CPS 512
Final Project Report
Dec 16th, 2017

Yuxi Yang
Yuan Yuan
Xuanyu Zhan

Introduction

The objective of this project is to study the geological distribution of Youtube CDNs, and try to demonstrate that lookups for mega-services like Youtube will return different addresses based on the locations where the requests are originated.
To implement our experiments, we wrote python scripts and sent http requests to visit Youtube videos from 47 Planet Lab servers distributed in 10 countries and 5 continents to collect our data. We also tried to send download requests to verify the locations of Youtube CDNs.

Background and Data Collection

Youtube is a renowned and leading video platform that people watch, share, and upload all kinds of videos. Youtube has its presence all around the world, therefore, it is a great choice for us to verify how mega-services handle requests from different regions. Another reason we choose Youtube to perform our experiment is because visitors do not have to login in order to watch a video, and this makes our experiment easier since creating dummy accounts is not necessary for the purpose of our project.
Youtube videos are identified by their unique video IDs, which comprised only by alphanumeric characters. With these video IDs, we firstly obtained the URL to the videos with a self-generated robust crawler. We have collected 1963 URLs in total by recursively crawl the next section of a random Youtube video page. Part of the URLs we generated are shown in Figure. 1. This randomized process eliminated the possibility that the CDNs we got are video dependent. Then our code sent requests from multiple vantage points (47 PlanetLab servers) in 10 different countries to these URLs to get the responses that contain the IP addresses of these URLs. We translated the ip addresses of those Youtube CDN servers contained in the responses into geolocations(longitudes and latitudes) using a http api provided by freegeoip.net. Then, we exported all the data we generated and collected into JSON format and saved in txt files(Sample JSON data shown in Figure. 2). Finally, we used a python graphing library called “Plotly” to generate and visualize the distributions of our nodes and Youtube’s data centers. As an extra, we also calculated the time delay between each request and response(Shown in Figure. 3) in order to try to get some senses of how well those distributed servers work.

```
https://www.youtube.com/watch?v=mQPjKSVe1tQ 173.194.142.186
https://www.youtube.com/watch?v=kJQP7kiw5Fk 74.125.7.74
https://www.youtube.com/watch?v=7mCjpYFu-J0 173.194.142.249
https://www.youtube.com/watch?v=sOqWJEbWrXs 173.194.142.250
https://www.youtube.com/watch?v=tRDjz-rtrtI 173.194.186.7
https://www.youtube.com/watch?v=UzJFrC2Te90 74.125.7.72
https://www.youtube.com/watch?v=JANApS0P4z8 173.194.142.42
https://www.youtube.com/watch?v=Uil3656ki2c 173.194.29.73
https://www.youtube.com/watch?v=hpZ4_ShYlJk 173.194.29.24
https://www.youtube.com/watch?v=oi8vSLpbjmk 173.194.186.10
https://www.youtube.com/watch?v=c64I9HNpiOY 173.194.144.8
https://www.youtube.com/watch?v=s00W64lIMBE 173.194.142.139
https://www.youtube.com/watch?v=fVF5v0RbGjo 173.194.186.7
https://www.youtube.com/watch?v=p3jUa3JOUBw 173.194.142.58
https://www.youtube.com/watch?v=7rb9eGANMt8 74.125.6.208
https://www.youtube.com/watch?v=WMpEp2Ss0pk 74.125.7.70
https://www.youtube.com/watch?v=LMUgGAdLPeE 173.194.142.41
https://www.youtube.com/watch?v=ZKErPftd--w 74.125.6.137
https://www.youtube.com/watch?v=L3yXkvYY-wY 173.194.11.210
```

