Xiufeng Xie and Xinyu Zhang *University of Wisconson-Madison*
Swarun Kumar *Carnegie Mellon University*
Li Erran Li *Uber Technologies*

Editors: Robin Kravets and Nic Lane

piStream:

Physical Layer Informed Adaptive Video Streaming Over LTE

Excerpted from "piStream: Physical Layer Informed Adaptive Video Streaming over LTE," from *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. http://dl.acm.org/citation.cfm?id=2790118

Mobile video streaming has witnessed a surge over the past few years, accounting for 70% of the mobile Internet traffic, with a compound annual growth rate of 78% [1]. The LTE cellular services have been massively deployed to match such growing traffic demand, with a peak downlink bit-rate of 300Mbps, almost 10× over 3G. However, user-perceived quality of experience (QoE) remains unsatisfactory. A recent worldwide survey [2] reveals that, even in regions with wide LTE coverage, LTE only increases video quality (bit-rate) by 20% over 3G. On the other hand, the average stalling time remains about 10 seconds for each minute's mobile video playback.

We have conducted extensive measurement studies and identified the root cause of the poor user experience – the inability of the streaming applications to track the network bandwidth changes under LTE link dynamics. Our investigation focuses on the popular HTTP-based adaptive streaming protocols (DASH) [3] used by many mainstream video content providers, such as YouTube and Netflix. DASH adapts the video bit-rate in real time to fit the varying network bandwidth. Specifically, a DASH server splits a video into a sequence of segments with uniform duration, and each segment is encoded into multiple copies with different encoding bit-rates. During streaming, a DASH client requests the quality or bit-rate on a per-segment basis, based on its historical downloading throughput. However, the throughput value may largely underestimate the network bandwidth unless the video segment saturates the end-to-end network path. On the other hand, as the LTE network condition fluctuates, history-based estimation may also lead to bandwidth overestimation and video stalling. As a result, the DASH clients suffer from the worst impact of either underestimation or overestimation of the available network bandwidth.

The above challenges arise from the inherent limitations of current DASH architectures, which can only obtain limited and lagged network measurements from a narrow network-layer scope. There has been a wide variety of DASH video adaptation algorithms that adopt throughput-based [7], buffer-based [8] and probing-based [14] bandwidth exploration strategies (to be detailed in Table 1). However, these adaptation algorithms still treat the end-to-end network path as a black box, so they fall short of agility and precision in the presence of high bandwidth variation in LTE.

In order to break such limitations, we propose piStream, a cross-layer video streaming framework compatible with current LTE infrastructures and MPEG-DASH standard [3]. piStream harnesses the information from LTE PHY-layer to obtain a more comprehensive view of the instantaneous network status, and hence enable more agile video adaptation than conventional protocols. The piStream design leverages the well-organized LTE radio resource structure, which divides the physical channel into a pre-defined number of Physical Resource Blocks (i.e., PRBs, the smallest resource unit in LTE's physical layer). piStream tracks the available network bandwidth based on the cell-wide PRB utilization status. Since the end-to-end link is typically bottlenecked by the LTE access link rather than the backbone network [4], the LTE link's PRB utilization alone can be used to approximate how much end-to-end bandwidth remains. In this way, piStream overcomes a fundamental limitation of conventional DASH protocols that are slow at exploring unused bandwidth.

We validate our piStream design by tethering a software radio that implements the radio resource monitor and PHY-informed rate scaling, to an LTE smartphone that implements the application layer video adaptation scheme. This piStream client prototype can directly play video in real-time from any server that follows the industry MPEG-DASH standard [4]. We benchmark piStream's performance against four state-of-the-art DASH schemes that have demonstrated superior performance over commercial DASH players. Under various experimental settings, piStream outperforms all other DASH schemes by achieving higher video quality and lower/comparable video stalling rate. Under typical static indoor environments, piStream achieves around 1.6× video quality gain over the runner-up while maintaining close to 0% stalling rate.

