

A video can be delivered using either: i) a constant bitrate level which requests all frames with the same quality (i.e., same number of points per frame); or ii) variable bitrate levels in which the video player switches between different quality levels for different frames. This decision depends on the network condition, its variability, and the buffer occupancy. Thus, instead of using the minimum constant bitrate level to avoid stalls as defined by Theorem 1, the video bitrate level can change over time according to the predicted throughput with the goal of eliminating stalls while maintaining video quality smoothness (i.e., avoid bitrate fluctuations which degrade user's QoE). For example, when the user mobility mode (still, walking, driving) changes, the mobility speed affects the available bandwidth. Hence, instead of prefetching frames with a very high quality, a more judicious decision can be made based on the predicted future bandwidth to decide which quality to use to avoid stalls. Thus, avoid requesting frames with the highest quality which yields only few frames in the buffer that will be depleted quickly. The goal is to develop an adaptive algorithm which can avoid stalls while at the same time deliver the highest possible quality with smooth quality variation instead of frequent changes from the highest quality to the lowest quality.

Theorem 1 not only demonstrates how to obtain bounds on achievable best video qualities, but also hints on how we may perform adaptive bitrate (ABR) selection for achieving best video qualities given bandwidth prediction for the upcoming X seconds. At the current time t, given the predicted network bandwidth $\tilde{b}(t)$ over $(t, t + \Delta t]$. Using the predicted total available bandwidth $B(t, t + \Delta t) = \int_t^{t+\Delta t} \tilde{b}(t)dt$, we employ Theorem 1 to determine the best video qualities for the next $\Delta k = \Delta t * F$ frames to be fetched. To account for uncertainty in the bandwidth prediction, a more conservative approach can be followed to assign priorities (or "deadlines") for fetching (future) content of different qualities: by prioritizing using the current (stable) available bandwidth to burst lower qualities of future Δk frames first than using it to increase the qualities of more recent frames. This will ensure a minimal video quality to users with *no* stalls while "smoothly" adapting to higher qualities whenever possible. This illustrates the power and utility of ML bandwidth prediction in enabling new mechanisms for 5G-aware applications to utilize the ultra-high bandwidth of 5G and overcome its wild fluctuation and dead zones. This deserves a separate paper to explore these design decisions and find the optimal algorithm.