Characterization and Modelling of YouTube Traffic in Mobile Networks

Géza Horváth, Péter Fazekas
Department of Networked Systems and Services
Budapest University of Technology and Economics
Budapest, Hungary
gezah@hit.bme.hu, fazekasp@hit.bme.hu

***Abstract*—Video streaming is one of the most data-intensive applications of today's Mobile Internet and YouTube generates 20% of mobile networks downstream traffic in several regions. YouTube employs the progressive download technique for video playback and therefore its traffic is bursty. We present the characteristics of the most important burst measures of traffic generated by YouTube when accessed via mobile broadband connections. Moreover, we also distinguish the characterization for non-optimized and optimized traffic since mobile operators are using media optimization platforms to effectively deliver video content. In this paper, we present our measurement-based analytical results to derive characterization of the traffic sources. As a result, we propose a generic model and its parameters according to the optimized and non-optimized traffic sources based on the experimental evaluation of the captured YouTube traffic. The proposed traffic models can be used in simulation of future work.**

Keywords-YouTube; video optimization; burstiness; traffic characterization; traffic model.

I. Introduction

Mobile data traffic is set explode in the upcoming years as consumers add more devices to the mobile networks and operators deploy faster networks. Mobile operators around the world are ramping up deployment of 4G LTE networks while subscribers are consuming more and more data. One of today's most data-intensive applications is video streaming. In this work, we present the characterization and examine the bursty nature of the mobile network traffic generated by one of the most relevant video streaming platform: YouTube.

According to the Sandvine Global Internet Phenomena Report from 2014 in most regions, YouTube is the application responsible for generating the most bandwidth; it accounts for around 20 % of mobile downstream traffic in North America, Europe and Latin America. While Netflix saw growth in the share thanks to the continued rollout of high bitrate Super HD content, in many regions YouTube continues to be the largest single source of Real-Time Entertainment traffic on both fixed and mobile access networks, which makes it the leading source of Internet traffic in the entire world. In North America 17.61% of the total downstream traffic is generated via YouTube, while in Europe it is 19.27% [1].

YouTube employs the progressive download technique; its video content is delivered by a regular HTTP web server rather than a streaming server. Video delivered using this technique is stored on the viewer's hard drive as it's received, and then it's played from the hard drive; it enables video playback before the content download is completely finished [6]. It also uses the HTTP/TCP (Hypertext Transfer Protocol/Transmission Control Protocol) platform to deliver data, which further distinguishes it from traditional media streaming [4].

The present paper aims at four main objectives: (1) to highlight the YouTube service traffic characterization when accessed via mobile broadband connection; (2) to present proper measures for catching the bursty nature of YouTube traffic; (3) to distinguish the effect of media optimization platform used in mobile networks; and (4) to propose traffic models for non-optimized and optimized YouTube traffic accessed via mobile broadband connection.

The rest of the article is organized as follows: Section 2 provides an overview of the traffic sources and the experimental framework used during the analysis and the capture method itself; Section 3 provides a summary of the most important burst and correlation measures and their numerical evaluation on the captured traces; Section 4 presents the analysis of the experimental results of the main characteristics of YouTube traffic accessed via 3G mobile network; Section 5 provides the traffic model state machine, its parameters and its mode of operation via algorithm; finally, Section 6 presents the main conclusions.

II. Traffic Sources

A. *Experimental framework*

In this section, we describe the experimental framework used to collect traces of data traffic generated by YouTube accessed via 3G mobile network. The framework is composed of a notebook connected to a mobile service provider's network via USB stick on the move. The modem is able to handle downstream traffic up to 42 Mbps via DC-HSPA+ (Dual Carrier-High Speed Packet Access) and 21 Mbps via HSPA+ access. Its upstream capability is 5.76 Mbps. The used mobile technology was at least HSPA+ in 98.4% of the samples, the remaining part was EDGE (Enhanced Data Rates for GSM Evolution). During pilot tests it was verified that neither the CPU (Central Processing Unit) nor the memory of the notebook impeded the normal playback of the video clips.

A playback monitor tool has been built with the main objective of analyzing YouTube traffic and collect available information about the videos. The tool includes a web application using the YouTube player API (Application

ISBN: 978-1-61208-398-8

Programming Interface) via JavaScript [11]. It is able to play videos sequentially based on *video_id* list and collect all the available information like duration, total bytes into a log file. The web application was stored on a public web server and was accessed from the notebook via Chrome browser 39.0. The other part of the framework is Microsoft Network Monitor 3.4 software installed on the notebook. However, Wireshark is more popular for trace collection, we had to swap it because it was not able to capture on mobile broadband interfaces. On the other hand the collected traces were compatible with Wireshark as well, so we could use its advantages in traffic analysis. It was enough to capture only the first 68 bytes of each package, with the help of the headers we were able to reproduce the traffic itself [9] [10] [12].

