Arjun D'Cunha
CS14BTECH11039
8 March 2017

Assignment 2

Downloading a File with Multiple Parallel TCP Connections using HTTP Range Requests

Design :

For the required implementation, I first input the URL of the file to be downloaded and the number of TCP Connections to open.

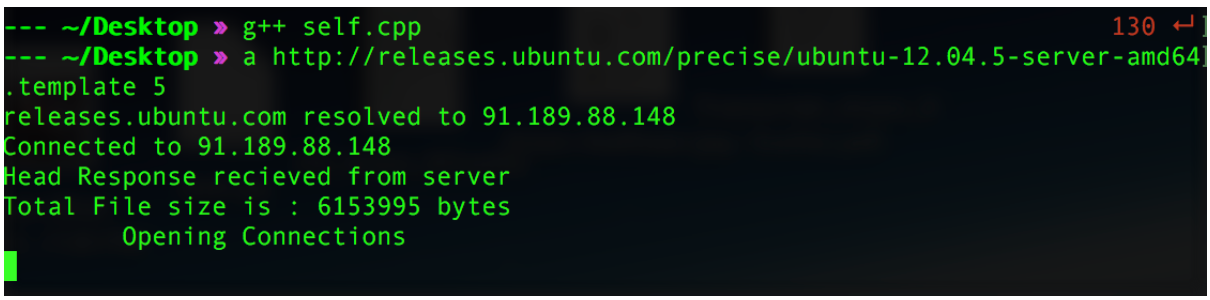
Separate functions are used to strip the given inputted URL and retrieve the required hostname and file path.

Initially a HEAD request is made to the server in order to get the size of the file from the server. The response received from the server is then parsed to get the file size.

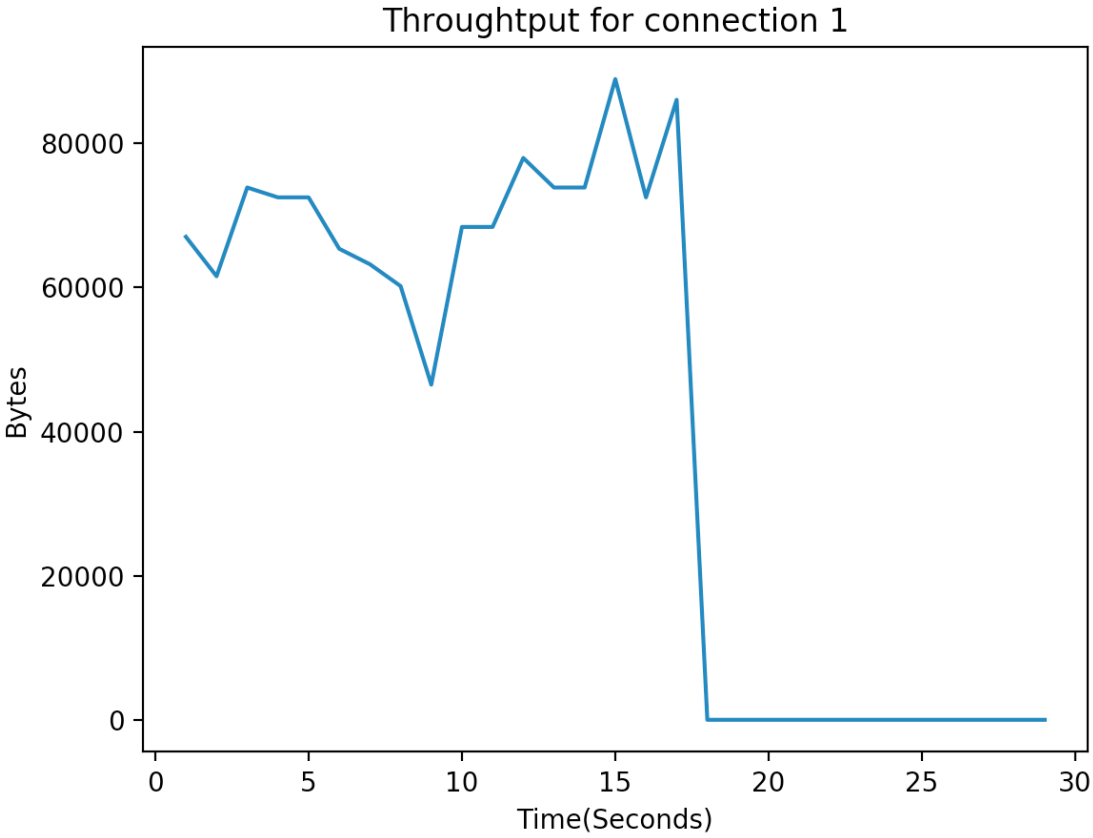
This file size is broken down into the number of connections so that each connection downloads (size/number) part of the file. Using HTTP range requests we can download only a particular part of the file to be downloaded from the server. Thus each connection only downloads a certain part/range of the file. Each part is then stitched into the final binary file.