Figure. 1 Sample urls generated with corresponding IP addresses


```
{"d_name": "https://www.youtube.com/watch?v=mxsgejDdFtQ", "city": "Mountain View",
"region_code": "CA", "region_name": "California", "ip": "173.194.61.23", "time_zone":
"America/Los_Angeles", "longitude": -122.0574, "metro_code": 807, "latitude": 37.4192,
"country_code": "US", "country_name": "United States", "zip_code": "94043"}
{"d_name": "https://www.youtube.com/watch?v=tvTRZJ-4EyI", "city": "Mountain View",
"region_code": "CA", "region_name": "California", "ip": "173.194.61.7", "time_zone":
"America/Los_Angeles", "longitude": -122.0574, "metro_code": 807, "latitude": 37.4192,
"country_code": "US", "country_name": "United States", "zip_code": "94043"}
{"d_name": "https://www.youtube.com/watch?v=5qxqo2CQakA", "city": "Mountain View",
"region_code": "CA", "region_name": "California", "ip": "74.125.172.23", "time_zone":
"America/Los_Angeles", "longitude": -122.0574, "metro_code": 807, "latitude": 37.4192,
"country_code": "US", "country_name": "United States", "zip_code": "94043"}
{"d_name": "https://www.youtube.com/watch?v=meRDk9GHwRU", "city": "Mountain View",
"region_code": "CA", "region_name": "California", "ip": "173.194.131.154", "time_zone":
"America/Los_Angeles", "longitude": -122.0574, "metro_code": 807, "latitude": 37.4192,
"country_code": "US", "country_name": "United States", "zip_code": "94043"}
{"d_name": "https://www.youtube.com/watch?v=AnSBbcAyQ1Y", "city": "Mountain View",
"region_code": "CA", "region_name": "California", "ip": "173.194.31.88", "time_zone":
"America/Los_Angeles", "longitude": -122.0574, "metro_code": 807, "latitude": 37.4192,
"country_code": "US", "country_name": "United States", "zip_code": "94043"}
{"d_name": "https://www.youtube.com/watch?v=jauZUjBms9U", "city": "Mountain View",
"region_code": "CA", "region_name": "California", "ip": "173.194.61.10", "time_zone":
"America/Los_Angeles", "longitude": -122.0574, "metro_code": 807, "latitude": 37.4192,
"country_code": "US", "country_name": "United States", "zip_code": "94043"}
```


Figure. 2 Data obtained in JSON format

```
https://www.youtube.com/watch?v=jauZUjBms9U 0:00:00.175246
https://www.youtube.com/watch?v=mQPjKSVe1tQ 0:00:00.095470
https://www.youtube.com/watch?v=kJQP7kiw5Fk 0:00:00.059200
https://www.youtube.com/watch?v=7mCjpYFu-J0 0:00:00.079337
https://www.youtube.com/watch?v=sOqWJEbWrXs 0:00:00.071970
https://www.youtube.com/watch?v=tRDjz-rtrtI 0:00:00.220400
https://www.youtube.com/watch?v=UzJFrC2Te90 0:00:00.071578
https://www.youtube.com/watch?v=JANApS0P4z8 0:00:00.051827
https://www.youtube.com/watch?v=Uil3656ki2c 0:00:00.086601
https://www.youtube.com/watch?v=hpZ4_ShYlJk 0:00:00.077813
https://www.youtube.com/watch?v=oi8vSLpbjmk 0:00:00.178042
https://www.youtube.com/watch?v=c64I9HNpiOY 0:00:00.052173
https://www.youtube.com/watch?v=s00W64lIMBE 0:00:00.054423
https://www.youtube.com/watch?v=fVF5v0RbGjo 0:00:00.036185
https://www.youtube.com/watch?v=p3jUa3JOUBw 0:00:00.254819
https://www.youtube.com/watch?v=7rb9eGANMt8 0:00:00.233039
https://www.youtube.com/watch?v=WMpEp2Ss0pk 0:00:00.046573
https://www.youtube.com/watch?v=LMUgGAdLPeE 0:00:00.173620
```

Figure. 3 Sample urls generated with request response delay

Data & Results

After collecting and pre-processing the data, we feed the data into “Plotly” library. Figure. 4 is the graph generated by the library. Red points show the location distribution of these vantage points. Blue points correspond to the Youtube Data centers that responded to our requests.

