Cross-layer Cooperation for Better Network Service

Mikhail Liubogoshchev
Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
liubogoshchev@wireless.iitp.ru

Abstract—**In modern networks, different applications struggle to generate traffic in such a way that maximizes the QoE of their users. For that, they try to estimate and predict the network state. However, these estimations are far from being accurate because of the fluctuations imposed by the way the networks operate. In the paper, we discuss why the estimations of the applications are so inaccurate and propose the applications to communicate with the networks to improve QoE for the end-users and the efficiency of the network operation.**

For a long time, the networks provided mainly the best-effort service [1]. As a consequence, the applications learned to adapt to QoS that fluctuates in a very wide range. For example, DASH video clients expect to watch high-quality videos without frequent quality changes, and playback interruptions [2]. Therefore, the applications try to satisfy these requirements and adapt to the changes in the network conditions. For that, they estimate the network state by measuring how its data is transmitted. The basic examples of the metrics estimating the network state are the bandwidth and delay (or round trip time). Typically, to measure bandwidth, the client would send a large portion of data at once and measure how fast it is transmitted. To measure delay (or round-trip time), the client would send small packets at a low rate to avoid increasing the network load.

These methods are OK, for example, for buffered video streaming of DASH applications, whose buffer allows to average the network measurements over rather long timescales, e.g., tens-seconds-average network throughput. However, for many emerging and future applications like cloud gaming or AR/VR, even additional 10 ms latency of a single packet can cause severe QoE degradation [3]. It means that they need to estimate the state of the network at very short timescales and quickly adapt to its fluctuations.

The fluctuations are especially severe in wireless networks, which are typically the bottlenecks at the traffic delivery route and significantly impact the end-to-end measurements. To increase the capacity of the wireless links, the overheads related to the channel access procedures need to be reduced. In both cellular and Wi-Fi networks, this is achieved with aggregation: more data is sent less frequently. For example, in cellular networks, transmissions have a relatively long period. In LTE, it is 1 ms and, in current 5G deployments, it is 0.5 ms, and in Wi-Fi, a typical duration of a transmission is a few milliseconds.

With a rather high probability, some of the aggregated packets will have to be retransmitted. To illustrate, the MCS (Modulation and Coding Scheme) in modern wireless networks is selected to provide packet error rate around 10%. A single transmission often involves a few dozens of packets. Therefore, the probability that at least one packet in transmission will require retransmission is close to 100%. Modern TCP-friendly wireless Media Access Control (MAC) layers try to ensure in-order delivery of packets in the flows. Therefore, if all the packets in the transmission belong to the same flow, some of them will be stalled by the MAC layer until the lost packets are successfully retransmitted. The retransmissions can happen only in several milliseconds (e.g., 8 ms for LTE or a random channel access procedure duration in Wi-Fi), which means that the data of a single flow is very often forwarded to upper layers aggregated in a window of several milliseconds. This aggregation and protocol-induced delays further increase the fluctuations of the network state and cause almost delusive measurements: they almost do not depend on the network load and because of aggregation they can happen relatively rarely. Moreover, they are hard to take into account because of the complicated and unknown to the end-nodes logic of the wireless transmissions control and frequently changing channel state.

An appealing option to solve these measurement problems is to establish the communication between the applications running on the end-nodes and the intermediate network equipment. This way, the applications will learn the important characteristics of the network explicitly and not with some guesses and estimations. Furthermore, this communication can provide the network equipment with accurate data of the application-layer QoE metrics and with the instantaneous QoS requirements of the applications. With such information, the network can use a wide range of techniques developed over the past years to utilize the network resources more efficiently while providing high QoE for the end-users [4].

If carefully designed, this communication can give the clients even higher privacy than they get in modern networks. Now Deep Packet Inspection (DPI) software or hardware is common in the networks of the Internet Service Providers (ISPs) [5]. Simultaneously with detecting malicious traffic and classifying the flows for improved QoS, the operators gain access to various data on the web services used by their clients. However, if the application-to-network communication protocol was employed, it could provide the metadata enabling the operators to improve QoS, but not revealing the contents of the traffic, and the visible now data (e.g., domain names in DNS requests) could be encrypted to provide higher privacy to the users.

By now, numerous application-to-network communication protocols or APIs have been developed. The examples are SAND [6], xStream [7], CAPIF [8] and xMB [9]. However, almost no support of these protocols is present in both modern networks and applications.

ACKNOWLEDGMENT

The research has been carried out at IITP RAS and supported by the Russian Science Foundation (Grant No 21-79-10431, https://rscf.ru/en/project/21-79-10431/)

REFERENCES

[1] J. Farkas, L. L. Bello, and C. Gunther, “Time-sensitive networking standards,” *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 20–21, 2018.

[2] N. Barman and M. G. Martini, “QoE modeling for HTTP adaptive video streaming–a survey and open challenges,” *IEEE Access*, vol. 7, pp. 30 831–30 859, 2019.

[3] GSMA, “Cloud ar/vr whitepaper,” Tech. Rep., April 2019.

[4] R. Trestian, I.-S. Comsa, and M. F. Tuysuz, “Seamless multimedia delivery within a heterogeneous wireless networks environment: Are we there yet?” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 945–977, 2018.

[5] (2019, Sept) The global internet phenomena report. Sandvine. [Online]. Available: https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/Internet%20Phenomena/Internet%20Phenomena%20Report%20Q32019%2020190910.pdf

[6] ISO, *ISO/IEC 23009-5:2017 Information technology – Dynamic adaptive streaming over HTTP(DASH) – Part 5: Server and network assisted DASH (SAND)*. ISO.

[7] I. F. Akyildiz, E. Khorov, A. Kiryanov, D. Kovkov, A. Krasilov, M. Liubogoshchev, D. Shmelkin, and S. Tang, “xstream: A new platform enabling communication between applications and the 5g network,” in *2018 IEEE Globecom Workshops (GC Wkshps)*, Dec 2018, pp. 1–6.

[8] 3GPP, *3GPP TS 21.222:Common API Framework for 3GPP Northbound APIs (Release 15)*. 3GPP, 2021.

[9] ——, *3GPP TS 26.348 Northbound Application Programming Interface (API) for Multimedia Broadcast/Multicast Service (MBMS) at the xMB reference point; (Release 16)*. 3GPP, 2020.