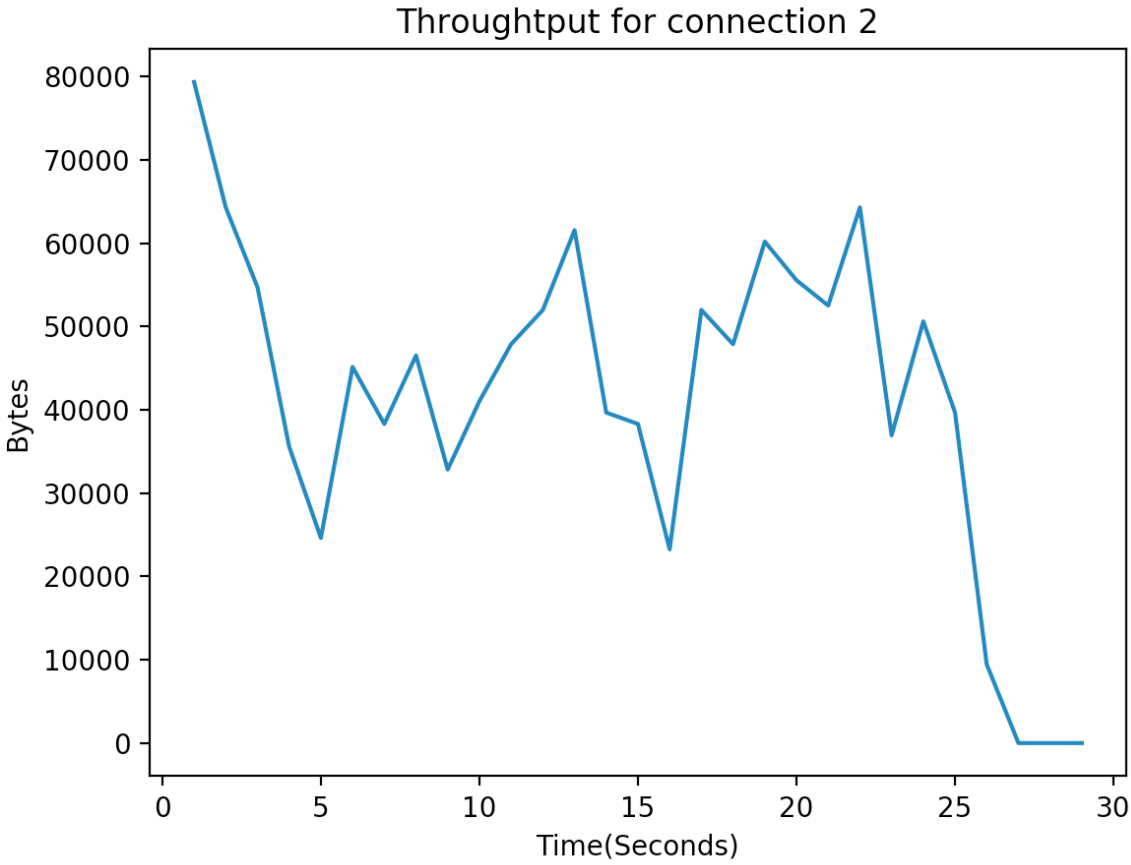
We can easily check that no data has been corrupted by using an MD5 hashing descriptor to check the file downloaded by my own code and the file downloaded from a browser.

