Intelligent Content Delivery over Wireless via SDN

Hyunwoo Nam^{*}, Doru Calin[†] and Henning Schulzrinne^{*}

*Department of Electrical Engineering, Columbia University, New York, NY, USA

[†]Bell Laboratories, Alcatel-Lucent, Murray Hill, NJ, USA

Abstract—On the Internet, most traffic is generated by content delivery applications such as video streaming, P2P and web surfing. With the proliferation of mobile applications that offer on-demand access to services, the ability to track end-users' perceived service quality has become a pivotal aspect for future networks such as 5G. Software defined networking (SDN) and network functions virtualization (NFV) in the mobile core offer unprecedented flexibility for opportunistic traffic steering, leading to advanced levels of quality-of-service (QoS) control. We leverage the SDN concept to dynamically control network traffic over wide area networks (WANs) from edge nodes of wireless networks (e.g., a PDN gateway in LTE), depending on changing network conditions and application types. Our SDN-based applicationaware routing system allows mobile network operators to achieve better utilization of their networks, service providers to improve customer satisfaction, and end-users to experience desirable service quality for various network applications. As for a prototype, we attach 802.11g to Mininet's virtual networks in order to show the feasibility of our approach¹.

Keywords—Software-defined Networking (SDN), Network Functions Virtualization (NFV), Quality-of-Service (QoS) in LTE, Content Delivery Network (CDN)

I. INTRODUCTION

Today's content providers take advantage of content distribution networks (CDNs) to speed up the delivery of their contents to clients. The purpose of using CDN is to assign a client to a geographically close delivery node. Once the connection is established, however, it is difficult to closely monitor the end-user's perceived quality-of-experience (QoE) and pinpoint a bottleneck in a network while content is delivered to the client. To provide an acceptable level of QoE to the end users, it is necessary to realize a tight cooperation between network operators and content providers, so as to a) have full access and visibility into end-to-end connectivity across all network segments, including content server side networks, transit Internet service providers (ISPs) and last mile networks of clients, b) dynamically control traffic flows based on network conditions, and c) monitor end-user's QoE in real time.

SDN/NFV have been proposed to advance a software-based approach to networking for more scalable, agile and innovative networks that can improve existing content delivery services. In this paper, we leverage an SDN platform to provide endto-end QoS from the perspective of LTE network operators, where they have full control over their mobile networks such as eNodeBs and packet data network gateways (P-GWs) but have limited access to the SDN controllers that are connected to the SDN-enabled WAN routers in the Internet. We assume that each WAN service provider operates its own SDN platform, and the SDN controllers are interconnected and exchange information about their respective domain. As shown in Figure 1, we assume that the WAN-SDN platform is already deployed in the Internet. Our proposed SDN application running on top of P-GW is designed to enforce the following two objectives:

- 1) Monitoring end-to-end network conditions: To monitor end-to-end network conditions, P-GW enables obtaining the outside network information in the Internet from the SDN controllers, using an existing SDN protocol such as OpenFlow [1]. It can also share the limited LTE network information such as LTE QoS parameters and Enhanced Packet System (EPS) bearer status with the SDN controller(s).
- 2) **Performing dynamic end-to-end flow control:** P-GW can dynamically control routing flows and perform flexible resource allocation at the wireless access nodes based on the end-to-end network conditions via SDN.

When a quality degradation is anticipated based on the endto-end network conditions provisioned via the SDN controllers. the SDN application in P-GW is designed to send queries to the controllers. Then the controllers diagnose the connected SDN enabled WAN routers to analyze network conditions (e.g., TCP throughput and packet loss rates) of the traffic flow, so as to pinpoint the bottleneck links in the network. In order to resolve the congestion, the SDN controller first determines the QoS budget, e.g., packet delay budget per network segment under each of the SDN managed area. Such criteria are used by the SDN controllers to determine whether any routing adjustment is needed or not. Further, the controller can dynamically *a*) update inbound routing paths among WAN routers in a network, b) change outbound routing paths by switching border routers in a network, or c) invoke a signal to redirect the client to another content delivery node with higher networking performance.

