Towards a Software-Defined, Fine-Grained QoS Framework for 5G and Beyond Networks

Zhi-Li Zhang, Udhaya Kumar Dayalan, Eman Ramadan, Timothy J. Salo
{zhzhang,eman}@cs.umn.edu,{dayal007,salox049}@umn.edu
Department of Computer Science & Engineering, University of Minnesota – Twin Cities, USA

ABSTRACT

5G offers a slew of new features and capabilities to support a whole gamut of new applications. On the other hand, 5G new radio (NR), especially, high-band mmWave radio, also poses new challenges, as shown by recent measurement studies of commercial 5G services. In order to effectively support new classes of application such as extra low-latency and/or high-bandwidth applications, we argue that *truly cross-layer* network-application integration that *exposes application semantics to enable 5G and beyond 5G (B5G) networks to make intelligent decisions*, e.g., for dynamic radio resource allocation, is needed. Unfortunately the existing 5G flow-based framework is inadequate to support such cross-layer integration. We therefore advocate a *software-defined*, *fine-grained* QoS framework. We use ultra-high resolution (UHR) volumetric video streaming as a use case and conduct very preliminary experiments to demonstrate the potential benefits of the proposed framework. This *position* paper serves as a *strawman* to call for new intelligent architectural designs for B5G networks and next-generation wireless systems.

CCS CONCEPTS

• **Networks** → Programming interfaces; *Application layer protocols*; **Mobile networks**; *Wireless access points, base stations and infrastructure.*

KEYWORDS

QoS framework, 5G and beyond, application semantics, software -defined, fine-grained

ACM Reference Format:
Zhi-Li Zhang, Udhaya Kumar Dayalan, Eman Ramadan, Timothy J. Salo. 2021. Towards a Software-Defined, Fine-Grained QoS Framework for 5G and Beyond Networks. In *ACM SIGCOMM 2021 Workshop on Network-Application Integration (NAI '21), August 27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3472727.3472798

1 INTRODUCTION

Unlike earlier generations of cellular technologies, emerging 5G networks are designed to enable a whole gamut of diverse new use cases from massive consumer/industrial IoT devices to control, safety and other V2X (vehicle-to-everything) applications for autonomous vehicles (AVs) and drones to ultra-high-resolution (8K & volumetric) live video streaming, augmented/virtual reality (AR/VR), telemedicine and healthcare. To this end, 5G New Radio (NR) incorporates a wide spectrum of radio bands, from sub-1GHz spectrum bands (5G *low-band*), to 1 GHz – 7.125 GHz radio bands (5G *mid-band*) and 24GHz – 60 GHz (5G *high-band* including mmWave radio bands), and introduces a number of innovations such as flexible numerology and frame structures, dynamic slot/min-slot scheduling, carrier aggregation and so forth to support ultra-low latency, ultra-reliable and ultra-high bandwidth applications.

All these ("real" and "yet-to-be-realized") new capabilities and features of 5G notwithstanding, they also give rise to two important (and *intertwined*) questions: 1) Are the current 5G architectural designs and associated new capabilities/features sufficient to support various *envisioned* 5G applications and services, some of which demand high bandwidth, other ultra-low latency and high reliability, yet other massive connectivity, or any combination of the above. 2) If deemed insufficient, what are the crucial missing pieces?

In this *position* paper we explore these questions by considering ultra-high-resolution (UHR) video streaming over 5G as a case study. This is partly motivated by our recent measurement studies of commercial (especially, mmWave) 5G services [14–16] and our findings that existing adaptive bit rate (ABR) algorithms for video streaming cannot effectively cope with the fast and highly varying mmWave 5G throughput, leading to poor user quality-of-experience (QoE) [16, 18]. In particular, we argue that *the current 5G QoS architecture is inadequate in effectively exploiting new capabilities of 5G NR to support UHR video streaming over 5G.* While improving upon the *fixed* data-radio-bearer (DRB)-based QoS service mapping of 4G LTE, the current 5G standard adopts a *flow-based* QoS framework and introduces a new *service data adaption protocol* (SDAP) sublayer to support QoS: the framework specifies a set of *predefined* QoS profiles with fixed QoS metrics, and allows user-plane functions (UPFs) to mark IP flows with appropriate SDAP QoS flow identifiers (QFIs) in accordance with their QoS profiles.

To exploit new capabilities such as diverse radio bands and carrier aggregation, we advocate the use of *scalable layered coding* (SVC) (see §3 for the rationale) to effectively cope with the high throughput variability posed by 5G high frequency bands such as mmWave which provides the ultra-high bandwidth needed for UHR video streaming. For example, when streaming SVC video over mmWave 5G, one can use a clear Line-of-Sight (LoS) beam (when available) to deliver the base layer video chunks, while at the same time utilize non-LoS beams to deliver enhanced layers based on available radio resources (see §3 for more discussion). However this is difficult, if not infeasible, to realize using the *flow-based* 5G

NAI '21, August 27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8633-3/21/08…$15.00
https://doi.org/10.1145/3472727.3472798

QoS framework: i) It is *too coarse* and *inflexible* – it cannot differentiate the "sub-flows" (the base and enhancement video layers and chunks) within the same video session (the "IP flow"). ii) With its pre-defined (and thus fixed) QoS profiles, it is also *carrier-centric* – it cannot exploit *application semantics* for *intelligent* decision making, e.g., dynamically allocate available radio resources to maximize user QoE.