B. *Media Optimization in Mobile Networks*

Video optimization refers to a set of technologies used by mobile service providers to improve the consumers viewing experience by reducing video start times or re-buffering events. The process also aims to reduce the amount of network bandwidth consumed by video sessions. While optimization technology can be applied to videos played on a variety of media-consuming devices, the costliness of mobile streaming and increase in mobile video viewers has created a very high demand for optimization solutions among mobile service providers [8].

Video optimization techniques used to improve network performance and subscriber experience include:

- Caching — local copies of video are stored for successive access
- Pacing — data is delivered “just in time” rather than all at once (and frequently abandoned)
- Transrating — changing the frame rate
- Transcoding — recoding the video codec for lower bitrate
- Enhancing TCP — for mobile handling to minimize back and forth signaling
- Compressing — data reduction of the content

Lossless Media Optimization offers a low-cost media optimization entry-point for operators. It adjusts the video transmission rate to match the actual video play rate, without changing the video content. Instead of clogging the network at the start of the video, the lossless Just-In-Time (JIT) technique “spreads” the video delivery over the entire video play time. With JIT, videos download as fast as needed, rather than as fast as possible, taking additional advantage over the fact that most videos are not watched to the end [8].

Lossy Media Optimization reduces the amount of data transmitted over the wireless network. Operators are using static data reduction or dynamic bandwidth shaping, which are performed in combination with Just-In-Time (JIT) lossless optimization. To achieve the best response time, lossy optimization is performed "on-the-fly". The data reduction of the media is a function of the media quality. The higher the data reduction of the content, the lower is the quality of the media provided [8]. Within the measured mobile network Lossy Media Optimization is in use. Figure 1 depicts the high-level architecture of the platform.

We have two types of traffic in the mobile network. In the first scenario video clips were delivered through the aforementioned media optimization platform, in the second scenario we collected traces without that. The mobile network was configured with two different Access Point Names to differentiate between the two scenarios. The only thing we had to do is to set the APN in use on the notebook before connecting to service provider.

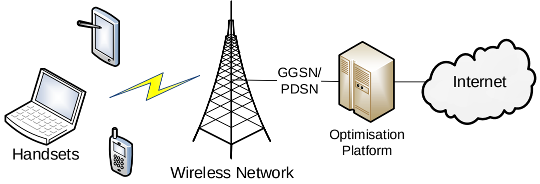

Figure 1. Architecture of Opimisation Platform

C. *Collected traces*

A trace set of 95 video clip downloads was collected with the recently described experimental framework. This set has been collected to understand the main traffic characteristics of YouTube traffic accessed via mobile network. All the video clips has been downloaded in their default format [4][5]. YouTube uses the MP4 container for high-definition (HD) clips and uses the Flash Video (FLV) as the default format for the majority of non-HD clips. YouTube adapts all the uploaded clips to the aforementioned formats before posting [4].

The trace set was captured two times as follows. We can assume that there was no bottleneck in the mobile networks backhaul or core network.

- SHAPED: This set has been collected using the recently mentioned mobile networks Lossy Media Optimization function; APN is set accordingly. Traffic was affected by the YouTube server, the Media Optimization platform and the access technology, radio conditions of the used mobile network.
- UNSHAPED: This set has been collected without using the recently mentioned mobile networks Lossy Media Optimization function; APN is set accordingly. Traffic was affected by the YouTube server and the access technology, radio conditions of the used mobile network.

III. BURST AND CORRELATION MEASURES

A simple class of burstiness measures takes only the first-order properties into account; they are each a function of the marginal distribution only of interarrival times (with resolution of 0.001 s). These measures can be taken into consideration as various characteristics of the marginal distribution of the inter-arrival time. The possible set of properties are the moments of that distribution [2].

More complex measures are utilizing the second-order properties of the traffic; they do take account of temporal dependence in traffic. From this class indices of dispersion is one of the most well-known methods. It includes the correlation properties of the traffic and can be very informative [2].

A. *Measures based on the first order properties*

One of the mostly used measures is the peak to mean ratio (PMR). Peak is defined in this paper as inter-arrival time between the two closest arrivals. In case of UNSHAPED traffic peak may be very high and likely to correspond to two arrivals in consecutive slots in practice. In the case of SHAPED traffic, the peak tells more about the shaper parameters than the traffic itself [3].

