Big Data Initiative: Effective Caching in Online Video Platforms Draft*

EdN:1

Rongrong Bao Atabak Hafeez Tom Wiesing
Jinbo Zhang

October 31, 2015

Abstract

Data on the internet grows by 50 percent annually. More than 90% of the data has been generated in recent years. This is the time for big data. How can we effectively transfer this huge amount of data?

We want to investigate caching techniques used by online video platforms and in particular by YouTube. YouTube is a leading online video provider worldwide. Before 2012, video streaming in YouTube was done using Real Time Messaging Protocol (RTMP)-based servers. This requires a streaming server and a near-continuous connection between the server and user. Requiring such a streaming server can increase implementation cost and RTMP-based video streaming is at risk of being blocked by firewalls. In 2012, this was replaced by HTTP (Hypertext Transfer Protocol) based servers known as MPEG DASH (Dynamic Adaptive Streaming over HTTP). HTTP is the protocol used by websites to bring their content to the users. By using this technology it was possible to use existing optimizations in the form of HTTP-Caching. This capability decreased total bandwidth costs associated with delivering the video since videos would be served from web-based caches rather than the origin server. This improved quality of service, since cached data is generally closer to the viewer and more easily retrievable.

The essay will explain and discuss different kinds of caching techniques, optimizations, data analysis and prediction techniques used by YouTube, including their advantages/disadvantages and potential social impacts.

*EdNote: Remove draft status

Contents

1 Introduction

[2] EdN:2

[2] EdNote: Write introduction

2 What is Caching?

Before we delve into the specifics of how Youtube does caching and the advantages and disadvantages of the different kinds of caching, we will explain how caching itself is and also explain the different possibilities available to use it in Youtube streaming.

Caching is a method to store data in cache. A cache is a basically temporary storage area on the local hard disk of a user. This storage area may contain data such HTML pages, images, files, and Web Projects in order to make it faster for the user to access it, which helps improve the efficiency of the computer and the overall efficiency of the task at hand. The important thing to note here is that occurs without the user knowing about it. For example, when a user returns to a Web page they have recently accessed, the browser can pull those files from the cache instead of the original server because it has stored the user?s activity. The storing of that information saves the user time by getting to it faster, and lessens the traffic on the network.

2.1 Caching and Buffering in Youtube

To explain caching in Youtube and how it has changed and updated, we also need to understand the notion of buffering. Buffering involves pre-loading data into a certain area of memory known as a "buffer". This is basically a a more specific kind of caching which Youtube uses to store the loaded video on to the local memory of the browser in use.

In 2013, Youtube made a design decision in their buffering system where they moved from Real Time Messaging Protocol (RTMP)-based Dynamic Streaming to MPEG DASH (Dynamic Adaptive Streaming over HTTP).

To a user, this is important because it changes the extent to which you can cache you Youtube video before viewing it. Basically, while Youtube was using RTMP-based Dynamic Streaming, if a user had a relatively slow connection, which would not allow them to view the video as smoothly as one would want, he could pause the video and view it later when the whole video is buffered or cached to the local storage of the browser. This technology required a near-continuous connection between the server - the original storage location where one's local computer is connected to retrieve the video - and the player on one's browser.

With Youtube's shift to using MPEG DASH, being able to buffer the whole video and then coming back to it was no longer possible. MPEG DASH use standard HTTP web servers to deliver streaming content, obviating the need for a streaming server. In addition, HTTP packets are firewall friendly and can utilise HTTP caching mechanisms on the web. To an average user, this means that now when he or she pauses a video because the video is not very smooth, the video buffers for a while and then stops

buffering. The cache, hence, does not at this stage store the whole video. The only condition in which the video may cache the whole video is if you start the video and watch the whole way through. In this case, if the user turns off the internet, it would be possible for the user to re-watch the video without having to reload it from the server.

