Performance Measurement Methods for OTT Media Streaming

ITU Workshop on Telecommunication Service Quality

Istanbul, September 2018

Ahmet DAĞLAR

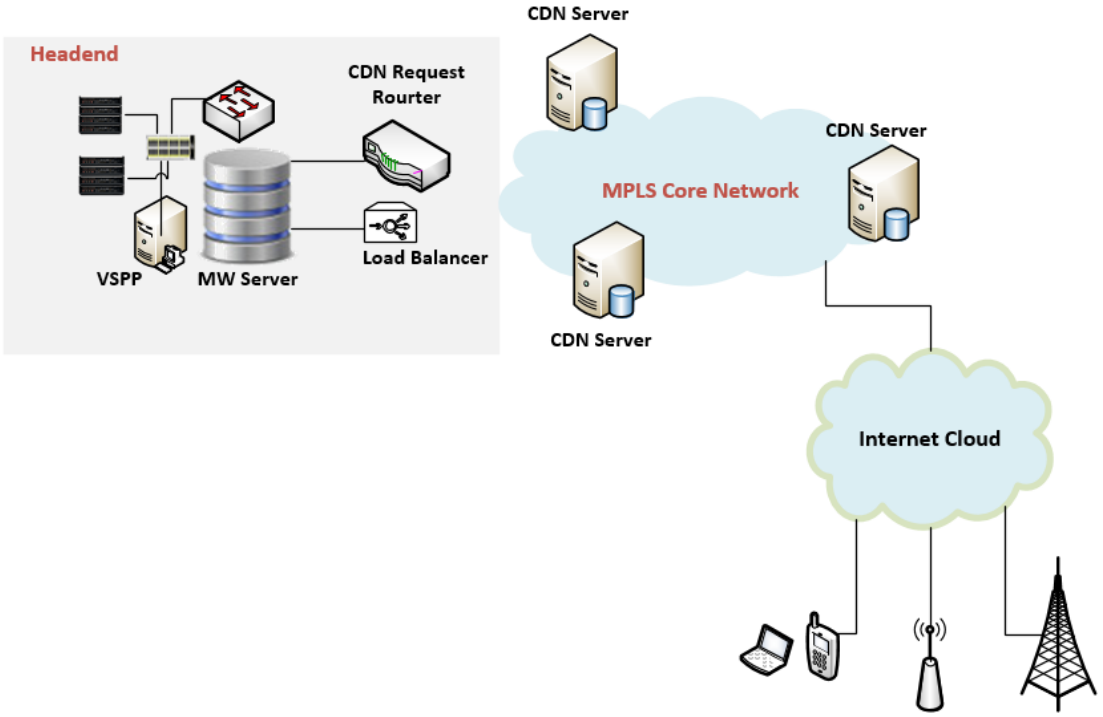
Türk Telekom

Overview

- OTT Network Architecture for Telecom Operators
- OTT Streaming Performance Measurements on Headend
- Adaptive BitRate Streaming Technique
- HLS Transport Protocol
- OTT Streaming Performance Measurements on CDN
- OTT Streaming Performance Measurements on Mobile Core Network

OTT Network Architecture for Telecom Operators

- ✓ Over the top (OTT) is a media distribution practice that allows a streaming content provider to sell audio, video, and other media services directly to the consumer over the internet.
- ✓ The below picture reflects basically required network domains in a Telecom Operator architecture to offer OTT service.



OTT Network Architecture for Telecom Operators

Following steps summarize end to end **logical flow** for a subscriber:

1- Initially client authentication process is evaluated by MW server.

2- Than if customer clicks the play button on the screen of application, this time MW server handles the received http get request. And the Request Router diverts the client through out appropriate CDN server.

3- Than client sends manifest request through CDN server and in response downloads an index file at the beginning of the streaming session.

4- The streaming is started (ABR based transport protocol) by the CDN server as unicast.