TABLE I. FIRST ORDER PROPERTIES

	PMR	***SCV***	***m_3***
UNSHAPED	0.00004349	281.555	25.265
SHAPED	0.00008560	254.040	24.411

Another widely used measure is the squared coefficient of variation (SCV) of the inter-arrival times. It includes information from the first two moments and is defined $C^2(X) = Var(X)/E^2(X)$ where X is the inter-arrival time [3]. As shown in Table I, when comparing the values of the SHAPED and UNSHAPED traffic, it is higher for the UNSHAPED indicating it is burstier because it includes sustained higher intensities. Also, the lower value of the SCV indicates a smoother process.

The SCV takes into account only the set of the inter-arrival times, the order of that is passed by. However, burstiness is mainly caused by two factors [3]: the distribution and especially the tail of the inter-arrival times and the correlation between them. The squared coefficient of variation from the first-order measures class takes into account only the inter-arrival distribution.

Suppose that *X* is a real-valued random variable. The variance of *X* is the second moment of *X*, and measures the spread of the distribution of *X*. The third and fourth moments of *X* also measure interesting features of the distribution. The third moment measures skewness, the lack of symmetry, while the fourth moment measures kurtosis, the degree to which the distribution is peaked.

Higher moments can also tell useful information about the traffic characteristics. For example, two traffic with same first two moments but different third moment can produce very different queueing behavior. The third moment tells about the long inter-arrivals. Although inter-arrival times are bounded from above for both SHAPED and UNSHAPED traffic, higher m_3 for UNSHAPED traffic means it has higher inter-arrival times.

B. *Measures based on the second order properties*

Indices of dispersion measures are useful because they show the traffic variability over different scales and they can capture the correlation structure. Two indices of dispersion measures are widely used: the index of dispersion for intervals (IDI) is related to the sequence of inter-arrivals; the index of dispersion for counts (IDC) is related to the sequence of counts of arrivals in consecutive time units [2].

The IDI is defined for a stationary inter-arrival sequence X as follows:

$$IDI = \frac{var(X_{i+1}+\cdots+X_{i+n})}{nE^2(X)} = C_j^2\left(1+2\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\rho_j\right) \quad (1)$$

$$n = 1, 2, \ldots$$

$$C_j^2 = \frac{var(X)}{E^2(X)}$$

$$\rho_n = cov(X_i, X_{i+n})/var(X)$$

In the definition, the sum of k consecutive inter-arrivals is taken. In the case of bursty process, the sort and long inter-arrivals are grouped together, and it causes and it causes the IDI to increase with increasing k. In fact, the increase or decrease in the IDI graph is directly related to the correlation of the inter-arrival sequence.

The IDC for a stationary process is defined as

$$IDC = \frac{V(t)}{E(t)} = \frac{V(t)}{tm} \quad (2)$$

where V(t) and E(t) are the variance and the expected number of the arrivals in an interval of length t, and E(t)=tm, where m is the mean intensity of arrivals. The IDC shows the variability of a process over different time-scales.