2.2 Caching on the Servers

So why did Youtube decide to make this significant change in their protocol? This is mainly because this makes streaming high quality videos more efficient because it can use caching on the servers.

Caching on the server means that when a user connects to a server, the connection is not direct. There is something in-between known as a caching server. The caching server acts as a web proxy server so it can serve those requests. After a web proxy server receives requests for web objects, it either serves the requests or forwards them to the origin server (the web server that contains the original copy of the requested information). Using MGEP DASH, Youtube was able to exploit this feature resulting in higher resolution videos being available to the user in a more efficient manner.

This form of caching also aims to make sure that the user can experience the best quality of video according to the available brandwidth speed. For example, a video may start playing at 360P resolution, but if the system detects that the brandwidth is now able to handle 720P, it will shift to that. Thus, using MPEG DASH, Youtube only caches chunks of the video and only a small chunk of them are loaded when a video is paused.

References:

3 Advantages of Caching

[3] EdN:3

[3] EdNote: Write advantages

4 Disadvantages of Caching

While caching videos in the users web browser has many useful effects as discussed above it also has several bad side effects. In this section we will first explain some of the technical side effects and then continue discuss some social implications[4]. EdN:4

4.1 Technical side effects

In order to take advantage of caching of videos the browser needs to create caches for all the videos that the user plays. Additionally every time a video is played the browser has to check if a cached version of the video already exists. If this is the case the video can be played from the cached version. If this is not the case however the browser has to either create a cached version of the video or play a direct version from the server. Both of these decisions have a negative performance impact since the video has to be downloaded from the server which is slower then playing a local version.

Furthermore a cached version might not always be up-to-date. It is conceivable that when a news station has uploaded a video to YouTube and later on finds an error in their report that they might update their video in order to provide the most correct and up-to-date information. If this video has been cached by a viewer of the video they might no longer have the newest version available in their cache. Thus in order to use caches properly every time the cache is used the web browser needs to check if it is still up-to-date. This process can be further complicated if only parts of the video are cached. Furthermore if the cache is invalidated (meaning it is found to no longer be up-to-date) it is inefficient to reload the entire video if only parts of it have changed.

It is also possible that some videos that are cached even though the user does not want to watch them. This is for example the case when adverts are played before the video. They are not desired by the user and might nonetheless be cached. Sometimes the users might also watch a few seconds of the beginning of the video and then decide they do not want to watch the remainder. This can cause both a waste of bandwidth and local hard drive space.

Since caching technically copies the videos from the server to the hard disks of the local user the video is no longer stored in only one central location. This means that even if the video has to be deleted (for example for legal reasons) it might still be available locally. This might mean legal hurdles if a license has to be obtained in order to show the video.

Additionally some of the videos the user watches might remain cached on the users hard drive even after they have watched the video. While a single video does not take too much space this can cause a problem if there

[4] EdNote: Better wording, maybe a longer introduction and link to the previous section

are many of these cached videos. It is especially important for mobile users as these typically have less space available on their devices. Furthermore the user might not want a record of the videos they watched for privacy reasons.

4.2 Social implications

When using predictive caching YouTube (or other online video providers) are using data from the user. They try to analyse the videos the user has watched in the past and want to predict which videos the user might watch in the future. Some users might not desire getting videos predicted for several reasons. Especially when on a connection with limited bandwidth the user might object to having this bandwidth used up with videos that he might not watch anyways.

Additionally several users might not want their video history to be recorded. This can easily be considered an intrusion into the users privacy. This information could also be used against the user [5] especially when the wrong videos are predicted by the algorithm. EdN:5

YouTube sells anonymised customer data to advertisers [6]. If sold to advertisers, this data is commonly used to predict products the user might be intersted in. Related adverts are then played on the next video the user watches. This brings money to the content provider but does not give any significant advantage to the user. EdN:6

[5] EdNote: Where?
[6] EdNote: Citation needed

5 Conclusion

6 References