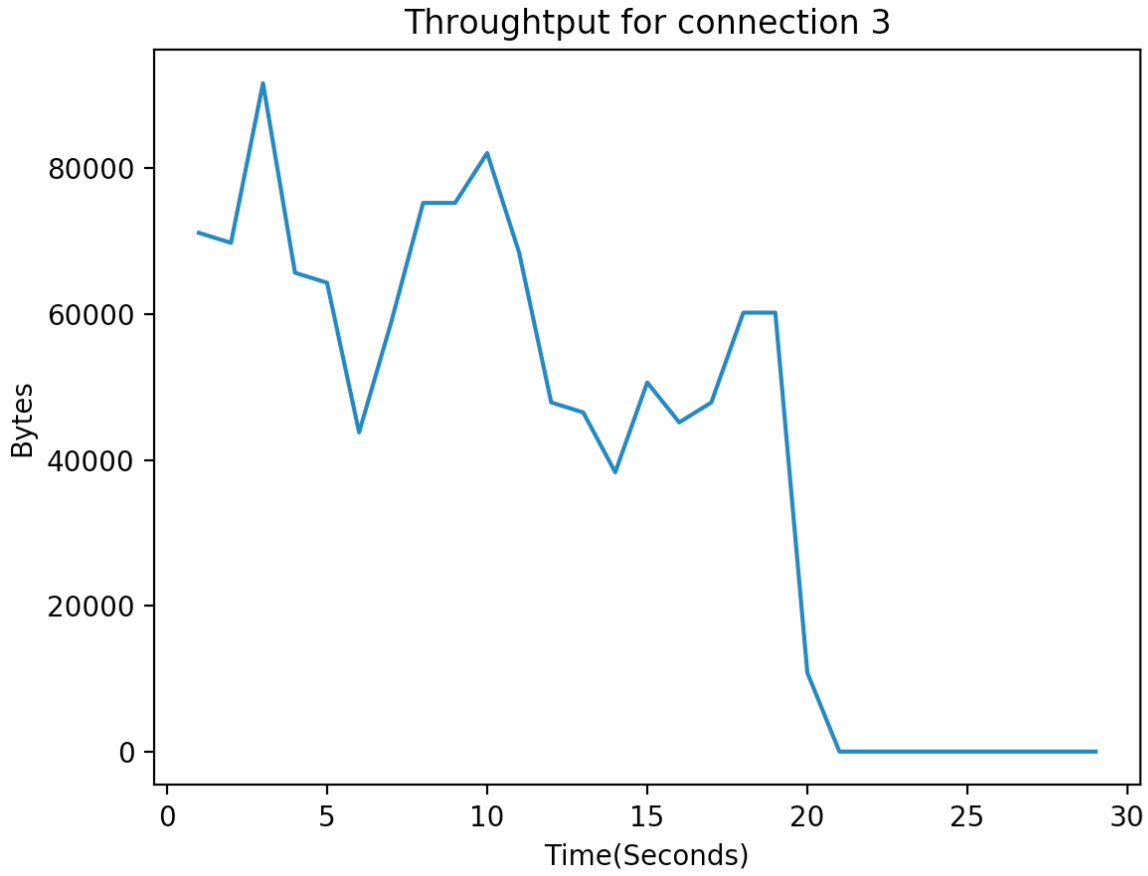
A sample from the running terminal is :