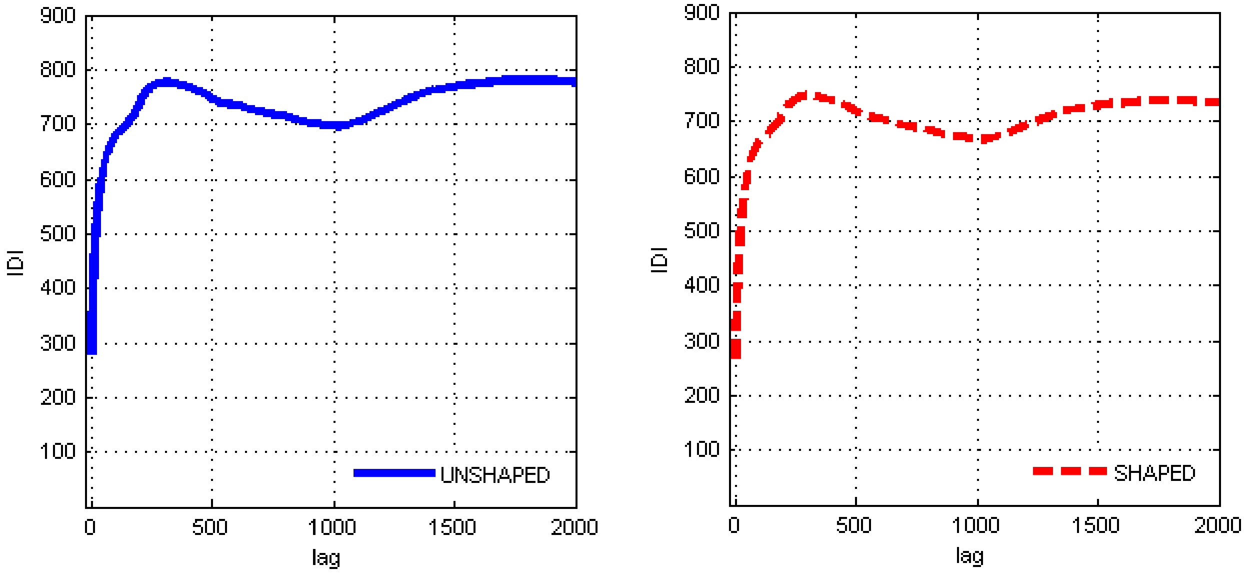

Figure 2. IDI Graph of traffic sources: (a) UNSHAPED, (b) SHAPED

ISBN: 978-1-61208-398-8

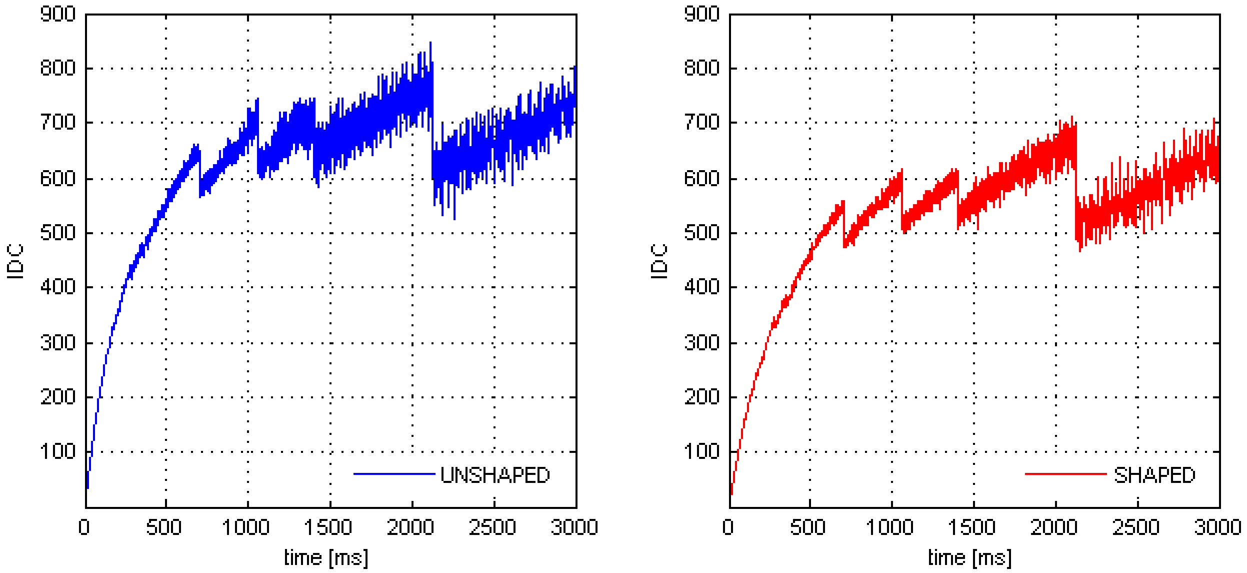

Figure 3. IDC Graph of traffic sources: (a) UNSHAPED, (b) SHAPED

As shown in Figure 2 and Figure 3, in case of UNSHAPED and SHAPED traffic the IDI and IDC curves both increase. Together they imply that the low burstiness in short scales increase over higher scales due to positive correlation. The quickly increasing curves and the high value at infinity imply that these are very bursty sources. From Figure 3, it can be seen, the SHAPED traffic has lower squared coefficient variation (start of IDI curve) and a bit lower IDI and IDC curves than UNSHAPED meaning lower burstiness of the traffic source.

To have accurate IDC curves, the maximal block size (t) should not exceed 10% of the sample size. Using non-overlapping blocks with size *t* we need at least 10 values in a block to calculate accurate variance. From Figure 3 it can be seen that increasing block size *t* implies more inaccurate IDC curves.

IV. Analytical Results

This section introduces the experimental results obtained to evaluate the traffic generated by YouTube captured via mobile broadband access. On the basis of the information provided by our experimental framework we depict the progressive download of a video clip, which belong to trace set UNSHAPED, as an example. Figure 4 plots the time evaluation of the instantaneous amount of data received by the player at the beginning of the download.