This leads us to advance a *software-defined, fine-grained* QoS framework to better support *network & application integration* over 5G (and beyond) networks. The basic idea is two-fold: The first is to enable an application (or application service provider) to specify application semantics tags (for data substreams or objects) and their associated QoS profiles or metrics, negotiate and signal them to the 5G carrier, and (dynamically) mark application data (with "semantic tags") accordingly. Secondly, using such (*service-specific*) semantics tags and QoS profiles, the 5G carrier can install appropriate UPFs in its 5G core network to classify/filter the application data packets and set QFIs; using these QFIs, the 5G radio access network can intelligently map radio bands, channels or beams with differing characteristics to appropriate data streams/objects, and dynamically allocate fast varying radio resources to transport the right (amount/type of) data for best user QoE while maximizing radio resource efficiency. We outline a *basic* design of the proposed framework and its key components in §4.

To demonstrate the potential benefits of the proposed framework, we conduct a preliminary evaluation using trace-driven emulation. We have implemented a few simple 5G core functions lifted from free5gc [8], where we emulate the radio network performance using real-world 5G throughput measurement traces collected in [15]. Admittedly, our design is still rather crude, with many details left unspecified; our evaluation is also very preliminary. Our goal of this position paper is to shed light on the need for a *software-defined*, *fine-grained* QoS framework that can truly enable *cross-layer*, network-application integration to better support diverse new applications and services over 5G and beyond networks.

2 BACKGROUND

As mentioned in the introduction, 5G New Radio (NR) supports a diverse range of radio bands and introduces a number of innovations, especially to accommodate high frequency bands. For example, while it retains the same OFDM waveforms and frame structure (10 ms frame with 1 ms subframe) as 4G, 5G NR supports flexible subcarrier spacings (SCS), dubbed *numerologies*, to allow carriers to meet varying bandwidth needs, and defines (SCS-dependent) slots and mini-slots to allow (dynamic & preemptive) data transmissions without preserving frame/slot boundaries to support low latency. 5G NR also introduces *bandwidth parts* (BWPs) so that each user equipment (UE) or end device can operate on only part of the 5G radio bands. We refer the reader to [2] and other 3GPP specifications for more detailed disposition. In the following we will provide a very brief overview of the 5G radio (access) network (RAN) protocol stack (see Fig. 1), focusing in particular on the 5G QoS framework and mechanisms that are most relevant to the theme of this paper.

The 5G radio network protocol stack resides below the OSI network layer ("IP layer"), and its functions are performed primarily by the (logical) 5G nodeB elements (gNBs). Compared with 4G

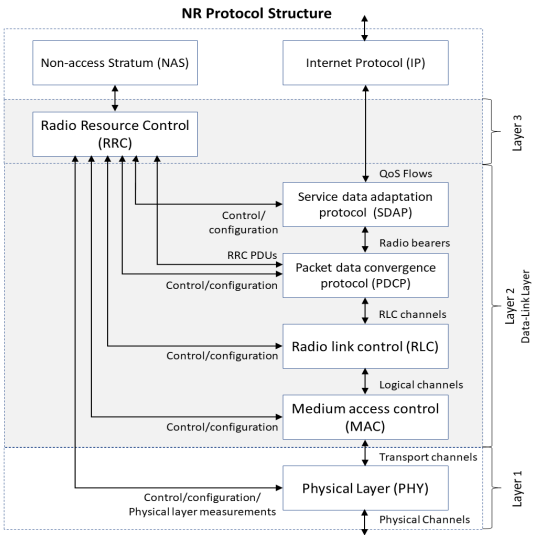


Figure 1: 5G Radio Network Protocol Stack.

LTE, 5G introduces a new SDAP sublayer to support its *flow-based* QoS framework, where the SDAP header includes a QFI (QoS flow identifier) field as well as ARP (allocation & retention priority) bits for admission control and/or resource preemption indication. A PDN (packet data network) connection or session carrying upper layer data (i.e., from an application in the form of IP flows that are transported over the 5G core network using NG-U tunnels) is classified based on a QoS profile. 5G defines a fixed set of QoS profiles (and thus the *service semantics*) using a set of parameters such as 5G QoS Identifier (5GI) with *pre-defined* values, indicating, e.g., the guaranteed flow bit rates, maximum bit rates, maximum packet loss rates. The 5G core network control plane functions as well as the RRC sublayer are responsible for selecting QoS profiles for IP flows (PDN sessions), setting up flow classification rules in the 5G core user plane and configuring UPFs to mark IP flows with QFIs. They are then mapped to appropriate logical data radio bearers (DRB) at the PDCP sublayer. After header compression and other operations, data packet units (PDUs) are passed down to the RLC/MAC/PHY sublayers for transmission over the radio network.