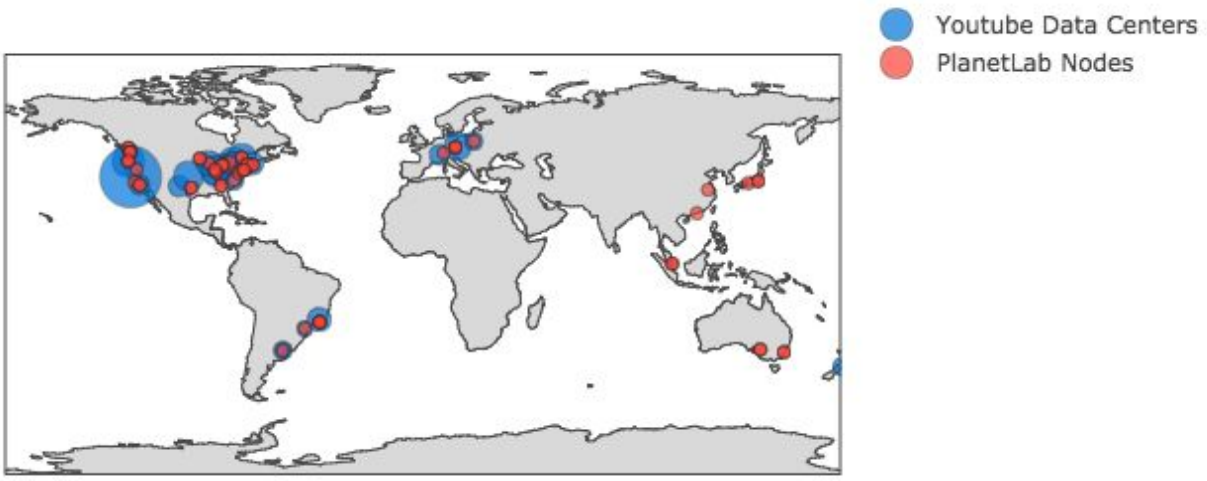


Figure.4 Geographical Distriution of CDN servers & Vantage points

Data Analysis

According to the data we have collected, first, we found that each PlanetLab node get data from 4-5 IP addresses nearby, which means Youtube has many CDNs around the world, multiple data centers may assigned to the same ip based on the workload of each, and client side get contents (requested videos) from the CDN server nearby.
With these data, we further calculated the distances between the PlanetLab nodes and its corresponding CDN servers. Then we group the PlanetLab nodes into different countries and different continents according to their geographical locations. For each group of nodes, we calculated the average distances between these nodes and the Youtube CDNs they requested.
Table 1 shows the average distance within the 10 different countries, and Table 2 shows the average distances within 5 continents. It is interesting to find out that the average distance from PlanetLab nodes to CDN servers in Canada is closer than America.

Table 1

Country	Average Distance(m)
Canada	1874133.89602
United States	2046119.37864
Czechia	3919446.40119
Brazil	6085722.7177
Switzerland	6450296.36907
Argentina	6776802.3145
Poland	7336852.6355
Japan	8539784.43415
Australia	9835841.57998
Hong Kong	11334266.238