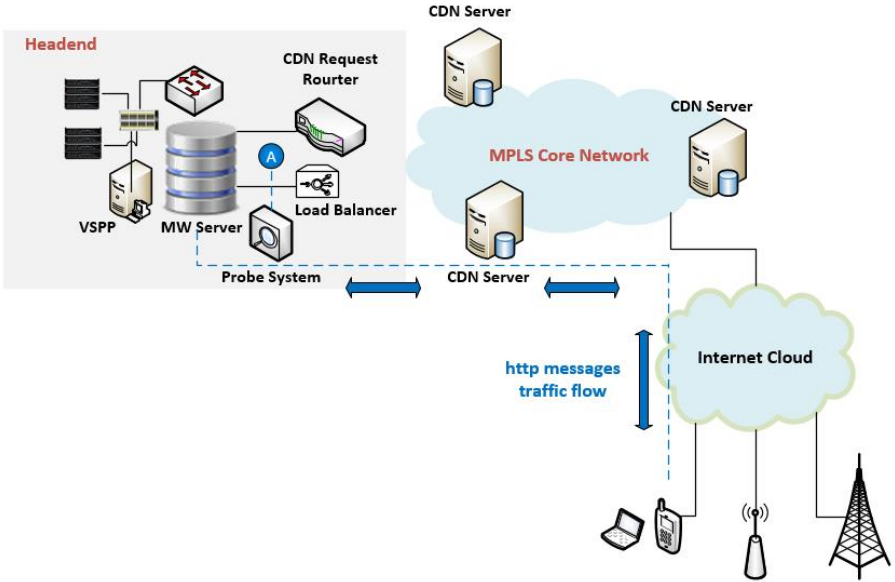
OTT Streaming Performance Measurements on Headend

Traffic Flow

Load Balancer located on Headend is the first equipment which meets application requests (VOD, TimeShift, Catchup, Live TV) of Clients. And it dispatches the received traffic to MW server within sequentially.

Measurement Point and Method

Probe system is located between MW server and Load Balancer *(refer to point A on picture)* so that the application URL traffic is replicated for each http session. The replicated traffic is analyzed at http level. Success rates are determined by calculating the rate of Client application requirements that have successfully answered by MW server. At the same time measure response times of MW server is calculated for performance analyses.



OTT Streaming Performance Measurements on Headend

Quality Metrics

***MW Server Login success rate*:** It's ratio between number of total Client authentication attempts and corresponding successful MW server responds based on http login messages.

MW Server VOD success rate: VOD is an application which allows customers to select and watch the video contents (movies or TV shows) whenever they choose from the special catalog rather than a scheduled broadcast time. VOD requests are sorted from whole http traffic based on URL extensions. And calculation is executed as follows:

$$= \frac{\text{number of successful MW server responds}}{\text{amount of http get messages for VOD aplicaiton}}$$

OTT Streaming Performance Measurements on Headend

Quality Metrics

***MW Server Timeshift TV success rate*:** TimeShift TV is a feature that gives customers making interactions during live media streaming. Through this feature, customers can make such as stop/resume, rewind or restart-up the content during streaming.

TimeShift TV requests are sorted from whole http traffic based on URL extensions. And calculation is executed as follows:

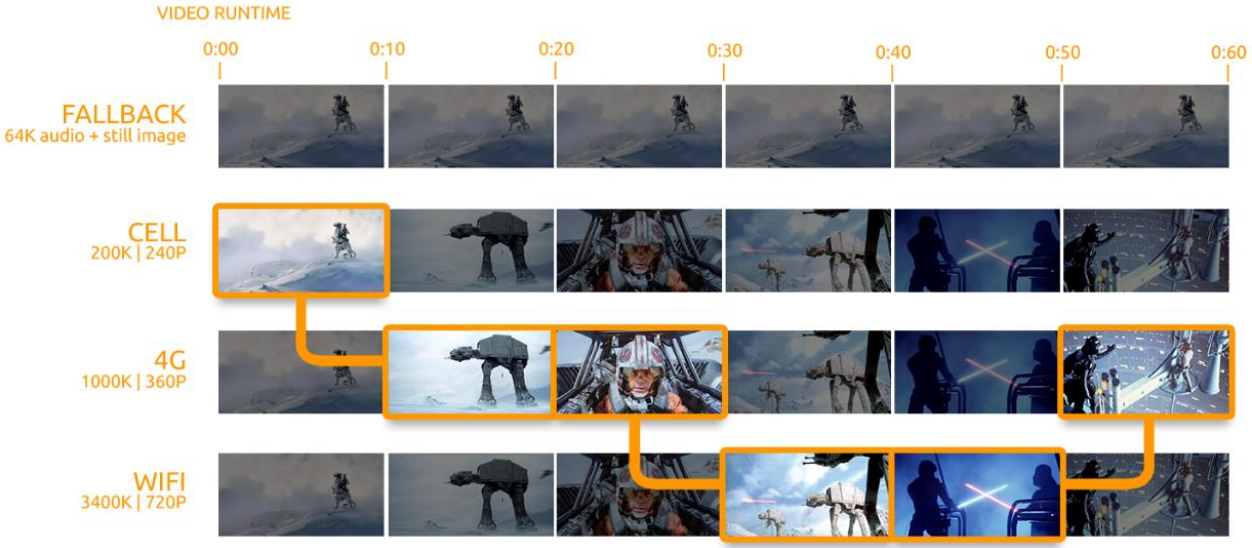
$$= \frac{\text{number of successful MW server responds}}{\text{amount of http get messages for Timeshift TV interactions}}$$

***MW Server Catchup TV success rate*:** Customers can watch old programs which are listed on electronic program guide whenever they want. We call this kind of interaction such as Catchup TV application. Similarly, Catchup TV requests are sorted from whole http traffic based on URL extensions. And calculation is executed as follows:

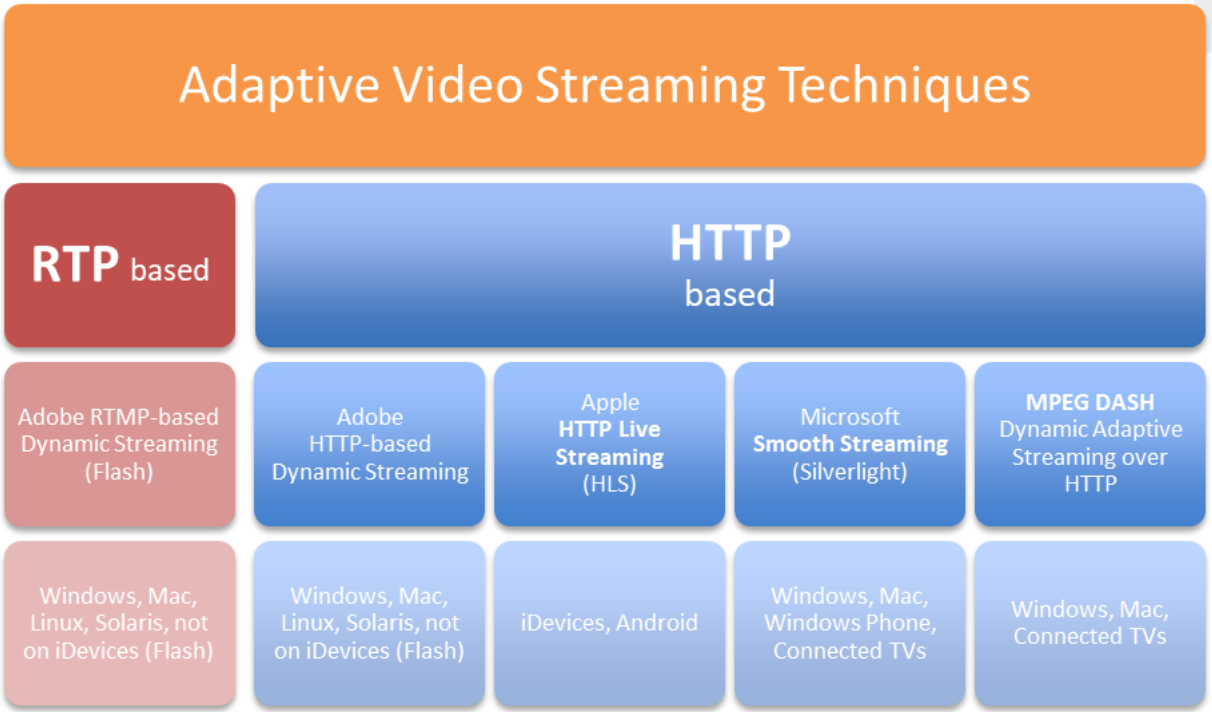
$$= \frac{\text{number of successful MW server responds}}{\text{amount of http get messages for Catcup TV interactions}}$$