The 5G RRC sublayer is responsible for radio resource signaling and allocation, and instructs the MAC sublayer to perform semi-persistent scheduling (SPS) and dynamic scheduling of radio resources and transmission time intervals (TTIs). The radio resources (over various radio bands and channels) are logically organized (one for each SCS) in grids of resource blocks (RBs) consisting of 12 contiguous subcarriers; a resource block (RB) consisting of 12 contiguous subcarriers is the basic unit of radio resource allocation and scheduling. To allow diverse deployment scenarios and *heterogeneous* networks (e.g., with macro and small cells), 5G supports both stand-alone mode (SA) and non-stand-alone (NSA) mode with 4G & 5G dual connectivity; each gNB can be configured with multiple cells organized in primary and secondary cell groups and cells. 5G allows *carrier aggregation* of multiple channels to achieve higher data rates, e.g., by aggregating *intra-band contiguous* channels, *intra-band non-contiguous* channels, or *inter-band* channels with diverse frequencies and radio characteristics.

3 CASE FOR INTEGRATING 5G AND APP

5G is designed with many new capabilities and features such as a wide range of radio bands, carrier aggregation, dynamic MAC and mini-slot/slot scheduling. It is – at least *in theory* – capable of delivering ultra-high (aggregated) peak bit rates and ultra-low latency. How can we leverage these capabilities and features of 5G, especially 5G NR, to effectively support a whole gamut of envisioned new applications and services? Before we answer this question, we first examine some *real-world* measurement results of commercial (especially, mmWave) 5G services (see [14–16] for more details).

Fig. **??** shows measured downlink mmWave 5G throughput as seen by a mobile device or user equipment (UE) when it is stationary and has a clear line-of-sight (LoS) path to a 5G antenna, as a function of the distance from UE to a server (the figure is from [16], where more details can be found). We see that when the server is fairly close-by (within 1000 km) or when multiple TCP connections are used, mmWave 5G can deliver more than 1 Gbps (up to 3.5 Gbps) bandwidth. However, high frequency radio bands such as mmWave radio are directional, have limited coverage ranges, and are highly sensitive to obstruction, interference and other environmental factors which may block, reflect, refract and attenuate the signals; user location, orientation and mobility modes thus greatly impact 5G performance he/she experiences (see [15] for a more in-depth measurement-based analysis on the impact of these factors on mmWave 5G throughput performance). Hence as shown Fig. **??**, in general 5G throughput performance can vary wildly, sometimes as high as 2 Gbps and other times dropping to near zero (5G "dead zones"), when there is no clear LoS path due to obstructions or when user/UE moves around. Frequent handoffs, e.g., between 5G and 4G under the NSA mode, may also occur.

The *high variability* in 5G throughput performance poses many challenges to many *bandwidth-intensive* new applications such as ultra-high resolution (UHR) volumetric video streaming and AR/VR. Today's video streaming applications primarily utilize an AVC (Advanced Video Coding) codec such as H.264 for video encoding: a video is typically segmented into chunks (e.g., of 2 seconds long), and each chunk is encoded separately into multiple quality levels with varying sizes. To adapt to the changing network bandwidth, an *adaptive bit rate* (ABR) algorithm is employed at the client side to dynamically select the quality level of a future video chunk to fetch from the video server. Unfortunately existing ABR algorithms based on (slow time scale) *application-layer* bandwidth estimation and bit rate adaption cannot effectively cope with the fast and wildly varying 5G radio channels, which lead to large video stall times and poor user QoE, as shown in [16, 18].

How can one effectively overcome the challenges posed by 5G high-band mmWave radio such as wildly fluctuating bandwidth and limited ranges to support *bandwidth-intensive* applications such as UHR video streaming? Clearly, relying on a single 5G high-band radio channel or a single directional mmWave beam – the condition (and thus bandwidth of) which may vary rapidly and significantly – cannot *reliably* deliver the ultra-bandwidth needed for UHR video streaming. On the other hand, our measurement studies [14, 15] also show when even without clear LoS to 5G antennas, mmWave radio may still be able to deliver far higher throughput (e.g., over 500 Mbps) than 4G LTE (see Fig. **??**), due to signal reflections along non-LoS paths. Hence a promising direction for tackling these challenges is to take advantage of the new capabilities offered by 5G such as diverse radio bands and carrier aggregation [18]. However, exploiting these capabilities to support today's AVC-encoded video streaming is challenging. First, as stated earlier, relying on the chunk level bit rate adaptation at the application layer is too slow to cope with rapidly varying channel conditions. Second, simply transmitting the video chunk data using an aggregation of diverse radio channels (with sufficiently high total bandwidth capacity) may not lead to an improvement of QoE performance at all, thanks to the fact that these channels (especially across different radio bands) may have very different characteristics, e.g., radio range, signal strength/directionality, block bit error rates, etc. Hence the channel with the *poorest* quality will determine the time that the entire chunk can be completely delivered, which may in fact render the performance worse than using a single high-capacity channel.