Our contributions can be summarized as follows. We first address the challenges of implementing intelligent content delivery on existing WAN-SDN and wireless access systems such as LTE and Wi-Fi. In accordance with the SDN concept, we then propose a novel application which unifies the control plane end-to-end, across the wireless access (e.g., LTE/Wi-Fi) and mobile core network, to dynamically control traffic flows based on changing network conditions and service types in real time. We design a prototype using Mininet [2]. As a proof of concept, our goal is to show the feasibility of how we can build an intelligent content delivery system using SDN over Wi-Fi networks.

¹This work was done during Hyunwoo Nam's 2014 summer internship at Bell Labs.



Figure 1: Intelligent content delivery system over LTE using SDN

II. RELATED WORK

With the advent of SDN/NFV, several researchers have been exploring the application-awareness and WAN routing controls in SDN, and the separation of control and data planes in mobile networks.

Application-awareness in SDN: Zafar et al. [3] propose a machine learning (ML) based application awareness that uses a crowd-sourcing approach to conduct traffic classification in SDN. Their prototype in HP Labs shows 94% accuracy on average on Android devices over wireless networks. Michael et al. [4] have shown an application awareness using deep packet inspection (DPI) in SDN-enabled networks. They track YouTube video streaming traffic, and dynamically change the routing paths while a video content is delivered to the client using SDN. Graham in Heavy Reading [5] has addressed the role of DPI in SDN. The white paper describes the challenge of implementing DPI and its use-cases in SDN.

WAN routing in SDN: Recently, a multi-protocol label switching traffic engineering (MPLS-TE) technology has received much attention in WAN-SDN. Ali et al. [6] have shown the feasibility of how MPLS-TE can be simply implemented over an SDN-enabled network using Mininet and NOX [7]. Saurav et al. [8] have shown a demo, where a networking application uses OpenFlow functions to provide application-aware aggregation and TE among WAN-routers in SDN. Sushant et al. [9] have shown an empirical study of WAN connecting data centers in SDN. They have implemented a centralized traffic engineering using OpenFlow, where the Internet backbone traffic is transferred among Google's data centers that are geographically distributed over the world.

SDN/NFV in mobile networks: Li et al. [10] have proposed software-defined cellular networks. They argue that SDN can be simply applied to cellular data networks, and elaborate

the challenges of providing scalability and QoS-based dynamic flow control in SDN-enabled mobile networks. Alcatellucent [11] has addressed the challenges and solutions of deploying NFV in a network. Siwar et al. [12] have proposed an OpenFlow-based control plane for LTE core networks. As for use-cases of their proposal, they argue that resiliency and load balancing are considered as key aspects of providing the seamless on-demand connectivity in the SDN-enabled LTE networks.

III. BACKGROUND

In this section, we briefly address the challenges of SDN/NFV for providing end-to-end QoS, and the existing QoS mechanisms in LTE.

A. Challenges of providing end-to-end QoS in SDN/NFV

The key function of SDN and NFV is the separation of control and data-forwarding planes. The control plane between service applications and network devices makes for more flexible and agile network service environments. The service providers can easily launch new applications, dynamically obtain network information from the control plane, and quickly control traffic flows for their needs. In order to provide endto-end QoS, however, there are still constraints that need to be addressed.

 Limitation on application-awareness: Such application recognition has become critical for enhancing QoS policybased flow control and security (e.g., preventing from spam and malicious applications) [5]. However, existing OpenFlow mainly focuses on conducting L2-4 inspection, lack of application awareness of higher layers such as L4-7. DPI has been proposed as one of possible solutions, but it requires edge node devices with high performance and may lead to high computation cost in terms of latency, especially for high bandwidth applications [3].

2) Lack of access to wireless networks: WAN-SDN has been proposed to ensure flexible control traffic flows among WAN routers in the Internet, so as to provide QoSbased services. However, although it has full control over the WAN routers, it has limited access to mobile networks such as 3G and 4G. Since most congestion of mobile Internet traffic is likely to occur in radio access networks (less likely on the fixed parts of the network such as transit networks and servers), the SDN controllers cannot fully provision the end-to-end network conditions without the support of the wireless core networks.

B. QoS in LTE and its challenges

To enhance service quality, an LTE system provides a comprehensive QoS, policy control framework and charging system, regarding user-profiles and application service types.