A. Initial burst

A video clip download commences with a significant burst of data, later the receiving data rate of the client's player is considerably reduced (see Figure 4 (a)). Initial burst is identified in each trace by determining the slope change in the accumulated data received by the player between the initial burst and the throttling phase. To eliminate the effect

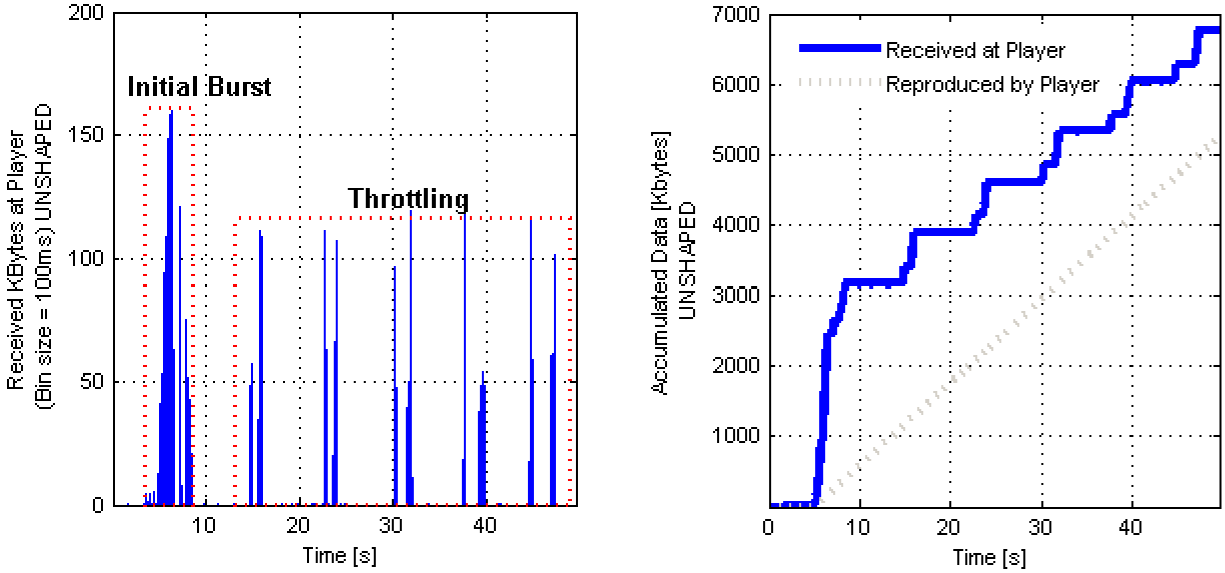

Figure 4. Examples received data at player: (a) instananeous, (b) accumulated

ISBN: 978-1-61208-398-8

of temporary slop changes caused by network bandwidth fluctuations, the accumulated data have been filtered with a 400 ms simple moving average. A slope approximation sequence computed as the difference between consecutive samples of the filtered series. Then, the maximum slop after the initial burst, during the throttling phase is computed by considering only the last 20 s of the trace [4]. The observed end of the initial burst is measured as the last instant of the trace when two consecutive samples of the slope approximation sequence surpass the maximum slope of the throttling phase.

Figure 5 (a) depicts the CDF of the amount of data (measured in seconds of video data) downloaded until the observed end of the initial burst for all downloads of the traces. The results show that the majority of the measured sizes amount to approximately 52 s and 65 s of the video data. For the remaining downloads, the empirical measurements of their initial burst slightly differ from the above mentioned two values, which is caused by short fluctuations in the mobile network’s available bandwidth that affect the empirical estimation.

Our examination highlighted that majority of the collected traces from UNSHAPED and SHAPED traffic sources has the same initial burst size measures in video seconds. However, in case of SHAPED traffic the deviation is a bit higher. Only the download duration of the initial burst is different: it is significantly higher in case of UNSHAPED traffic than SHAPED. It is caused by the traffic shaping function of the aforementioned media optimization platform, limiting the available bandwidth for a given video clip, therefore helping the network not be overloaded.

It should be noted that in opposite with the results of [4] setup parameter burst is not sent via the HTTP request by the YouTube client anymore. Our experimental result show that the initial burst size is not limited to 40 s but can take on different values, higher than earlier.