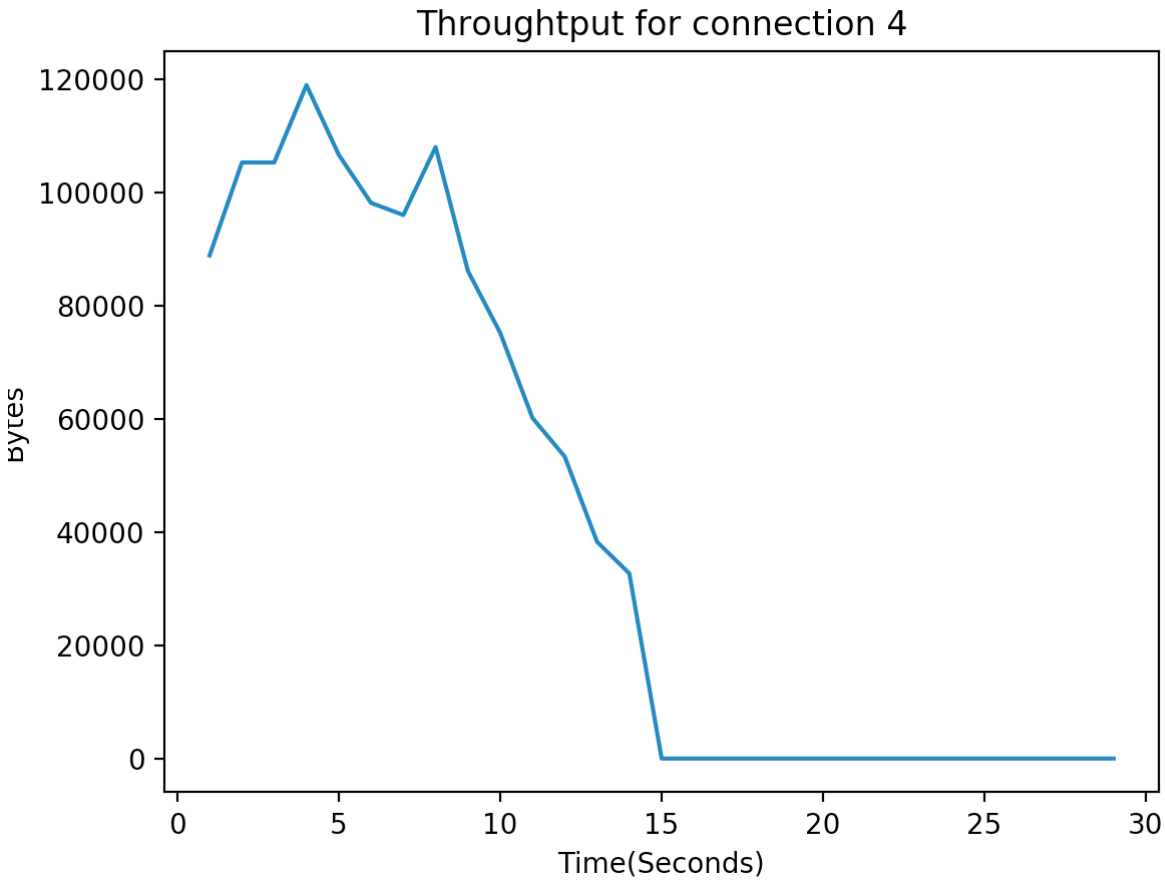


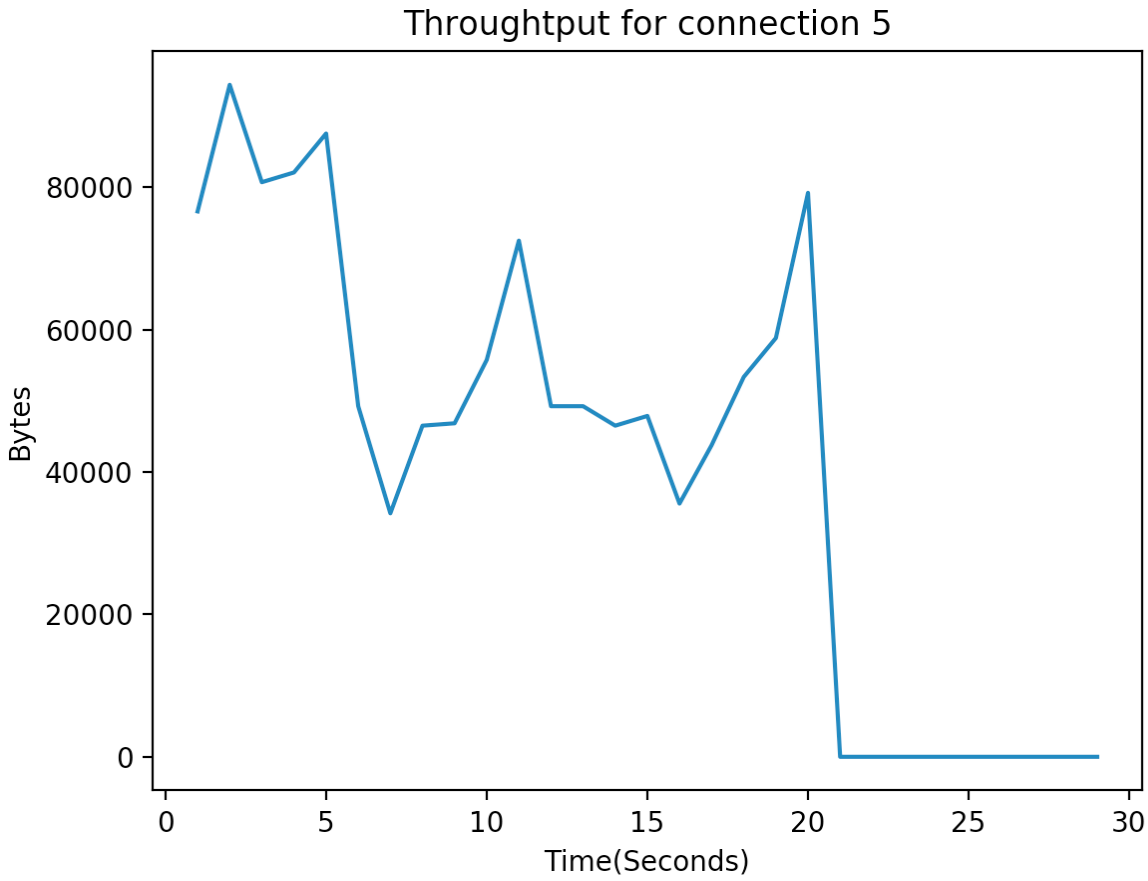
Let us look at the throughput of each connection for a file size of around 6 Mb with 5 parallel TCP Connections.



