Internet Behavior at Different Times

Ethan Nelson, Jiacheng Qiao

Team

Ethan Nelson
Jiacheng Qiao

Objective

The goal of this project is to paint a clear picture of the world wide web through its latency and connection paths to different sites around the world.
The objectives of this research come fivefold.

1. Comparing RTT (Round Trip Time, otherwise known as latency) with geographical distance from websites all over the world.
2. Show patterns of RTT throughout an average day in reference to websites spread across the globe.
3. Show patterns of average RTT throughout a given week in reference to geographically close and distant websites
4. Show the patterns of hop distance for each website
5. Discover the patterns of unresponsiveness for each website

Literature Survey

A Tutorial On Network Latency and its Measurements
Minseok Kwon. (2015). A Tutorial On Network Latency and its Measurements.

Within this paper contains a list of useful terms to become familiar with, many of which are used frequently in this paper. A few of them are: network latency, round-trip time (RTT), and hop (pg 19-20, Kwon).

Challenges Faced & Subsequent Solutions

1. When searching for geographically diverse websites to conduct the experiment, many websites had servers hosting their website within the US (using a CDN - Content Distribution Network). This was difficult in regards to conducting the experiment but very useful in finding a solution to the distance related latency. Unfortunately, the only solution that remained was searching for other websites that had our desired server location.
2. The other challenge that we faced involved the length of the routes to each website. On occasion, servers on route to the website did not respond to the probe from traceroute, limiting the amount of information recorded. In addition, the length of the route exceeded what is deemed acceptable. While there is no real solution to this problem, we increased

the limit of hops from the default number of 30 to 60 to reduce the number of routes that did not complete in the original 30 hop limit.

3. Another small issue we originally did not think of expecting was the MTU (maximum transmission unit). When testing out different packet sizes for the experiment, any packet size over 1500 bits resulted in no response from the site. Due to this, we limited the range of packet sizes from 100 - 1200 bits.

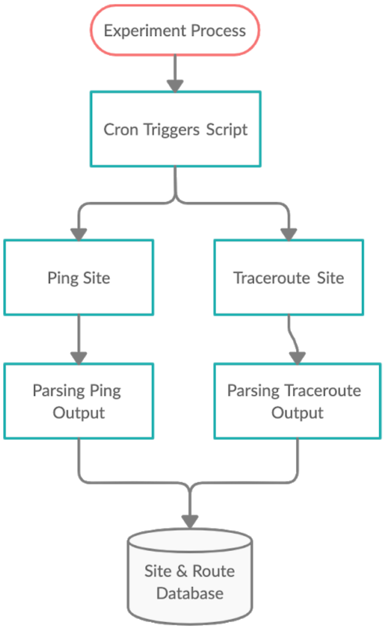
Experiment Set-Up

The main experiment that took place was the utilization of the tools Traceroute and Ping to record RTT (round time trip), path (as well as RTT to each server hop) to each website.

Tools Used:

- Traceroute
- Ping
- PingPlotter
- SQLite
- Bash
- Cron
- Raspberry Pi 3B+

Experiment Diagram



1. 10 websites were selected across the globe to give a variety of data to the experiment. The websites that were chosen include:

1. Cam.ac.uk (Cambridge, UK)

2. Politico.eu (Brussels, Belgium)
3. Yahoo.co.jp (Tokyo, Japan)
4. unimelb.edu.au (Melbourne, AU)
5. Uni.edu (Cedar Falls, IA, US)
6. Apple.com (Cupertino, CA, US)
7. hawaii.gov (Honolulu, HI, US)
8. www.msu.ru (Moscow, Russia)
9. www5.usp.br (San Paulo, Brazil)
10. www.tsinghua.edu.cn (Bejing, China)

    UPDATE:
    Due to a very unfortunate overlooked issue of updating the source code, the sites that were recorded over a period of a week were an old compiled list of websites with much less geographic diversity:

    1. Cam.co.uk (Cambridge, UK)
    2. Politico.eu (Brussels, Belgium)
    3. Yahoo.co.jp (Tokyo, Japan)
    4. Unimelb.edu.au (Sydney, AU)
    5. Uni.edu (Cedar Falls, IA, US)
    6. Ucla.edu (LA, CA, US)
    7. Apple.com (Cupertino, CA, US)
    8. Amazon.com (Seattle, WA, US
    9. IBM.com (Research Triangle Park, NC, US)
    10. Yahoo.com (New York, NY, US)

2. We chose a range of packet sizes to send to each website: 100 bits, 300 bits, 600 bits, 900 bits, and 1200 bits.
NOTE: We specifically did not choose higher numbers since the MTU (maximum transmission unit) is limited to 1500 bits over ethernet.

3. We built a bash script to execute Ping & Traceroute, parse the output, and store it in an SQLite database. The settings used for Ping and Traceroute are as follows:

a. Ping -c10 -q -s(packetSize) (site)
~ c10: pinged the website 10 times,
~ q: quiet output,
~ s: specifies packet size

b. Traceroute -n -q1 -I -w1 -m60
~ n: only returns IP names (ex: 192.168.1.1 not google.com),
~ q: set the number of probes to 1,

~ I: sets the packet type as ICMP,
~ w1: set maximum response time wait to 1 second,
~ m60: set the maximum amount of hops to 60

a. The schema for the database is the following:
Site(date, time, siteName, packetSize, hops, RTT)
Route(date, time, siteName, packetSize, hopNum, hopServer, hopRTT)

4. The bash script is managed by Cron, set to execute every hour over the period of 7 days.

5. The set of tools and software can be used on any up to date Linux computer system. We chose to use a Raspberry Pi 3B+ to conduct our experiment.