- 1) **QoS mechanisms in LTE:** When a subscriber requests QoS-enabled services such as an online gaming and VoIP, a traffic flow template (TFT) including a set of QoS policies is applied to the EPS bearer between the subscriber and P-GW (Figure 1) for policy and charging control enforcement. QoS parameters such as QCI, ARP, GBR, MBR, APN-AMBR and UE-AMBR are typically used for the wireless resource allocation at transport nodes (e.g., eNodeBs) when congestion occurs in the network.
 - QCI (QoS Class Identifier): The QCI specifies IP packet forwarding treatment such as scheduling weights, queue managements in transport nodes in LTE and link layer protocol configuration. Nine QCI values in total are standardized in terms of resource type (GBR and Non-GBR), priority, packet delay budget and packet error loss rate, depending on application service types.
 - ARP (Allocation and Retention Priority): When an LTE network is heavily congested and a new EPS bearer has been established, the LTE entities such as P-GW, S-GW and eNodeB take account of the ARP of the EPS bearer. Then, it decides to remove some existing bearers or accept / refuse a new bearer in the network.
 - GBR (Guaranteed Bitrate) and MBR (Maximum Bitrate): The dedicated bearer consists of two types, GBR and Non-GBR. In a GBR mode, the bandwidth of the bearer is guaranteed. It is serviced with MBR that indicates the maximum bitrate allowed in the LTE network. The bearer with Non-GBR is typically used for the best-effort services such as file downloading and web browsing. It has a lower priority when the network experiences congestion.
 - APN-AMBR (APN Aggregate Maximum Bitrate) and UE-AMBR (UE Aggregate Maximum Bitrate): In a Non-GBR mode, there is no bandwidth limit that controls the maximum bandwidth usage for each bearer. Instead, the APN-AMBR is used to limit the total maximum bitrate for all Non-GBR bearers in the same packet data network. The UE-AMBR represents the maximum possible bitrate that the subscriber can use in the LTE network. It is especially designed to pre-

vent the subscriber from consuming all the available bandwidth from others attached to the same eNodeB.

2) Lack of access to WANs in the Internet: Although LTE provides QoS services to an EPS bearer, it is still unable to provision the network conditions between the P-GW and the content delivery nodes in the Internet.

IV. INTELLIGENT CONTENT DELIVERY OVER WIRELESS USING SDN

We propose to implement an SDN northbound application on P-GW in LTE that communicates with the SDN controllers in order to deliver content intelligently. We take advantage of an existing LTE QoS system and combine with the strength of future SDN/NFV networks to deliver unprecedented QoE to the end users.

- Building an application-awareness: In an LTE network, various QoS parameters such as QCI, ARP, GBR and MBR are used to classify application types (e.g., video, voice, gaming and email) and provide QoS as needed. The key point of our solution is to share these LTE QoS parameters with the SDN controllers that have difficulty implementing the application-awareness. The recent studies have proposed various approaches using Machine Learning (ML) and crowdsourcing [3] to differentiate applications depending on traffic characteristics. However, implementing those techniques has not been proved as practical in an SDN network. With the QoS parameters defined in LTE, an SDN controller can provide more accurate and consistent end-to-end QoS, which can lead to a truly agile application-aware network.
- 2) Monitoring end-user's QoE in real time: To provide end-to-end QoS, it is necessary to track end-user's perceived quality while delivering contents to the clients. This part is challenging since there is no method to accurately measure the QoE from the point of view of network operators that have limited access to client applications and content server side networks. In an LTE, we can estimate the QoS and QoE factors using DPI, which is now a standard option in 3G gateway GPRS support nodes (GGSNs) and 4G LTE P-GWs [5]. With DPI, for example, P-GW enables running L4-7 inspection that provides not only context-awareness but also extracting metadata attributes (e.g., URL, file name, browser type, cookies, DNS queries, video codec, IMSI, SIP caller/callee, device type and operating system) and other data that is used to calculate networking performance statistics (e.g., delay, jitter and application response time) for each flow.
- 3) **Performing a dynamic end-to-end flow control from P-GW:** In the Internet, the SDN controller dynamically changes flow paths based on the real-time network conditions periodically reported from P-GW and SDN-enabled WAN routers. In an LTE network, P-GW also takes advantage of the outside network information obtained from the SDN controller, so as to enforce end-to-end QoS.
 - Optimal content server selection: P-GW sends queries to the SDN controller(s) to obtain network conditions on each path between the P-GW and each content delivery node. The information can be used for the applications on clients' devices to select the best available content server at the moment.