To our knowledge, piStream is the first protocol to facilitate LTE adaptive video streaming using PHY-informed network bandwidth estimation. The contributions of piStream can be summarized as follows: (i) a lightweight PHY-layer radio resource monitor and rate scaling mechanism that enables an LTE client to efficiently estimate available bandwidth; (ii) a video adaptation algorithm that harnesses the PHY-informed bandwidth estimation to optimize the video quality and stalling rate; (iii) a real-time piStream prototype implementation that demonstrates significant performance gains over state-of-the-art DASH algorithms.

SYSTEM DESIGN

Current adaptive video streaming systems face several fundamental challenges in LTE. First, network-layer measurements cannot accurately reflect the available network bandwidth unless the bandwidth is saturated, so conventional DASH protocols that adapt based on historical throughput measurement are slow at bandwidth exploration. Furthermore, DASH client needs to adapt the bit-rate for the next (future) video segment based on currently available bandwidth, but the LTE link fluctuations can easily invalidate the rate adaption decision. To address these challenges, a piStream client first uses a **radio resource monitor** to estimate the PHY-layer resource (PRB) utilization by sensing the LTE downlink channel. It then performs **PHY-informed rate scaling**, i.e., scaling up its instantaneous throughput measurement proportionally by the PRB utilization, in order to obtain an estimation of the potential network bandwidth available for video streaming. To accommodate possible bandwidth fluctuations [10,11], piStream adopts a stochastic adaptation protocol that estimates how likely current bandwidth is going to last for the next video segment, based on the long-range-dependency (LRD) feature of the traffic patterns in typical packet switching networks [12, 13]. Accordingly, the client selects a bit-rate to maximize the video quality while minimizing the risk of stalling. We refer to this protocol as **LRD-based video adaptation**. Below we provide more details on piStream's design components:

1. Radio Resource Monitor

The radio resource monitor acts as a PHY-layer daemon in each LTE client that monitors the utilization status of downlink radio resources (PRBs) in the current cell. The PRB utilization essentially reflects the ratio between current downlink traffic rate and the potential network bandwidth. Note that an LTE basestation performs centralized resource allocation and each client is only informed of the resources assigned to itself rather than the overall cell-wide resource utilization. A straightforward way for an LTE client to obtain the cell-wide PRB utilization is to decode the downlink control messages dedicated to all other clients [5]. However, this approach is not compliant with the LTE standard and will fail LTE's bit-error control mechanism. In contrast, piStream requires a resource monitor that is accurate, efficient, and deployable on LTE phones. We meet this challenge by leveraging the well-organized LTE resource structure. Our radio resource monitor inspects the signal energy on each PRB to assess whether it is idle or utilized for data transmission. The utilized PRBs carry higher energy than the noise level. We then obtain the PRB utilization as the fraction of the utilized PRBs among all available PRBs in the LTE resource structure.

Although such a resource monitor is not yet available in current LTE phones, it can be realized without any hardware modification. In effect, it only requires the phone to expose its channel state information (CSI), which only needs a firmware upgrade. Furthermore, recent work [6] has shown that the PRB utilization can be estimated, though less accurately, from energy-based PHY layer metrics readily accessible on current smartphones.

2. PHY-Informed Rate Scaling

Given the monitored PRB utilization, a piStream client then scales up its throughput measurement proportionally by this PRB utilization to obtain a network bandwidth estimation. Behind this design are two rationales: (i) The end-to-end network bandwidth is typically bottlenecked by the LTE access link [3]; (ii) The access link capacity of a client is determined by the amount of allocated PRB resources since more allocated PRBs obviously support higher traffic bit-rate. As a result, the cell-wide PRB utilization essentially indicates the amount of surplus resources remaining at the LTE basestation, which can be potentially consumed by the piStream client to enable higher traffic bit-rate. Based on these insights, we derive the following principle for a client to scale up its measured throughput to obtain an estimation of its potential network bandwidth: Suppose current cell-wide PRB utilization ratio is u, and the client's current throughput is R, then at least a traffic bit-rate of R/u can be supported without over-utilizing the network bandwidth.