B. Throttling algorithm

We continue our discussion of the experimental analysis by focusing on the traffic received by the player after the initial seconds of a progressive download. As shown in Figure 4 (b), after the initial burst, the slope of the accumulated received data at the player was reduced because of a decrease in the receiving data rate. Figure 4 (b) shows that after the initial burst the steepness of the slope remains approximately constant until the download is completed. It is caused by the server, which throttles down the traffic generation rate increasing the total time required to complete file download. From the collected traces throttling factor can be calculated easily:

$$Throttling\ factor = \frac{Total\ data\ rate}{Throttling\ data\ rate} \qquad (3)$$

From the results of Figure 4 it can be concluded that after the initial burst, the media server throttles down the traffic generation rate, thereby avoiding transferring the data at the maximum available bandwidth. A throttling algorithm is applied with a throttling factor of 0.92 of the video total data rate. Figure 6 also depicts that there is no difference in throttling factor of UNSHAPED and SHAPED traffic sources.

This throttling procedure is also used in other platforms. It saves bandwidth of media files that might not be played to the end [7]. Additionally, it prevents congestion both at the server and the network because the data transfer is not performed at the Internet’s maximum available bandwidth.

C. Chunk size

In the previous section we highlighted that traffic generation rate is constant during the throttling period. If we magnify Figure 4 (a), it is clearly visible that traffic consists of small chunks. Figure 4 shows that during the throttling phase, the pattern of reception of data alternates between the reception of data chunks and short periods without packets.

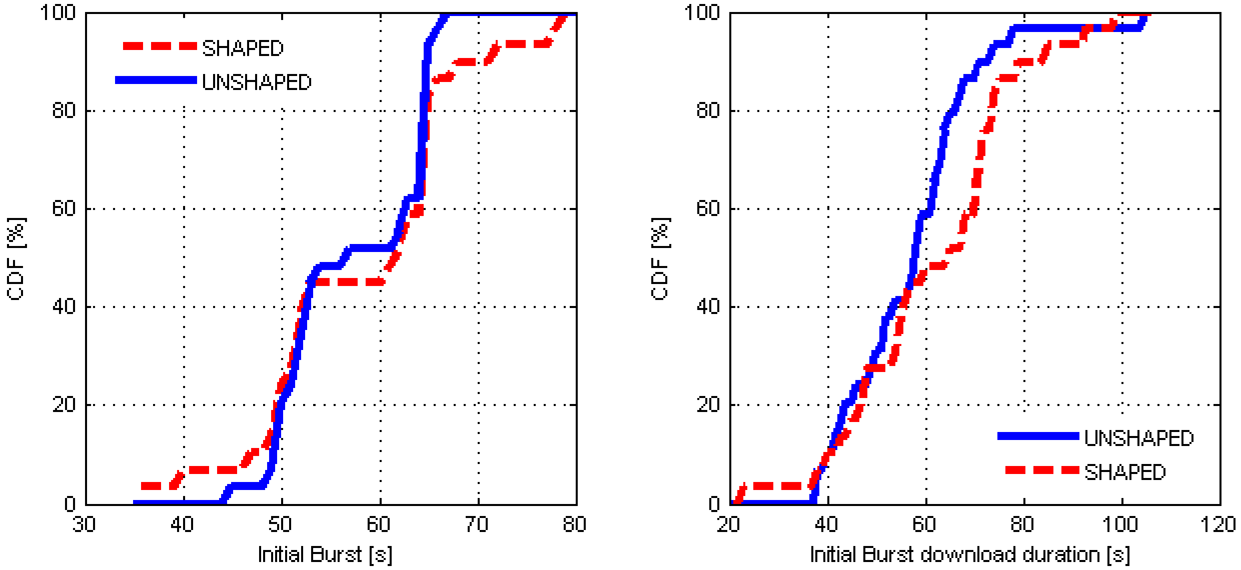

Figure 5. Initial burst measures: (a) Size in video seconds, (b) Download duration

ISBN: 978-1-61208-398-8

To further analyze this characteristic, all video clips of trace set UNSHAPED and SHAPED were postprocessed. It eliminates the initial burst of each download and additionally, groups packets into chunks so that two consecutive packets belong to the same chunk if the difference between their arrival times does not exceed a given time threshold. If the difference is longer than the time threshold, the two consecutive packets are assumed to belong to different chunks. Thus, the size of a chunk can be calculated simply by aggregating the size of the payloads of all of its TCP packets. The time threshold used to decide if two consecutive packets belong to the same chunk is selected to be 200 ms.