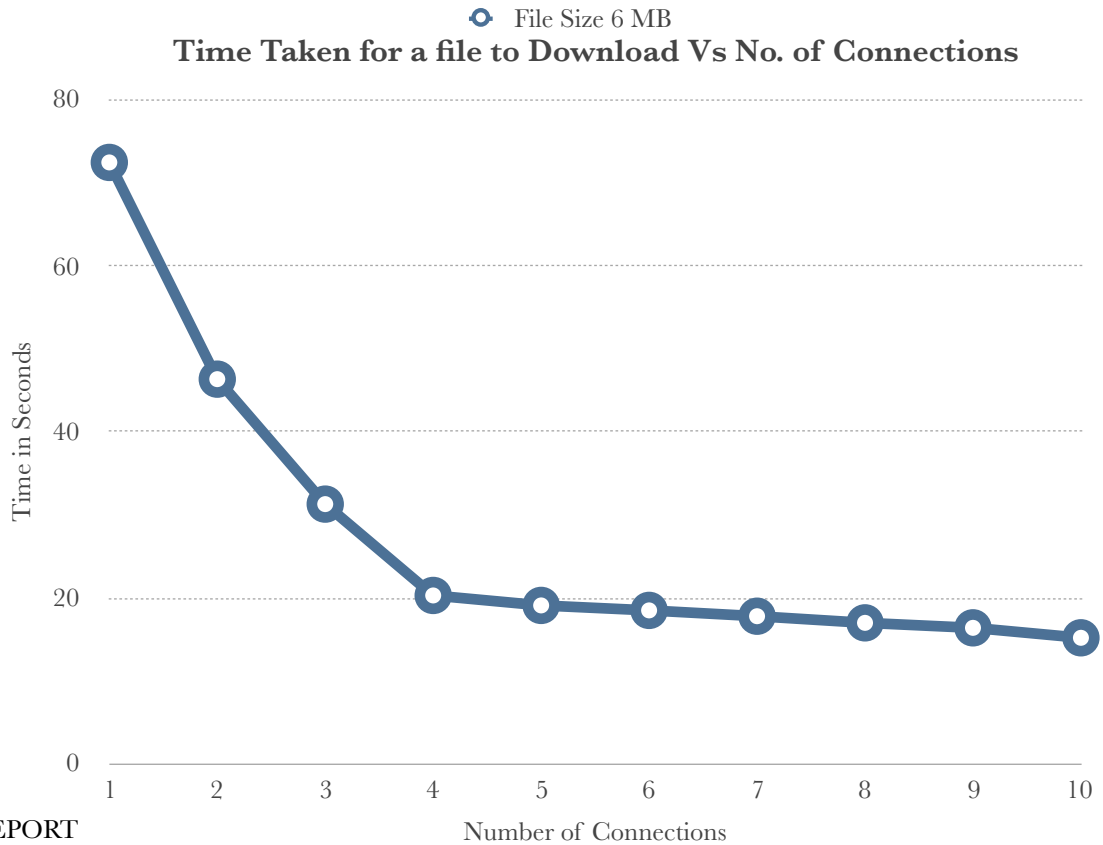



Throughtput for connection 3
Bytes
80000
60000
40000
20000
0
0
5
10
15
20
25
30
Time(Seconds)