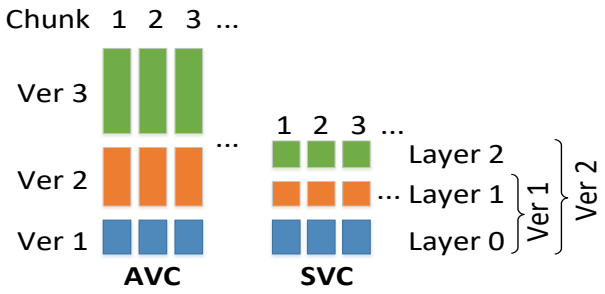


Figure 2: AVC vs. SVC.

We advocate instead the use of *Scalable Video Coding* (SVC) – which encodes each video (and video chunks) progressively into multiple layers (see Fig.2) – for streaming UHR videos over 5G to cope with the high variability of 5G throughput performance. However, *SVC by itself is not sufficient!* SVC-based ABR algorithms have been proposed in the literature (see, e.g., [6, 11, 26]) – operate purely *at the application layer* to select video layers and deliver them *sequentially*, and thus suffer the same issues facing the AVC-based ABR algorithms; in fact, due to the large encoding overhead, (application-layer) SVC-based ABR algorithms rarely outperform AVC-based ABR algorithms. In contrast to existing (*radio-agnostic*) SVC-based ABR algorithms, we advocate a truly *cross-layer* approach to take full advantage of the new capabilities of 5G: i) Instead of using *sequential* (layer-by-layer) delivery as in existing SVC-based ABR algorithms, the application will transmit multiple layered video chunks (i.e., belonging to the base layer plus one or more enhancement layers) *simultaneously*; the number of layers may be decided either based on the estimated bandwidth provided by the 5G network or predicted by the application [15]. ii) The 5G radio network will intelligently match and map the layered video chunks of differing utility to diverse radio channels of varying qualities and dynamically allocate radio resources for their transmission. For example, a clear LoS beam is allocated for the base layer video chunk delivery, while non-LoS beams are used for "best-effort" delivery of enhancement layers, when a user is stationary. As another example, when a user is mobile, a 5G mid-band channel with higher quality and larger coverage range is allocated for transporting the

base layer video chunks, while high-band channels with fast varying conditions and higher bit errors are assigned for enhancement layer video chunks. (See § 6 for other potential use cases.)

In order to support such *cross-layer* network and application integration and allow the application to signal the *utility* of application data, it is imperative to *expose application semantics to 5G (and beyond) networks*, and enable RAN radio resource control to *make intelligent, swift decisions in delivering the right type/amount of data with maximum utility to applications*. Unfortunately, the existing flow-based 5G QoS framework with pre-defined QoS profiles is too coarse and too rigid to allow such *cross-layer network and application integration*. More specifically, video data belonging to different layers all belong to the same IP flow, thus assigned with the same QFI. As a result, the 5G RAN cannot distinguish the (finer-grained) layered video data to provide differential QoS treatment. This motivates us to propose a new QoS paradigm for 5G and beyond 5G (B5G) networks.

4 FINE-GRAINED QOS FRAMEWORK

We advance a *software-defined*, *fine-grained* QoS framework for 5G/B5G networks. In the following we outline our proposed framework and present its key components.

4.1 Framework Overview

The overall architecture of our framework is schematically sketched in Fig. 3, where we have depicted the relations between a 5G carrier (with its constituent core network and radio access network) and an application service provider (with its service controller and server/client end points). This figure also shows how our QoS framework can be integrated with the 5G core architecture and RAN protocol stack. The basic premise is that an application service provider enters into a *cooperative agreement* (i.e., a business relationship with certain financial or other arrangements) with a 5G carrier to *collaboratively* provide *(application-)semantics-aware, cross-layer* support for its application or service over 5G.

Our proposed QoS framework is *software-defined* in that it follows the same principles of software defined networking (SDN), where the behavior of data/user plane functions is controlled and programmed by the control plane – via *service-specific QoS tables*, extending the SDN flow tables. On the other hand, our framework is far more *flexible* and *fine-grained*: instead of using solely (predefined) *flow headers*, both the QoS control and data/user functions are *service-specific* and specified using *application semantics* (via "semantic tags") in addition to standard flow headers; this also allows the applications (controllers and service end points) to dynamically push and update the semantics manifests and QoS profiles (and thus QoS tables) to the 5G control and user planes. To these ends, we take advantage of "softwarization" of 5G/B5G networks and network function virtualization (NFV) to support the needed functionality. Our design goal is two-fold: i) to enable the 5G radio network (and 5G core network) to intelligently allocate and match available radio resources (e.g., channels or beams with appropriate qualities) to application semantics and QoS requirements, perform smart scheduling and other adaptive mechanisms; and ii) to allow application/service end points (either at the server or client side or both) to not only dynamically adapt to varying network conditions, but also fully take advantage of available network resources to meet service QoS objectives and deliver the best QoE to users.

4.2 Key Components

We now briefly describe the roles of the application service provider and 5G carrier in our proposed QoS framework, and the basic functions of the key components shown in Fig. 3.