Figure 2: Policy and charging control (PCC) architecture

- Fast end-to-end connection control from P-GW: Via the SDN application, P-GW can directly send signals to the SDN controller to dynamically add/remove a flow in advance.
- Performing flexible resource allocation at the wireless access node: Based on the outside network information obtained from the SDN controller, P-GW can dynamically change the OCI values of some traffic flows, providing it is acceptable by the corresponding applications, to flexibly reallocate wireless resources among subscribers. For an example, an LTE user is experiencing slow Internet speed. If the congestion occurs from the outside network in the Internet and there are no alternative paths or content delivery nodes available at the instant, this indicates that there is no way to resolve the congestion neither inside the LTE networks nor through the outside networks in the Internet. In this case, P-GW may degrade the QCI value (e.g., to a less stringent value or even changing GBR mode to Non-GBR whenever applications can tolerate it), so other competing flows attached to the same eNodeB can have more chance at using the network resource.

A. Building an SDN Application on P-GW

In our proposed architecture, the SDN application is designed to take account of various flow information that can be obtained from P-GW in EPC (Evolved Packet Core) and shared with the SDN controllers distributed in the Internet.

- 1) P-GW and PCRF (Policy and Charging Rules Function): The P-GW is the interconnection between the EPC and the external IP networks. The primary roles of a P-GW are to conduct QoS policy enforcement for the service data flows (SDFs), packet filtering using DPI and charging support. Policy and Charging Enforcement Function (PCEF) in P-GW communicates with the PCRF that is in charge of policy control decision-making and flow-based charging functionalities. The PCRF provides the PCEF the Policy and Charging Control (PCC) rules that include SDF identification, policy control, and flowbased charging rules for each SDF. For applications that need dynamic policy and charging control, it may obtain additional network elements from the Application Function (AF). The PCEF monitors the traffic events, according to the PCC rules for each SDF.
- 2) **SDN application:** The SDN application can be deployed in the same machine where the P-GW is running on.



Figure 3: Flow chart of QoS-based flow control

Otherwise, it can be implemented in other machines and communicate with the P-GW via the Gx interface that is used for charging and policy control between P-GW and PCRF (Figure 2). The SDN application is designed to share the traffic information related to QoS enforcement with the SDN controllers using existing northbound APIs such as Representational state transfer (REST). To dynamically control traffic flows in SDN and LTE, we design a simple decision tree, as shown in Figure 3.

B. Possible Scenarios

In this section, we elaborate possible scenarios for our proposal, which represent the advantages of providing end-toend QoS services in LTE over WAN-SDN enabled networks.

Scenario 1) Application-awareness flow control from P-GW: Client A is playing an online game, and Client B is watching an HTTP-based adaptive bitrate video streaming. P-GW enables sharing the application-awareness information with the connected SDN controllers. Based on the QoS parameters (e.g., QCI) in LTE, the controllers assign the low-latency paths among WAN routers for Client A. For Client B, the controllers enable obtaining additional video streaming information (e.g., video bitrate requested from Client B) from P-GW using DPI in LTE. It can be used to monitor the downloading datarate of the flow. If a connected WAN routing path does not meet the necessary datarate required to play the video without viewing interruption, the controllers automatically change it to another alternative path with better networking conditions.

Scenario 2) User-device and dynamic network conditionaware video delivery node selection: Clients A and B are connected to the same eNodeB. Client A, using a smartphone with low resolution, downloads a video from the video delivery node 1, while Client B, using a tablet with high resolution,



Figure 4: Testbed architecture

receives the same video from the video delivery node 2 (Figure 1). Considering the available bandwidth on the paths at the moment, the video delivery node 1 can provide enough bandwidth for Client A but not for Client B. P-GW is able to assist the video application to find the best available video delivery node based on the real-time network conditions retrieved via the SDN controller, and hardware specifications of the client's device obtained using DPI.