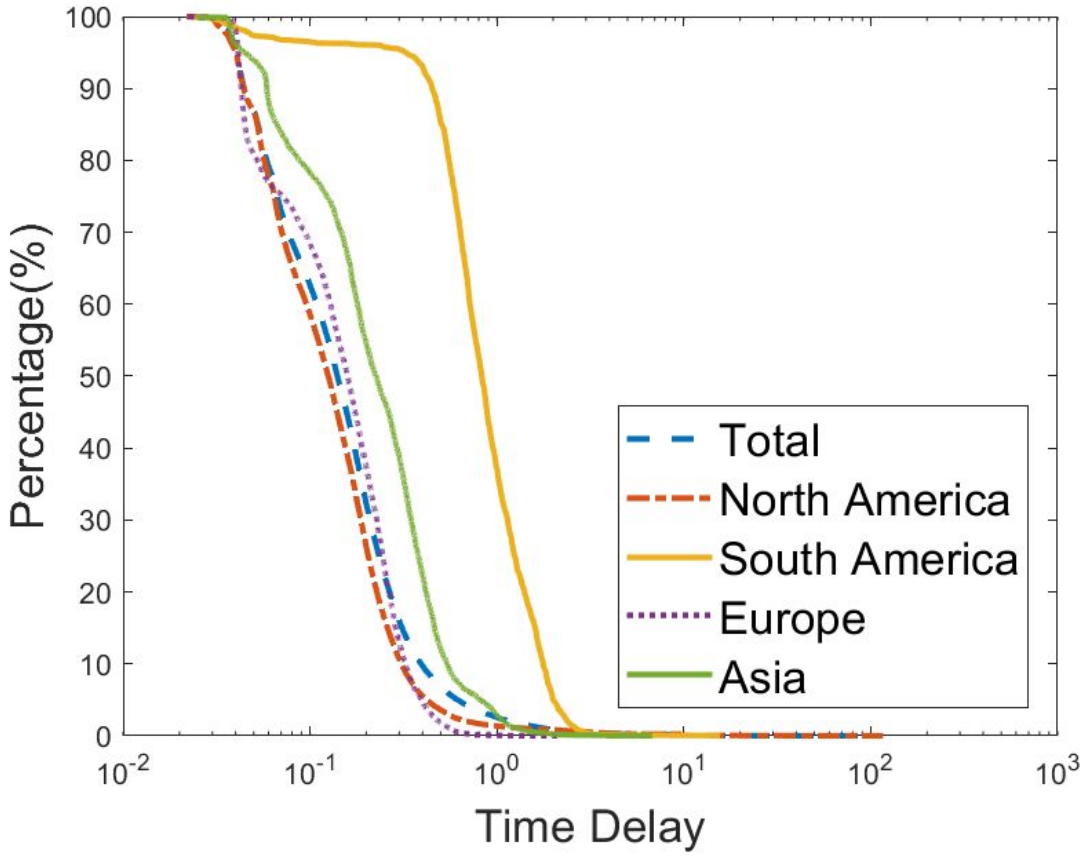
From Table2 we can see that North America has the least average distance, whereas Oceania and Asia have the farest average distance. This phenomena can be caused by the distribution of Youtube CDN servers.

Table 2

Continent	Average Distance(m)
North America	2026989.78587
Europe	5407537.22949
South America	6315960.18167
Asia	9471938.76857
Oceania	9835841.57998

Second, based on the delay time data, we generated the Complementary Cumulative Distribution Function(CCDF) of the time delay of video download requests. Based on the location vantage points who sent the request, we group the requests into 4 continents, i.e. North America, South America, Europe and Asia, as stated in Fig 2.

Fig2 CCDF of the Time Delay



Conclusion

In conclusion, Youtube CDNs are distributed around the PlanetLab nodes. We have a deeper understanding of the wide distribution and extensive use of Youtube CDN servers. However, due to lack of data from other continents other than North America, our conclusion to the worldwide spread of Youtube CDNs can be biased. The populariy of the videos and traffic competing may cause redirection.

The IP-geolocation mapping APIs may give you inaccurate results like they may return the location of the headquarters of the company that owns the IP address instead of the location of the CDN. While some more precise academic IP-geolocation databases like CBG tools and Alidade are not available to us. Besides, Google makes it hard for you to get the exact locations of their CDNs. So, IP-geolocation can also lead to bias.

Reference

[1] Ruben Torres, Alessandro Finamore et al. Dissecting Video Server Selection Strategies in the YouTube CDN, ICDCS2011