•**Application Service Provider (ASP).** We assume that the application service provider operates within a mobile edge cloud (MEC) within or close to the 5G carrier network for reduced latency. In other words, it has servers running inside the MEC with direct connectivity to the 5G core network. The application service provider negotiates with the 5G carrier control plane (via an ASP *controller*), and provides it with (*service-specific*) application semantics manifests and QoS profiles, e.g., in the form of XML or JSON files (similar to the manifest files used in video streaming applications). These files specify *semantic tags* which will be used by its applications to define (*fine-grained*) data objects or data streams such as layered video frames or chunks and mark them with appropriate semantic tags for the (desired) QoS metrics and treatments over 5G. The ASP may also supply the 5G carrier with *application functions* (AFs) for service-specific data processing, e.g., deep packet inspection, data conversion, classification, filtering and tagging. The semantics manifests and QoS profiles may be dynamically updated and pushed to the 5G network using out-of-band signalling [29], e.g., by altering the number of data streams such as SVC layers based on bandwidth prediction or feedback from the 5G network.

•**5G/B5G Control Plane.** Based on the business agreement with the ASP, the 5G core network control plane of the 5G carrier will define QFI values for the desired service objectives and QoS metrics specific to this ASP and its service, and institute appropriate control functions, such as SMFs and AMFs to set up and authenticate PDN sessions for application flows and track the mobility of mobile client end points, and PCFs (policy control functions) to install *service-specific* QoS tables at the relevant UPFs in the user plane. Similar to the flow tables used by SDN switches, the QoS flows extends them to include (service-specific) semantics tags to map IP flows into finer-grained QoS "subflows" or QoS data streams. Each entry of a QoS table is of the form ⟨ flow header; semantic tags | QFI ⟩ (including perhaps also ARP bits). The 5G core control plane also instructs and configure the 5G RAN for *intelligent* radio resource control (RRC) functions.

•**Application Server Endpoint.** The ASP Server Endpoint generates and (re-)factors application data into fine-grained data streams, marking them with the corresponding semantic tags specified by the application semantics manifests. These semantic tags can be implemented using, e.g., IPv6 flow labels or extension headers [20].

•**5G/B5G Core User Plane.** Upon receiving the data packets from the application, the UPFs in the 5G user plane, assisted by service-specific AFs, will process them based on the (service-specific) QoS tables for the desired QoS treatments – in particular, convert and encapsulate them to 5G packet data units (PDU) with appropriate QFI values (and ARP bits).

•**5G/B5G Radio Network.** Once reaching the 5G RAN, the RRC functions will intelligently allocate radio resources (channels, beams, transmission intervals) based on channel quality and other characteristics (e.g., coverage range and reliability), invoke dynamic MAC

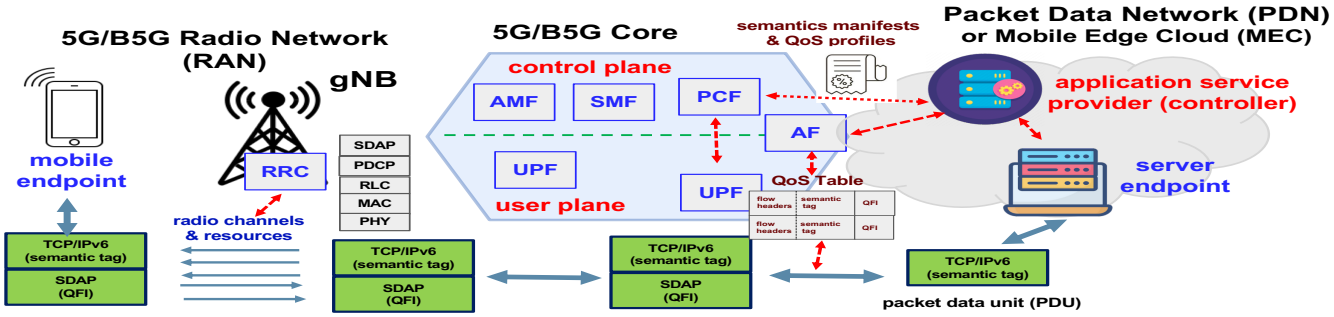


Figure 3: Software-Defined, Fine-Grained QoS Framework for 5G and Beyond 5G (B5G) Networks. As stated in the text, semantic tag can be implemented using IPv6 flow labels or extension headers, or other mechanisms.

scheduling algorithms, adaptive PHY-layer mechanisms, e.g., rate matching, MCS selection and hybrid ARQ (HARQ) to best match application data streams of differing utilities and QoS requirements with diverse and often fast varying radio characteristics.

•**End Device and Mobile App Endpoint.** The (mobile or fixed) end device, such as a smart phone, an autonomous vehicle or an IoT device, plays dual roles: on one hand, the ASP mobile app running on the device functions as the client endpoint; on the other hand, the end device is also equipped with an 5G radio interface that executes the radio protocol stack. For example, for uplink data transmission, the end device may mark application data using *reflective* QoS as in 5G [19]; it may also signal the data semantics in-band to the 5G RAN to effect changes in data transmission [20].