Adaptive BitRate Streaming Technique

- ✓ Adaptive bitrate (ABR) streaming technique is a combination of server and client software that detects a client's bandwidth capacity and adjusts the quality of the video stream between multiple bitrates and/or resolutions.
- ✓ Thus video streaming keeps running by switching to another defined lower speed profile.
- ✓ When a mobilized customer moves to different coverage zone where network quality is lower, despite poor network performance ABR allows video sessions to continue. Because the client can dynamically request the available best bit rate option, given their specific bandwidth during the same session.



Adaptive BitRate Streaming Technique

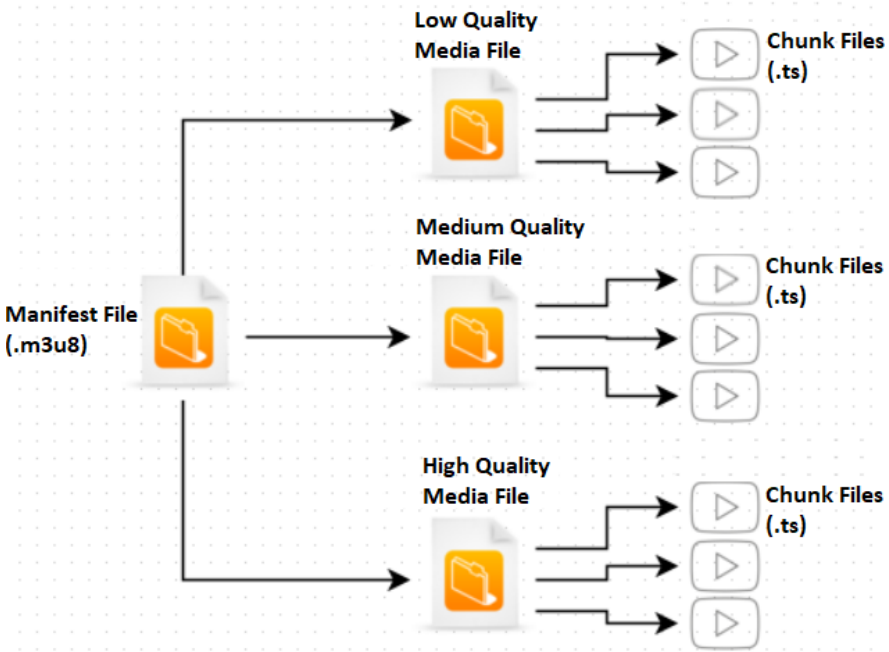


(based on Jan Ozer: 'Video Compression for Flash, Apple Devices and HTML5' [2011])

HLS Transport Protocol

HLS stands for HTTP Live Streaming. It is an ABR-based media streaming communication protocol implemented by Apple. HLS consists of three types file:

- ✓ The stream is presented within small ***chunk files***. Each chunk contains 5-10 seconds video.
- ✓ These chunk files are encoded within various bit rates and sorted in ***media files***. Thus each video content is replicated with different profiles.
- ✓ The file that holds all possible profile information for a content is called the ***manifest file***.