Scenario 3) User-profile based traffic priority: Clients who use a wireless priority service (WPS) or pay more for high speed Internet services can get better end-to-end services. The traffic generated by the users can be prioritized among WAN routers in the Internet by the SDN controllers, in addition to the per-traffic priority management specific to LTE.

Scenario 4) Flexible resource allocation at eNodeB based on dynamic QoS parameters: The current standardized QoS platform provides the static QCI table, where nine QCI values are used for traffic classification based on application types. Since P-GW is able to visualize the overall network resources via the SDN controller, the QCI characteristics can be more flexible in terms of packet delay budget and packet error loss rate. For instance, suppose that Clients A and B watch the same video through the same eNodeB. The video delivery node 1 provides a reliable connection while the video delivery node 2 offers a less-reliable connection to P-GW. P-GW has Client A connected to the video delivery node 1 with QCI7 (based on the static QCI table, 300 ms packet delay and 10^{-6} loss rate) while Client B accessed to the video delivery node 2 with different QCI properties (200 ms packet delay and 10^{-4} loss rate) to compensate for the less reliable connection to the delivery node 2. Based on the differentiated QoS parameters, eNodeB is expected to efficiently allocate the resources, which results in improving user satisfaction. This will also improve the load balancing in the server side networks.

V. BUILDING A PROTOTYPE

To prototype an SDN, we use Mininet [2], [13] that allows users to emulate complete network hosts, links and switches on a single machine. The platform created in Mininet can be easily transferred to the real OpenFlow hardware with minimal changes. In the Mininet network, we generate different types of packet flows that simulate the real network traffic such as video streaming, emails and general web traffic.



Figure 5: Testbed setups for evaluation

A. Implementation

Figure 4 shows the system architecture in our testbed. We implement a Mininet network on a Linux machine, and attach a physical networking interface to the virtual network. The connected interface (*eth1*) is wired to the 802.11g wireless access point (Cisco WRT54G). This system allows the real Wi-Fi clients to access any hosts in the Mininet network. Our SDN application in Mininet is designed to periodically update network statistics to the monitoring server. Then, the monitoring system provides the visualization of the traffic in the network over the Internet in real time. The detailed testbed setups are as follows:

- Creating WANs in Mininet: We implement six different subnet networks where virtual routers (R1 through R5) and switches (S1 through S6) are used to forward IP network traffic among the hosts in the network.
- Emulating network conditions on links: We implement seventeen links (L1 through L17) in the virtual network. We assume that each WAN link has 100 Mb/s bandwidth capacity. For a realistic setup, we emulate background traffic on the links based on recent mobile traffic statistics [14].
- Attaching a real 802.11g network to Mininet: We add an 802.11g network to the virtual network. As shown in Figure 4, the Wi-Fi client attached to the access point enables communicating with the clients and the servers in the Mininet network.
- Deploying an SDN controller over the network: Our QoS enabled application runs on the top of the SDN controller (*C0*). Using OpenFlow, it has been designed to communicate with the open virtual switches so as to obtain network statistics and control routing paths based on the QoS policy rules. It periodically updates the network information to our monitoring server.
- Building a flow database: We assume that the edge devices of last mile networks (e.g., P-GW in LTE) dynamically access and update the flow database in the WAN SDN controllers. The database contains traffic information for each traffic flow. In our testbed, for example, the database includes client identification, application types (e.g., video, voice chat and emails), priority, and required



Figure 6: TCP throughput on Client A without our SDN solution



Figure 7: TCP throughput on Client B with our SDN solution

network conditions in terms of bandwidth and latency. The SDN controller refers to the database for dynamic flow control, based on the provisioned network conditions in the network.

• Implementing a monitoring server: Our monitoring system is designed to provide network status in real time to clients over Internet. Our demo and experimental results can be found on our web page². It shows the dynamic flow change in the proposed video streaming scenarios.