5 SVC VIDEO STREAMING USE CASE

We take SVC video streaming over 5G as a use case to illustrate the potential benefits of our proposed software-defined, fine-grained QoS framework by conducting a preliminary evaluation. As argued in §3, by exposing application semantics (via "fine-grained" data substreams marked by semantic tags, in this case, *layered video chunks*), our proposed framework will enable the 5G network to intelligently match and map the layered video chunks of differing utility to diverse radio channels of varying qualities, and dynamically allocate radio resources for their transmission so as to deliver the best user QoE. In this preliminary experimental evaluation, we consider only a very simple channel allocation scheme.

5.1 Preliminary Experimental Evaluation

We implement and set up a simple emulation environment to carry out a *preliminary* evaluation of our proposed framework. For the application service endpoints, we re-purpose IPv6 flow labels as "semantic tags", based on a similar implementation in [20]. To emulate a 5G core network, we originally based our implementation on free5gc [8] which implements the basic 5G core functions. But running on virtual machines (VMs) it can only process data at the rate of below 200 Mbps; it also does not support SDAP. Instead, we use a bridged network environment with a basic implementation of SDAP and a simple UPF mechanism which maps application semantic tags (IPv6 flow labels) to QFIs carried in SDAP. We use the real-world measurement traces of commercial mmWave 5G services collected in [1] to emulate different channel conditions and throughput performance.

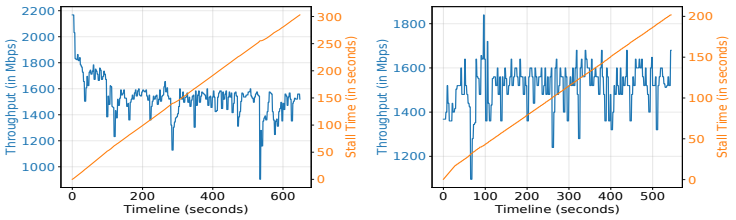
We consider volumetric video [10] streaming where video frames are represented as 3D point cloud. We encode the video frames using SVC with a total of 5 layers, a base layer with a resolution corresponding using roughly 1/5 of the total points per frame, and 4 enhancement layers which progressively enhance the frame resolution. We simulate and allocate up to 5 radio channels with differing channel qualities. In the case of the current **flow-based QoS** framework, data packets from the same frame will be *striped* across the allocated channels – i.e., *independent of the layers the data packets belong to*, as they all carry the same QFI value. In the case of our **fine-grained QoS** framework, we assign and transmit the base layer data packets using the radio channel with the best quality, the next (enhancement) layer data packet using the radio channel with the next best quality, and so forth.

5.2 Initial Results

In the following we present some initial results we have obtained. In Fig. 4 we show the throughout (bit rates) over time and the *cumulative* video stall time for streaming 500-sec volumetric video with 1750K point resolution, where the 5G channels qualities fluctuate wildly and constantly as in Fig. **??**. We see that using our fine-grained QoS framework we can reduce the total stall time from nearly 303 seconds to below 201 seconds. Table 1 summarizes the results of three experiments using volumetric videos with three different maximum point resolutions. We remark that our evaluation is still *very preliminary*. For example, in these experiments we use a very *basic* fixed channel assignment strategy where we match the base layer to the channel of the best quality *regardless of* the base layer throughput requirement and channel capacity. As a result, the base layer data may take longer to get delivered or "stall" when the average channel capacity of the best radio channel is below the required throughput of the base layer. Hence these initial results do not fully reflect the potential benefits of the proposed QoS framework. We are designing more intelligent radio resource control schemes for further experimentation and evaluation.

6 DISCUSSIONS AND RELATED WORK

Internet QoS has been extensively studied in 1990s (see [7] for a survey of Internet QoS theory. IETF has proposed two QoS architectures, Integrated Service (IntServ) [21] and Differentiated Services (DiffServ) [3], but neither is widely adopted and deployed. In [30] a (logically) centralized control framework was introduced for managing QoS, predating the notion of SDN. Following SDN, 5G adopts



(a) Flow-based QoS (b) Fine-Grained QoS

Figure 4: QoS Performance of Streaming a Video with 1750K Points/Frame.

Table 1: Stall Time for Video Playback.

Points/Frame	1500K	1750K	2000K
Required Tput	3240 Mbps	3780 Mbps	4320 Mbps
Flow-based QoS	136 sec.	303 sec.	348 sec.
Fine-Grained QoS	84 sec.	201 sec.	228 sec.

a flow-based QoS architecture (see [19] for an overview). We note in particular the recent IETF "Diffserv to QCI Mapping" effort, the goal of which, as stated in the current Internet draft [9], is to provide guidance to "help maintain a consistent QoS treatment between cellular networks and the Internet." In other words, this effort aims to "retrofit" IETF DiffServ with the *existing* 5G QoS architecture, whereas we have proposed a completely new *software-defined, fine-grained* framework, with the goal to enable a (cloud/application) service provider/network to *directly signal* 5G and B5G networks for *cross-layer network-application integration.*

Cross-layer design has been a major theme in wireless networks, but most studies largely rely on passing relevant information up or down the protocol stack to address specific problems, e.g., congestion control and related issues [4, 12, 17, 22, 23, 25, 27, 28]. None of them target a cross-layer QoS architecture. In [20] we propose a *semantics-aware* framework for mission-critical networks.