3. LRD-based Video Adaptation

After the rate-scaling algorithm estimates the potential bandwidth currently available, a piStream client can correspondingly select

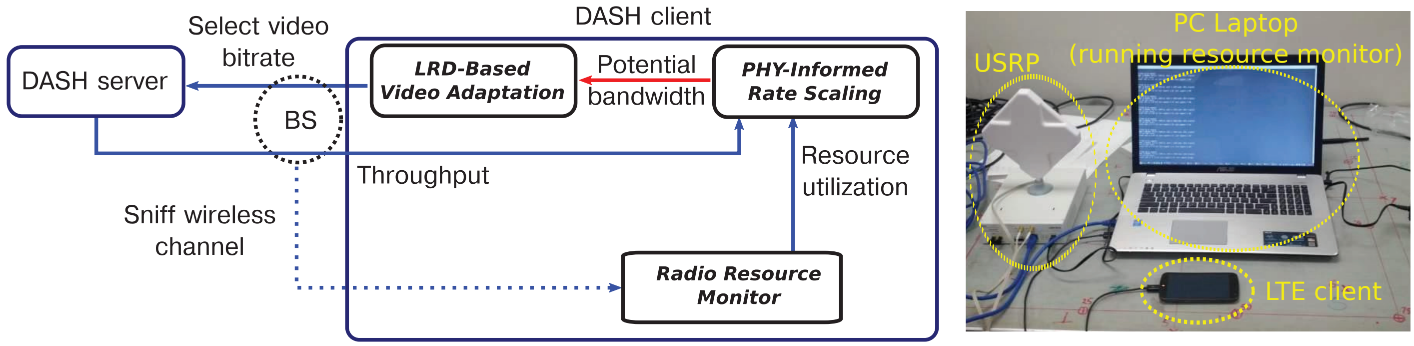

FIGURE 1. (a) piStream Architecture Implementation. (b) piStream Testbed.

the bit-rate for the next video segment. Theoretical work [11] has shown that DASH performance can improve significantly if future network bandwidth is known. However, the wireless channel fluctuation and random traffic demand makes bandwidth prediction very challenging in LTE. Instead of attempting to predict the exact value of future bandwidth, the piStream client takes advantage of a statistical feature, i.e., long-range-dependency (LRD), of LTE downlink traffic. It estimates how likely its current bandwidth-level is likely to remain consistent for the next video segment, and then chooses the per-segment video bit-rate in a probabilistic way. In addition, to alleviate the impact of an occasional improper decision, piStream also takes the client's playback buffer level into consideration. When the amount of video bits in the playback buffer is low, an improper video bit-rate choice can easily lead to an empty buffer and freeze the video, thus the client acts more conservatively in this case by reducing the video bit-rate for every bandwidth decrease it experiences.

TABLE 1. Implemented DASH algorithms.

Algorithm name	BBA	FESTIVE	PANDA	GPAC	piStream
Bandwidth estimation	N/A	Average throughput	Incremental probe	Latest throughput	PHY-informed
Buffer knowledge	Yes	No	No	No	Yes
PHY information	No	No	No	No	Yes

PROTOTYPE

We implement a prototype of piStream that works in real time. Figure 1(a) and (b) illustrate the implemented system architecture and the prototype testbed, respectively.

Radio Resource Monitor: We implement piStream's radio resource monitor over the USRP N210 software radio. It processes the LTE downlink signal samples captured by the software radio, and extracts the cell-wide PRB utilization in real time. The PRB utilization statistics are computed over a moving-average window whose length equals the video segment duration (2000ms in our experiments), and then the statistics are fed to the video adaptation logic to facilitate the bit-rate selection for the next video segment.