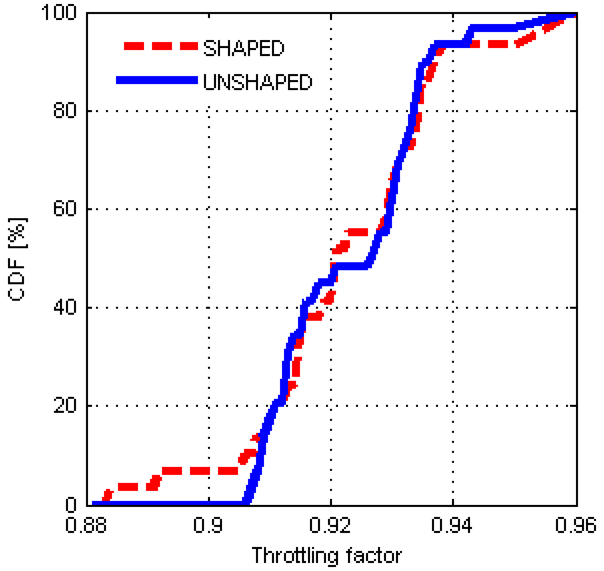

Figure 6. Throttling factor

From the empirically measured chunk sizes, we observe that in case of audio majority of the measured chunk sizes are equal 240 KB. Marginally, chunks with size of 182 KB were also found. We cannot observe significant difference between UNSHAPED and SHAPED traffic sources. In case of video we observe chunks with sizes between 400 and 1500 KB. Very small chunks (40-52 bytes) are ignored since these are identified as ACK messages in the TCP sessions.

V. Traffic Model

On the basis of the experiments presented in previous sections, a common synthetic model of the UNSHAPED and SHAPED traffic sources are proposed by means of a basic state machine (see Figure 7.) with pseudo-code describing its mode of operation (see Figure 8.). The algorithm provides the time instants and burst sizes in bytes to send on the TCP layer. The two examined traffic sources are distinguished with the parameters of the state machine as follows.

TABLE II. . Parameters of initial burst length distribution

	a	*b*	*μ1*	*σ1*	*μ2*	*σ2*
UNSHAPED	51.72	48.28	50.54	3.77	64.29	1.32
SHAPED	44.83	55.17	48.24	5.62	66.11	5.17

The generic traffic model consists of two states: an initial burst and throttling state. At the start of a video playback the algorithm needs *d* (video total duration in seconds) and *s* (video total size in bytes) as input parameters.

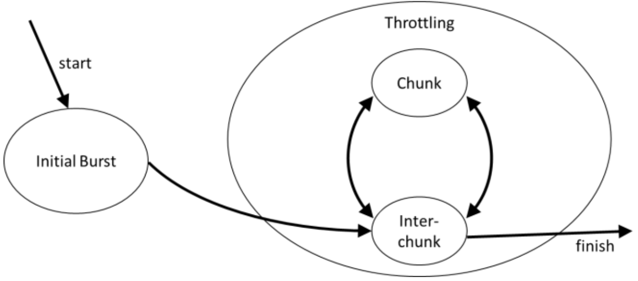

Figure 7. Generic traffic model

To set up the initial burst state first we need to calculate the size of the burst to send in bytes. To achieve this, we recall that distribution of d_{ib} (initial burst size in video seconds) shows two distinct values with high probability. According to Equation (4) we use the weighted sum of two distinct normal distributions as approximation. Parameters of the distribution formula of d_{ib} can be obtained from Table II. The s_{ib} parameter (size of the initial burst in bytes) can be calculated based on d_{ib} easily with Equation (5).

$$d_{ib} = a * \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + b \ * \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \qquad (4)$$

$$s_{ib} = \frac{d_{ib}}{d} * s \ [bytes] \qquad (5)$$

While having *b* available bandwidth, in the initial burst state the algorithm has to send the first s_{ib} bytes of the video to the player with *b* available bandwidth. Algorithm maintains a variable $s_{remaining}$, which contains the bytes still has to send to the player out of *s*. Figure 8. presents to pseudo code of each state of the state machine.