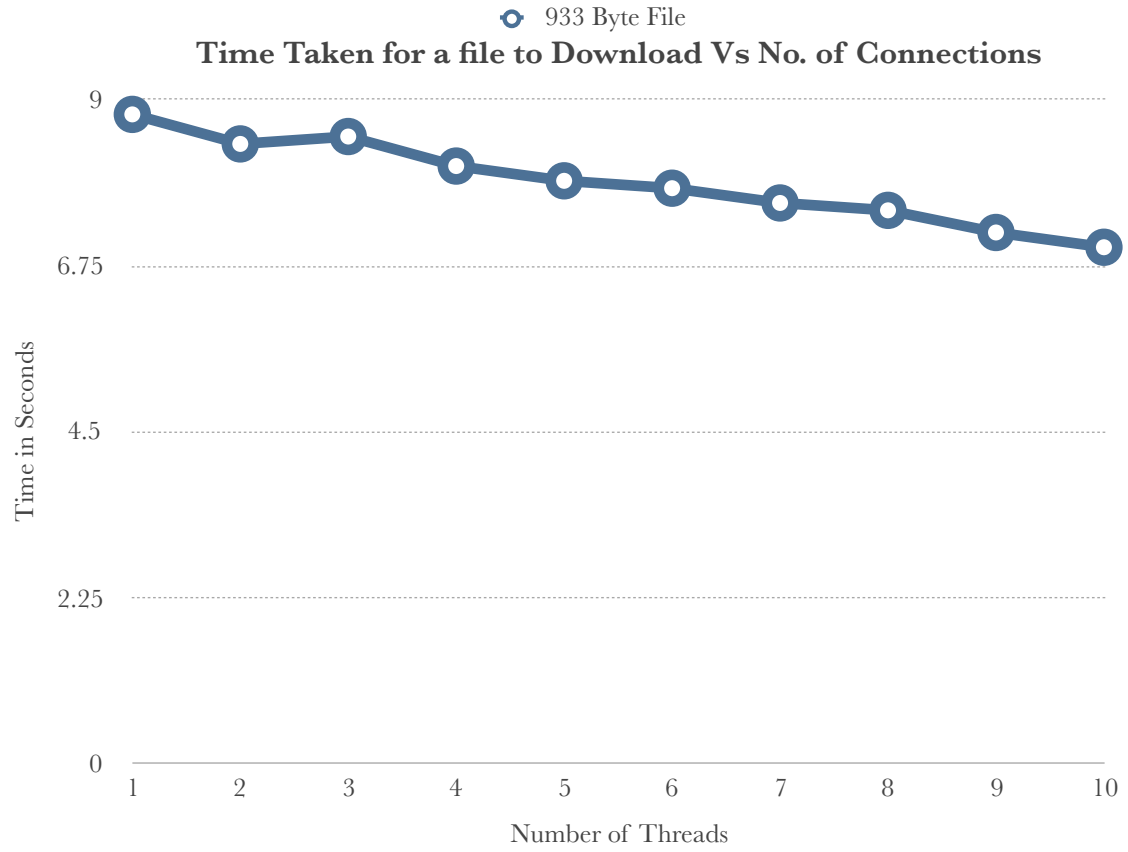

Throughtput for connection 4
Bytes
120000
100000
80000
60000
40000
20000
0
0
5
10
15
20
25
30
Time(Seconds)



The above graphs show us the throughput for each connection. The erratic behaviour we observe is due to various priorities given to certain threads/processes and the internet speed which varying quite a lot some times.



The above graph is for a reasonably big file of around 6 Mb size. We can easily see that as the number of connections go on increasing, the total time to download the file goes on deceasing.



Here we see that making multiple connections is not very helpful.

Hence multiple connections are much better to download large file.

With smaller files, overhead also adds up, reducing efficiency.