DASH Client: We implement piStream's PHY-informed rate scaling and LRD-based video adaptation algorithms by extending the DASH player from the GPAC project [8]. To interface the application layer with the radio resource monitor, we connect the LTE smartphone running the DASH player and the USRP software radio monitoring the PHY channel to the same PC laptop, and then set up a named pipe between the resource monitor and the DASH player.

DASH Server: We set up a standard MPEG-DASH server on Amazon EC2 using Apache2. The bandwidth to the server is 66 Mbps measured with iperf over a wireline network, which is much higher than the maximum 15Mbps LTE downlink throughput in our measurements. This follows typical scenarios where the LTE downlink, instead of the wireline backbone, acts as the bottleneck for video streaming. The server hosts multiple DASH video data sets with typical 2s video segment length.

Benchmark Protocols: To benchmark piStream's performance, we implement state-of-the-art DASH adaptation protocols including buffer-based adaptation (BBA [8]), optimization-based adaptation using historical throughput (FESTIVE [9]), and TCP-like bandwidth probing (PANDA [14]). The implementation builds on GPAC's client-centric DASH player [7], but substitutes GPAC's default video rate adaptation algorithm. Table 1 summarizes the features of these algorithms.

EVALUATION

With our testbed implementation, we perform a thorough system-level test under various network conditions. Due to the space limit, we only present the experiment results from a static video client in populated commercial LTE network. More evaluation settings and results can be found in [15]. From the results in Fig. 2, we have the following observations:

(i) piStream achieves the highest video quality among all algorithms. In particular, it improves the average video bit-rate (quality) by 64% over GPAC, 125% over FESTIVE, 61% over BBA, and comparable to PANDA. This performance boost is mainly attributed to the PHY-informed rate scaling,

which enables piStream to select higher video bit-rate than other algorithms without overusing the bandwidth. Meanwhile, piStream achieves a low video stall rate of close to 0 thanks to the LRD-based video adaptation, which effectively evaluates the probability of bandwidth drop and thus minimize the risk of buffer drain even under piStream's aggressive bit-rate selection.
(ii) The GPAC client uses instantaneous throughput to determine the video quality of the next video segment, which may severely underestimate the available bandwidth. Therefore it achieves a much lower video quality than piStream which has a similar level of video stalling rate.
(iii) The FESTIVE algorithm relies on the harmonic mean of the historical throughput to control the video bit-rate. Since the harmonic mean value updates slowly even under the high bandwidth variation of LTE, it results in severe bandwidth underutilization, despite a comparable stalling rate with piStream.
(iv) Given a high buffer level, the BBA algorithm selects a high video bit-rate regardless of current throughput. Such uninformed aggressive choice causes bandwidth overuse, network congestion, and finally a higher video stall rate (23%) than piStream.
(v) The PANDA algorithm keeps requesting higher video bit-rate to explore the network bandwidth until observing a throughput decrease as an congestion indicator. However, the client can detect the decreased throughput only after the congestion, and the video adaptation via feedback takes effect even later, thus in most cases the video freezes before the video server reduces the bit-rate.

CONCLUSION

Video streaming applications keep growing and dominating the cellular network traffic today. However, current video streaming protocols fall short of QoE and leaves the LTE bandwidth highly underutilized. Inspired by the rich information contained in the PHY-layer statistics, we present piStream, a DASH-compatible adaptive video streaming framework that exploits LTE's PHY layer information to facilitate video rate adaptation. piStream's PHY-informed design enables a more accurate bandwidth estimation and agile video rate adaptation. Extensive experiments on a