OTT Streaming Performance Measurements on CDN

Traffic Flow

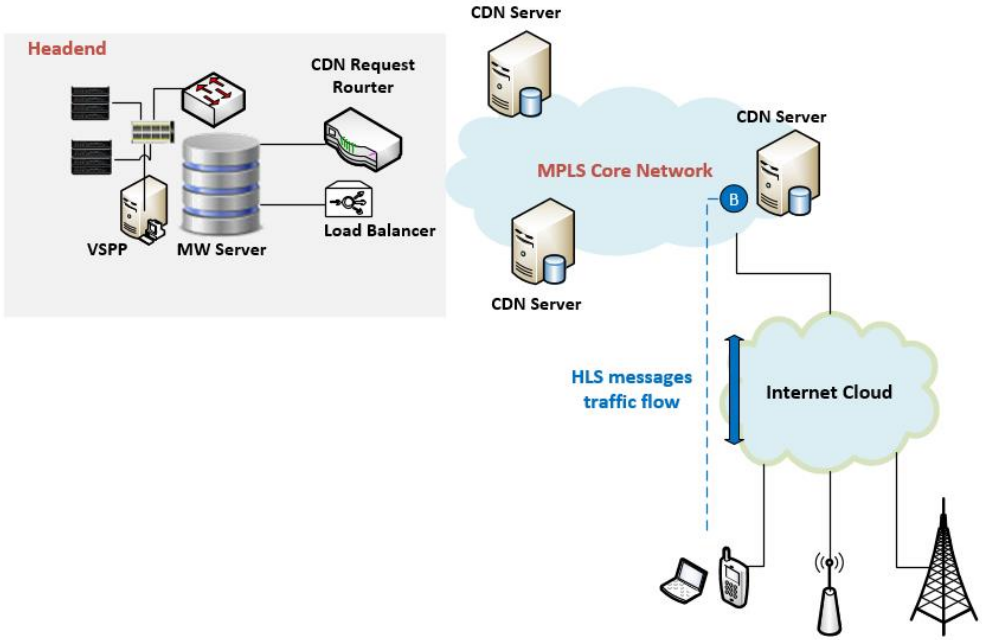
Once a client (WEB, Smart Phone etc.) received successful response for the playback request by MW server, it's directed through most appropriate CDN server which is located in Fixed Core network.

Than client sends Master manifest request through CDN server and in response downloads an index file at the beginning of the streaming session. This file contains the multiple media files of the requested content at different data rates and they are divided into short chunks. So streaming starts and continue over these short chunks (5-10 second).

Measurement Point and Method

The logs of CDN servers *(refer to point B on picture)* are collected periodically on a big data platform to make http url type based analyze. Total Master Manifest requests by clients are identified and calculation is made to clarify corresponding successful CDN server responses against theme.

Basically if the manifest file at the beginning of a session was not successfully downloaded by client, this mean as customer failed to start streaming of the requested content.



OTT Streaming Performance Measurements on CDN

Quality Metrics

Basically following function will be executed for each application type (LiveTV, VOD etc.)

$$= \frac{\text{number of successful CDN server responds}}{\text{amount of HLS Manifest requests sent by clients}}$$

LiveTV streaming launch success rate: Only Live TV specific http messages are reserved from whole CDN server logs . And CDN server responds to Manifest requests sent by clients are investigated based on http status codes.
Then the ratio between number of total Master Manifest http messages (.m3u8 extension) and corresponding successful server responds is calculated.