In this paper we have focused on SVC layered video streaming as the main use case for the proposed fine-grained QoS framework. There are many other important use cases. For example, HTTP-based web services often employ a persistent HTTP/TCP connection where various objects (e.g., embedded in multiple frames in a web page) are transported between a client and a server [5, 24]; our proposed QoS framework will allow a web service to signal the differing importance and utility, and intelligently transport them, e.g., using channels of different qualities, to improve the page loading time and overall service response time. Perhaps more exciting applications will be emerging V2X applications and other IoT services. For instance, instead of (statically) slicing the data into fixed categories, e.g., ULLRC, mMTC and eMBB, with (fixed) radio band allocation, our proposed QoS framework would allow more flexibility to enable autonomous vehicles and radio networks to decide what channel to be used for what service depending on the specific context, e.g., dynamically allocating mmWave radio bands as well as other channels for videos and other sensory data *all* used for a critical *safety* application, e.g., during a road accident. More generally, as inexpensive IoT devices proliferate, many of which are not 5G capable, IoT gateways (i.e., N3IWF gateways [13]) that can communicate with a diverse array of IoT devices on the one side, and with 5G radio & core networks on the other side will play a crucial role. Our proposed QoS framework will enable the N3IWF IoT gateways and 5G radio/core networks to dynamically signal data criticality and utility, and intelligently allocate radio resources based on IoT service semantics, as well as the latency, bandwidth and reliability requirements.

7 CONCLUSIONS AND FUTURE WORK

In this position paper we have laid out a *basic* design of a novel software-defined, fine-grained QoS framework for *truly cross-layer network and application integration.* Our framework enables applications to expose application semantics and allow 5G/B5G networks to exploit such application semantics for intelligent decision making, e.g., to dynamically match and allow radio channels of varying qualities to application utilities and QoS requirements. Clearly, our design as well as evaluation are still very preliminary. Significant efforts are needed to realize the proposed framework and demonstrate its full potential. This *position* paper serves as a *strawman* to call to the research community for new intelligent architectural designs of B5G networks and next-generation wireless systems.

ACKNOWLEDGMENTS

We thank our shepherd Qiao Xiang and the anonymous reviewers for their insightful comments and suggestions. This research was in part supported by NSF under grants CNS-1617729, 1618339, CNS-1814322, CNS-1836722, CNS-1831140 and CNS-1901103.

REFERENCES

[1] 2020. *University of Minnesota "5Gophers" project website.* Retrieved July 2021 from https://5gophers.cs.umn.edu/
[2] 3rd Generation Partnership Project. 2019. *Release 15.* Retrieved July 2021 from https://www.3gpp.org/release-15
[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. 1998. *IETF RFC 2475: An Architecture for Differentiated Services.* Retrieved July 2021 from https://datatracker.ietf.org/doc/html/rfc2475
[4] Neal Cardwell, Yuchung Cheng, C Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2016. BBR: Congestion-based congestion control. *Queue* 14, 5 (2016), 50.
[5] Yingying Chen, Ratul Mahajan, Baskar Sridharan, and Zhi-Li Zhang. 2013. A Provider's Perspective on Search Response Time. In *Proc. ACM SIGCOMM'13.*
[6] Anis Elgabli, Vaneet Aggarwal, Shuai Hao, Feng Qian, and Subhabrata Sen. 2018. LBP: Robust rate adaptation algorithm for SVC video streaming. *IEEE/ACM Transactions on Networking* 26, 4 (2018), 1633–1645.
[7] V. Firoiu, J.-Y. Le Boudec, D. Towsley, and Z.-L. Zhang. 2002. Advances in Internet Quality of Service. *Proceedings of IEEE,* Special Issue on Internet Technologies and the Convergence of Telecommunications Services, Vol.90, No.9 (2002), 1565 –1591.
[8] free5GC. 2021. *free5GC: an open-sourec 5G core network.* free5GC. Retrieved July 2021 from https://www.free5gc.org/
[9] J. Henry, T. Szigeti, and L. Contreras. 2020. *IETF Internet Draft: Diffserv to QCI Mapping (draft-henry-tsvwg-diffserv-to-qci-03).* Retrieved July 2021 from https://tools.ietf.org/id/draft-henry-tsvwg-diffserv-to-qci-03.html
[10] Mohammad Hosseini and Christian Timmerer. 2018. Dynamic Adaptive Point Cloud Streaming. In *Proceedings of the 23rd Packet Video Workshop (PV).* 6 pages.
[11] Yunzhuo Liu, Bo Jiang, Tian Guo, Ramesh K Sitaraman, Don Towsley, and Xinbing Wang. 2020. Grad: Learning for Overhead-aware Adaptive Video Streaming with Scalable Video Coding. In *Proceedings of the 28th ACM International Conference on Multimedia.* 349–357.
[12] Feng Lu, Hao Du, Ankur Jain, Geoffrey M Voelker, Alex C Snoeren, and Andreas Terzis. 2015. CQIC: Revisiting cross-layer congestion control for cellular networks. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications.* ACM, 45–50.