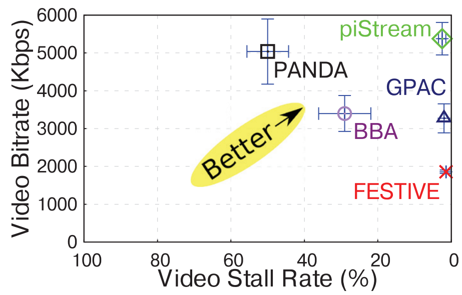

FIGURE 2. System-level Video Performance Comparison

real-time prototype show that piStream outperforms state-of-the-art video streaming algorithms in both video quality and playback smoothness. Under typical streaming environments, piStream achieves around 1.6× video quality (bit-rate) gain over the runner-up algorithm while maintaining a lower video stalling rate. Since future cellular network architecture is evolving to expose more detailed PHY-layer statistics to the user space, we believe piStream's PHY-informed design principle can be applicable to a wider range of mobile network services in the long term. ■

Xiufeng Xie is a PhD student in the ECE department at the University of Wisconsin-Madison. His research interests include mobile computing systems, MIMO systems, and cellular networks.

Xinyu Zhang is an assistant professor in the ECE department at the University of Wisconsin-Madison. His research interests include wireless networking and mobile computing. He received a PhD in EECS from the University of Michigan-Ann Arbor in 2012.

Swarun Kumar is an assistant professor of the ECE department at Carnegie Mellon University. His research interests are in next-generation wireless network protocols and services. He received a PhD in EECS from MIT in 2015.

Li Erran Li is a researcher leading the Machine Learning Platform development in Uber Technologies. He is also an adjunct professor of the Computer Science department at the Columbia University. His research interests include machine learning systems and mobile computing. He received a PhD in Computer Science from Cornell University in 2001.

REFERENCES

[1] Cisco Systems, Inc., "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017," 2012. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html
[2] Ctrix, "Mobile Analytics Report," 2014. [Online]. Available: http://www.citrix.com/content/dam/citrix/en us/documents/products-solutions/citrix-mobile-analytics-report-september-2014.pdf
[3] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," in IEEE Multimedia, 2011.
[4] Y. Xu, Z. Wang, W. Leong, and B. Leong, "An End-to-End Measurement Study of Modern Cellular Data Networks," in Proc. of Passive and Active Measurement Conference, 2014.
[5] S. Kumar, E. Hamed, D. Katabi, and L. Erran Li, "LTE Radio Analytics Made Easy and Accessible," in Proc. Of ACM SIGCOMM, 2014.
[6] Chakraborty, Abhijnan, et al. "Coordinating Cellular Background Transfers using Loadsense." ACM MobiCom, 2013.
[7] ENST, "GPAC." [Online]. Available: https://gpac.wp.mines-telecom.fr
[8] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, andM. Watson, "A Buffer-based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service," in Proc. of ACM SIGCOMM, 2014.
[9] J. Jiang, V. Sekar, and H. Zhang, "Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE," in Proc. of ACM CoNEXT, 2012.
[10] X. Yin, V. Sekar, and B. Sinopoli, "Toward a Principled Framework to Design Dynamic Adaptive Streaming Algorithms Over HTTP," in Proc. of ACM HotNets, 2014.
[11] X. K. Zou, J. Erman, V. Gopalakrishnan, E. Halepovic, R. Jana, X. Jin, J. Rexford, and R. K. Sinha, "Can Accurate Predictions Improve Video Streaming in Cellular Networks?" in Proc. of ACM HotMobile, 2015.
[12] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic," IEEE/ACM Transactions on Networking, vol. 2, no. 1, 1994.
[13] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web Traffic: Evidence and Possible Causes," in Proc. of ACM SIGMETRICS, ser. SIGMETRICS '96, 1996.
[14] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. Begen, and D. Oran, "Probe and Adapt: Rate Adaptation for HTTP Video Streaming At Scale," IEEE Journal on Selected Areas in Communications, vol. 32, no. 4, 2014.
[15] Xie, Xiufeng, Xinyu Zhang, Swarun Kumar, and Li Erran Li. "piStream: Physical Layer Informed Adaptive Video Streaming over LTE." ACM MobiCom, 2015.