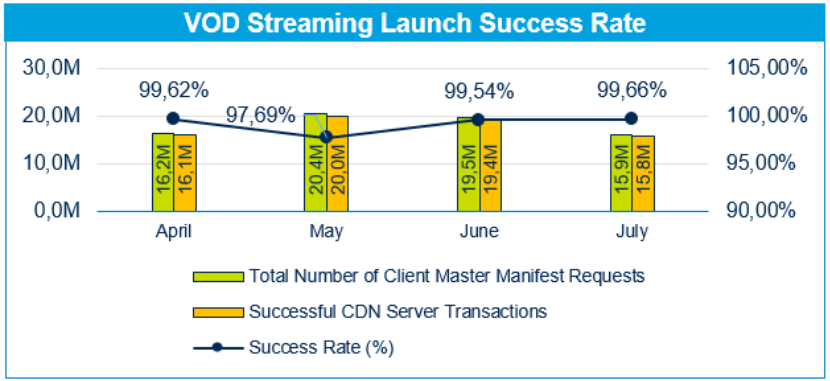
Timeshift/Cathup TV streaming launch success rate: Only for TimeShift/Catchup interaction specific http messages are reserved from whole CDN server logs . And CDN server responds to Manifest requests sent by are investigated based on http status codes.
Then the ratio between number of total Master Manifest http messages (.m3u8 extension) and corresponding successful server responds is calculated.

OTT Streaming Performance Measurements on CDN

Quality Metrics

VOD streaming launch success rate: Only VOD content specific http messages are reserved from whole CDN server logs . And CDN server responds to Manifest requests sent by clients are investigated based on http status codes.

Then the ratio between number of total Master Manifest http messages (.m3u8 extension) and corresponding successful server responds is calculated.



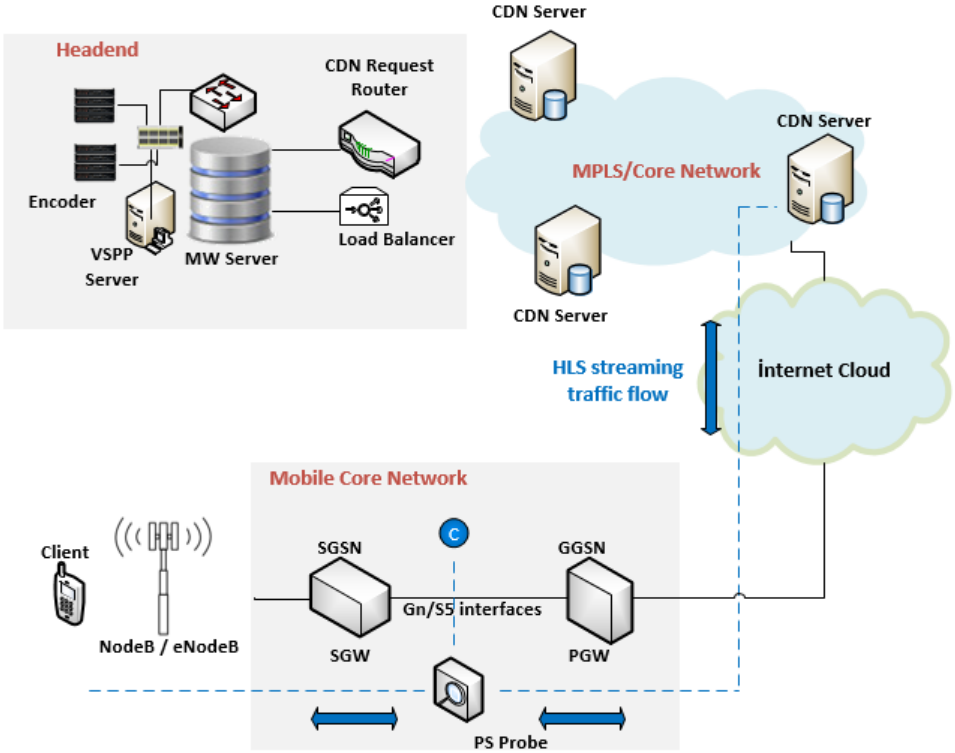
OTT Streaming Performance Measurements on Mobile Core Network

Traffic Flow

Once a mobile client successfully downloaded the manifest file, it requests chunk files periodically from the CDN server. This HLS traffic flows through the mobile core network and is transmitted to the RF location of the subscriber. Meanwhile the buffer of client player begins to full fill and then video playback has started on customer screen.