[13] 5G Security Assurance Specification (SCAS); Non-3GPP InterWorking Function (N3IWF). 2020. *Technical specification (TS): Release 17, Reference 33.520.* Retrieved July 2021 from https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3748

[14] 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

[15] Arvind Narayanan, Eman Ramadan, Rishabh Mehta, Xinyue Hu, Qingxu Liu, Rostand A. K. Fezeu, Udhaya Kumar Dayalan, Saurabh Verma, Peiqi Ji, Tao Li, Feng Qian, and Zhi-Li Zhang. 2020. Lumos5G: Mapping and Predicting Commercial mmWave 5G Throughput. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*. ACM, 176–193. https://doi.org/10.1145/3419394.3423629

[16] 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).

[17] Thomas Nitsche, Carlos Cordeiro, Adriana B Flores, Edward W Knightly, Eldad Perahia, and Joerg C Widmer. 2014. IEEE 802.11 ad: directional 60 GHz communication for multi-Gigabit-per-second Wi-Fi. *IEEE Communications Magazine* 52, 12 (2014), 132–141.

[18] Eman Ramadan, Arvind Narayanan, Udhaya K. Dayalan, Rostand A. K. Fezeu, Feng Qian, and Zhi-Li Zhang. 2021. Case for 5G-Aware Video Streaming Applications. In *Proceedings of the ACM SIGCOMM Workshop on 5G Measurements, Modeling, and Use Cases (5G-MeMU'21)*.

[19] Stefan Rommer, Peter Hedman, Magnus Olsson, Shabnam Sultana, and Catherine Mulligan. 2020. *5G Core Networks: Powering Digitalization.* Academic Press, San Diego, CA, Chapter 9, 203–216.

[20] Timothy J. Salo and Zhi-Li Zhang. 2020. Semantically Aware, Mission-Oriented (SAMO) Networks: A Framework for Application/Network Integration. In *Proceedings of the Workshop on Network Application Integration/CoDesign* (Virtual Event, USA) *(NAI '20)*. Association for Computing Machinery, New York, NY, USA, 41–42. https://doi.org/10.1145/3405672.3409490

[21] S. Shenker, C. Partridge, and R. Guerin. 1997. *IETF RFC 2212: Specification of Guaranteed Quality of Service.* Retrieved July 2021 from https://datatracker.ietf.org/doc/rfc2212

[22] Sanjib Sur, Ioannis Pefkianakis, Xinyu Zhang, and Kyu-Han Kim. 2017. Wifi-assisted 60 ghz wireless networks. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 28–41.

[23] Sanjib Sur, Xinyu Zhang, Parmesh Ramanathan, and Ranveer Chandra. 2016. BeamSpy: enabling robust 60 GHz links under blockage. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. 193–206.

[24] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. 2016. Speeding up Web Page Loads with Shandian. In *13th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2016, Santa Clara, CA, USA, March 16-18, 2016*, Katerina J. Argyraki and Rebecca Isaacs (Eds.). USENIX Association, 109–122. https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/wang

[25] Keith Winstein, Anirudh Sivaraman, and Hari Balakrishnan. 2013. Stochastic forecasts achieve high throughput and low delay over cellular networks. In *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*. 459–471.

[26] Siyuan Xiang, Min Xing, Lin Cai, and Jianping Pan. 2015. Dynamic rate adaptation for adaptive video streaming in wireless networks. *Signal Processing: Image Communication* 39 (2015), 305–315.

[27] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. 2020. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*. ACM, 479–494.

[28] Yasir Zaki, Thomas Pötsch, Jay Chen, Lakshminarayanan Subramanian, and Carmelita Görg. 2015. Adaptive congestion control for unpredictable cellular networks. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 509–522.

[29] Yunfei Zhang, Gang Li, Chunshan Xiong, Yixue Lei, Wei Huang, Yunbo Han, Anwar Walid, Y. Richard Yang, and Zhi-Li Zhang. 2020. MoWIE: Toward Systematic, Adaptive Network Information Exposure as an Enabling Technique for Cloud-Based Applications over 5G and Beyond (Invited Paper). In *Proceedings of the 2020 Workshop on Network Application Integration/CoDesign, NAI@SIGCOMM 2020, Virtual Event, USA, August 14, 2020*. ACM, 20–27. https://doi.org/10.1145/3405672.3409489

[30] Z.-L. Zhang, Z. Duan, L. Gao, and Y. T. Hou. 2000. Decoupling QoS Control from Core Routers: A Novel Bandwidth Broker Architecture for Scalable Support of Guaranteed Services. In *Proc. of ACM SIGCOMM 2000*. Sweden.