B. Evaluation

For evaluation, we conducted video streaming experiments. In the network, two Wi-Fi clients connected to the same AP via the 802.11g network, and downloaded high-definition (HD - 1080p) video using HTML5 from the same video delivery node. Client A received the video packets via static path while Client B downloaded the same video but via dynamically changing routing paths, depending on network conditions in the network. During experiments, traffic congestion is created along the static path, and the SDN controller has the ability to determine better alternative routing paths. Figures 6 and 7 show the TCP throughput, and Figure 8 represents accumulated TCP packet sequence numbers measured on the clients' devices during the experiments. Client B with our OoS-enabled solution took 132 seconds to download the video (with 128 seconds of the playback) while Client A without our solution took 160 seconds to download the same video clip.

VI. CONCLUSIONS

We have shown the feasibility of an intelligent content delivery system over wireless using SDN. It relies on a novel application which unifies the control plane end-toend, across the wireless access (e.g., LTE/Wi-Fi) and mobile core network, to dynamically control traffic flows based on changing network conditions and service types in real time. We demonstrated the feasibility of the idea by building a prototype



Figure 8: Accumulated TCP packet sequence number on Client A and B as time elapsed

testbed, which emulates multiple WAN nodes and adds real Wi-Fi clients in the Mininet's virtual network. According to the experimental results, our SDN-based QoS solution enables achieving dynamic flow control based on changing network conditions and application types, which leads to improving end-user's perceived service quality.

REFERENCES

- [1] OpenFlow Switch Specification Version 1.4.0. [Online]. Available: https://www.opennetworking.org/
- [2] Mininet: An Instant Virtual Network on your Laptop. [Online]. Available: http://mininet.org/
- [3] Z. A. Qazi, J. Lee, T. Jin, G. Bellala, M. Arndt, and G. Noubir, "Application-awareness in SDN," in *Proceedings of the ACM SIG-COMM Conference*, Hong Kong, China, Aug. 2013.
- [4] M. Jarschel, F. Wamser, T. Höhn, T. Zinner, and P. Tran-Gia, "SDNbased Application-Aware Networking on the Example of YouTube Video Streaming," in 2nd European Workshop on Software Defined Networks (EWSDN), Berlin, Germany, Oct. 2013.
- [5] G. Finnie, "The Role of DPI in an SDN World," QOSMOS, Tech. Rep., Dec. 2012.
- [6] A. R. Sharafat, S. Das, G. Parulkar, and N. McKeown, "MPLS-TE and MPLS VPNS with Openflow," in *Proceedings of the ACM SIGCOMM Conference*, Toronto, Ontario, Canada, Aug. 2011.
- [7] About NOX. [Online]. Available: http://www.noxrepo.org/
- [8] S. Das, Y. Yiakoumis, G. Parulkar, N. McKeown, P. Singh, D. Getachew, and P. Desai, "Application-aware Aggregation and Traffic Engineering in a Converged Packet-Circuit Network," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC) and the National Fiber Optic Engineers Conference*, Los Angeles, USA, Mar. 2011.
- [9] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, and A. Vahdat, "B4: Experience with a Globally-Deployed Software Defined WAN," in *Proceedings of the ACM SIGCOMM Conference on SIGCOMM*, Hong Kong, China, Aug. 2013.
- [10] L. Li, Z. Mao, and J. Rexford, "Toward Software-Defined Cellular Networks," in *Software Defined Networking (EWSDN)*, 2012 European Workshop on, Darmstadt, Germany, Oct 2012.
- [11] Alcatel-Lucent, "Network Functions Virtualization Challenges and Solutions," Alcatel-Lucent, Tech. Rep., Jun. 2013.
- [12] S. Ben Hadj Said, M. Sama, K. Guillouard, L. Suciu, G. Simon, X. Lagrange, and J.-M. Bonnin, "New control plane in 3GPP LTE/EPC architecture for on-demand connectivity service," in *Cloud Networking* (*CloudNet*), 2013 IEEE 2nd International Conference on, San Francisco, USA, Nov 2013.
- [13] B. Lantz, B. Heller, and N. McKeown, "A Network in a Laptop: Rapid Prototyping for Software-defined Networks," in *Proceedings of the 9th* ACM SIGCOMM Workshop on Hot Topics in Networks, Monterey, California, Oct 2010.
- [14] Citrix, "Bytemobile Mobile Analytics Report," Citrix Systems, Tech. Rep., Jun. 2013.

²http://dyswis.cs.columbia.edu/belllabs/experiments/