Measurement Point and Method

OTT streaming traffic is captured and as well as DPI analyzes are performed by using PS probes located in the mobile network *(refer to point C on picture).* After than calculations are executed based on the mathematical model.



OTT Streaming Performance Measurements on Mobile Core Network

Quality Metrics

Operators offer their own Client application software for customers' use. So customers download the application on their equipment (mobile phone, tablet, notebook etc.) over internet.

With this application, player codec parameters such as profile types, startup profile, client buffer amount (which are used in ABR based protocols) can be set as fixed values.

So that all Clients will follow the same rules throughout the ABR-based transmission, regardless of equipment brand/model.

P (kbps): startup profile,

T (s): the video lenght that is carried in a chunk file,

C : the number of chunk files transferred by the server until the amount of client buffer is full

Let's imagine that the application software is configured with following values:

P:500kbps

T: 5s

C: 3

Profile Types	Resolution	Video (kbps)	Audio (kbps)
Highest Profile	720x576	2500	128
High Profile	720x576	1500	128
Low Profile	640x480	1000	128
Lowest Profile	480x360	500	64

OTT Streaming Performance Measurements on Mobile Core Network

Quality Metrics

Average Streaming response delay: This metric indicates the amount of time a subscriber waits until video playback is launched on the screen after clicking the play button. According to previously fixed values, it requires 3 chunk files to fill client buffer before the playback can be started. If so, corresponding the amount of video data that must be transferred by clients can be easily calculated as follow:

= 500kbps (startup profile) x 3 (http chunk quantity) x 5s (video duration of each chunk)
= 938 kbyte

The DPI analyzes will be carry out to measure how much time it takes to download 938 kbyte of video data for each session.

Streaming start success rate: This metric indicates the rate of how many media content are successfully played on screen after customer clicks the play button. With the DPI analyze, total number is specified for the sessions which transmitted at least 938 kbyte media stream. And than following ratio is calculated:

$$= \frac{\text{number of sessions} > 938 \text{ kbyte}}{\text{amount of customer sessions}}$$

OTT Streaming Performance Measurements on Mobile Core Network

Quality Metrics

Average Streaming throughput: Session based Customer throughput experience is achieved by following function:

$$= \text{average}\left(\frac{\text{amount of downloaded streaming data}}{\text{duration of streaming session}}\right)$$

The calculation is carried out for the sessions that at least one minute playback has been exist by customers. So that negative affect of TCP protocol at beginning phase of sessions can be excluded for calculations. Corresponding value to one minute playback time can be easily calculated as follow:

= 500kbps (startup profile) x 12 (http chunk quantity) x 5s (video duration of each chunk)