After sending an initial burst with s_{ib} bytes, in the throttling state the procedure write blocks of *cs* (chunk size in bytes) of data into the TCP socket with a period controlled by the *tf* (throttling factor). Chunks are generated based on identified chunk sizes.

```
//initial burst
While $s_{remaining}$ > s - $s_{ib}$
        Send initial burst with b bandwidth;
Endwhile;

//throttling state
While $s_{remaining}$ > 0
        Send chunk with size cs
        Sleep for cs / [(s/d) * tf] seconds
Endwhile;
```

Figure 8. Pseudo-code of the states

ISBN: 978-1-61208-398-8

We have compared the download rates from the original YouTube and synthetic model traces for every video clip. For the comparison, the instantaneous relative error of the accumulated amount of data has been computed at every sampling instant *n* as:

$$\varepsilon[n] = \frac{\left|\hat{A}[n] - A[n]\right|}{A[n]}$$

where $\varepsilon[n]$ denotes the instantaneous relative error, $A[n]$ represents the amount of the accumulated data received by the player's buffer in the case of the download from the original YouTube server and $\hat{A}[n]$ represents the amount of the accumulated data from synthetic model. It has to be noted that period between consecutive samples of the discrete-time sequence $A[n]$ and $\hat{A}[n]$ was set to 100 ms. Finally the 90th percentile of the discrete-time sequence $\varepsilon[n]$ has been computed and denoted as $\hat{\varepsilon}$. The results show that the relative error $\hat{\varepsilon}$ does not exceed 8%.

VI. Conclusion

In this paper we described the characteristics of YouTube traffic from the viewpoint of mobile broadband access. It is very valuable for predicting the video quality perceived by end-users and enhancing network design. The characterization is based on our executed experiments.

The present results have shown, that YouTube traffic has a bursty nature when accessing via mobile broadband connection. It has been also identified, that media optimization has its effect on bursty YouTube traffic: using lossy media optimization with just-in-time delivery function can visibly decrease the burstiness of the traffic. We have verified this via first order properties like SCV, PMR and m_3, and second order properties like IDI and IDC.

We have depicted the differences between UNSHAPED and SHAPED traffic sources: we have presented parameters of initial burst, throttling factor and chunk size for both traffic sources based on our experiments. It was also justified that SHAPED traffic source has the same amount of bytes in the initial burst as UNSHAPED, but is consumes less bandwidth, it takes longer for SHAPED traffic to download the initial burst. It was highlighted that initial burst size parameter is not sent via the HTTP request by the YouTube client anymore; based on our experimental result it is not limited to 40 s.

We proposed a generic traffic model for the examined traffic sources and we also presented its parameters for UNSHAPED and SHAPED YouTube videos. The model is given with its formulas and can be easily implemented in network simulation tools to evaluate service performance and end-user quality. In future work we plan to extend our model with the differentiation between audio and video streams.

References

[1] Sandvine Global Internet Phenomena Report 1H 2014 [Cited 2014 Oct 22]. Available from: https://www.sandvine.com/downloads/general/global-internet-phenomena/2014/1h-2014-global-internet-phenomena-report.pdf. May 15, 2014.

[2] S. Molnár, Gy. Miklós, "On Burst And Correlation Structure of Teletraffic Models (extended version)" 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, 21-23 July 1997, West Yorkshire, U.K.

[3] V. S. Frost, B. Melamed, "Traffic Modelling For Telecommunications Networks", IEEE Communications Magazine, March, 1994

[4] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analysis and modelling of YouTube traffic", Transactions on Emerging Telecommunications Technologies, 2012

[5] Characterization of trace sets T1 and T2. [Cited 2014 Sept 1]. Available from: http://dtstc.ugr.es/tl/downloads/set_t1_t2.csv

[6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: a view from the edge", In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, 2007, DOI: 10.1145/1298306.1298310.

[7] Microsoft Corporation. IIS Media Services. [Cited 2014 February 27]. Available from: http://technet.microsoft.com/en-us/library/ee729229(WS.10).aspx. June 10, 2010.

[8] Citrix ByteMobile. Applying Adaptive Traffic Management: Improving Network Capacity and the Subscriber Experience. [Cited 2014 October 10.]. Available from: https://www.citrix.com/content/dam/citrix/en_us/documents/products-solutions/applying-adaptive-traffic-management-improving-network-capacity-and-the-subscriber-experience.pdf. 2013.

[9] Wireshark Corporation. Wireshark network protocol analyizer. [Cited 2014. October 10.]. Available from: https://www.wireshark.org/

[10] Microsoft Corporation. Microsoft Network Monitor 3.4 [Cited 2014. October 10.]. Available from: http://www.microsoft.com/en-us/download/details.aspx?id=4865

[11] Youtube Corporation. YouTube APIs and tools.[cited 2014 April 25]. Available from: http://code.google.com/intl/en-US/apis/youtube/overview.html.

[12] G. Horváth, "End-to-end QoS Management Across LTE Networks", In the proceedings of SoftCOM Conference, 2013 DOI: 10.1109/SoftCOM.2013.6671871