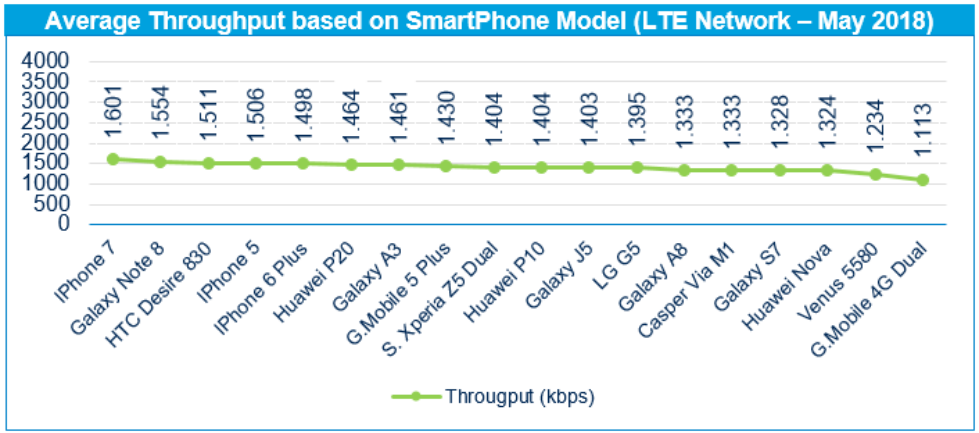
= 3750 kbyte

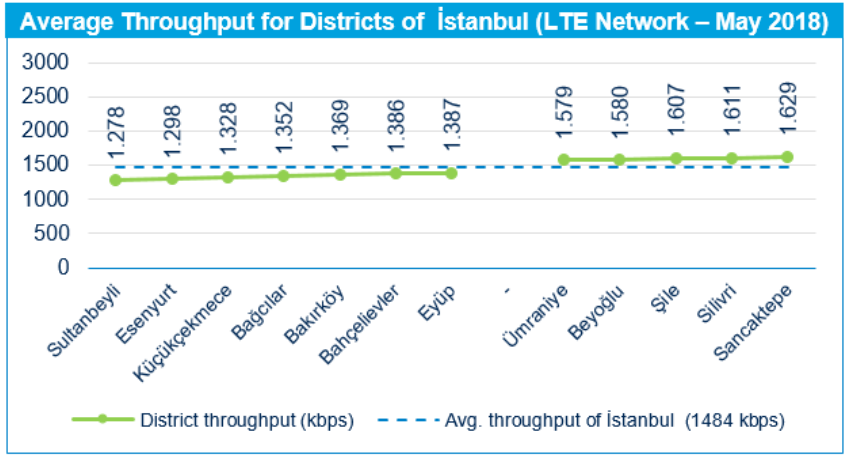
OTT Streaming Performance Measurements on Mobile Core Network

Quality Metrics

Also each metric can be broken on bases of region, equipment brand/model or access type (3g, LTE) for Operator network.

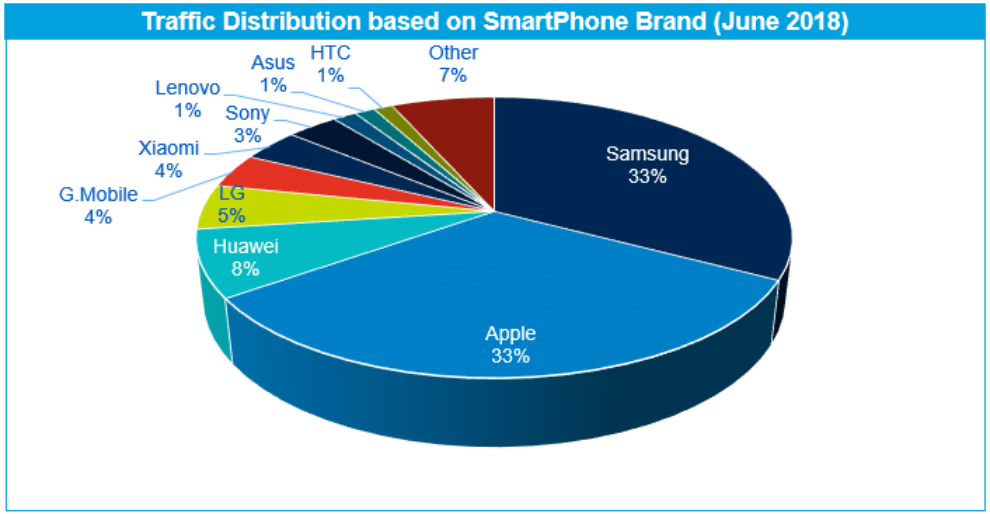
Below pictures reflect measured sample throughput realizations for OTT media streaming in Türk Telekom network.





OTT Streaming Performance Measurements on Mobile Core Network

Quality Metrics



Thanks