Experiments Planned

There was only one experiment planned. It entails recording the response time from a list of 10 websites, and the corresponding route to each website. From recording the combination of this information, the pattern of response time, routes to each site, and relation to geographical distance can become more apparent.

The first half of the data that was collected included 8,286 entries of response times over the course of a week. Each entry is composed of the time of connection, website name (in IPv4), the packet size used, the number of hops to the site, and the response time to the site.

The second half of the data that was collected included 141,023 entries composed of the broken up paths to each site. Each entry contains the time of connection, the final site attempting to be reached, the packet size used, the specific hop number in the path, the IP address of the server being 'hopped' to, and the response time of the server being 'hopped' to.

The experiment was conducted over a 7 day period, with data being recorded every hour.

Results

Figure 1

Line graph depicting average latency at each hour for each website location

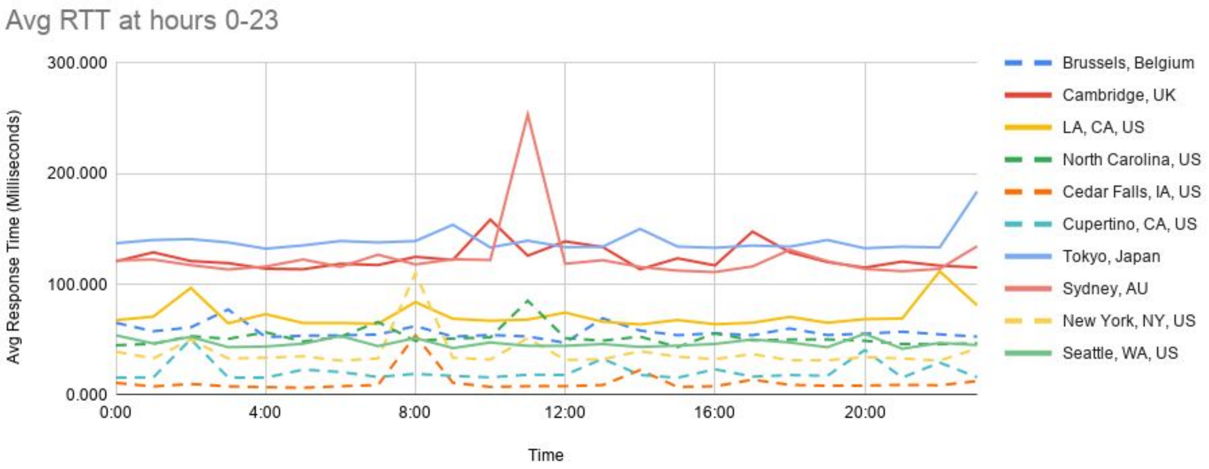
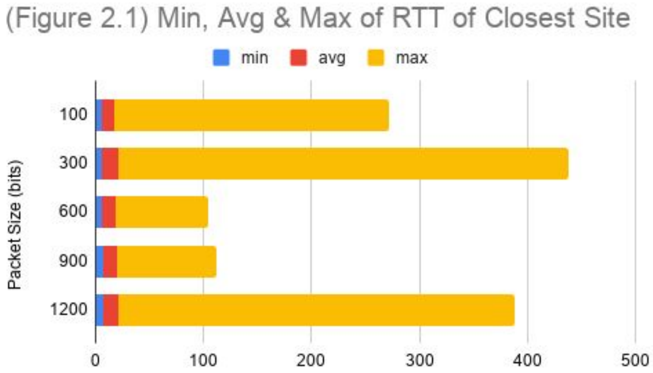


Figure 2.1 & 2.2

Stacked bar charts of latencies with different packet sizes for the closest & furthest site

(The closest site is in Cedar Falls, IA, US - Uni.edu)

(The furthest site is in Sydney, AU - unimelb.edu.au)



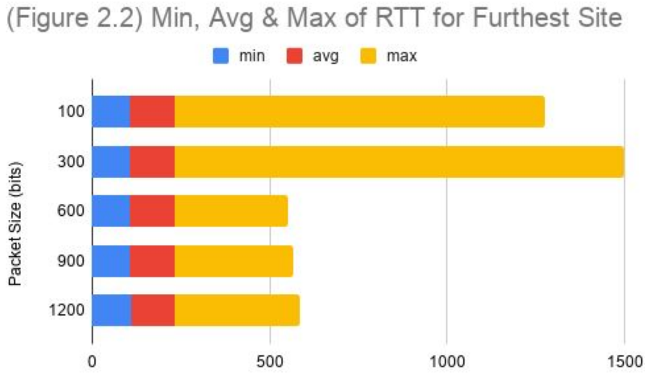


Figure 3
Line graph of daily average latencies of each website location

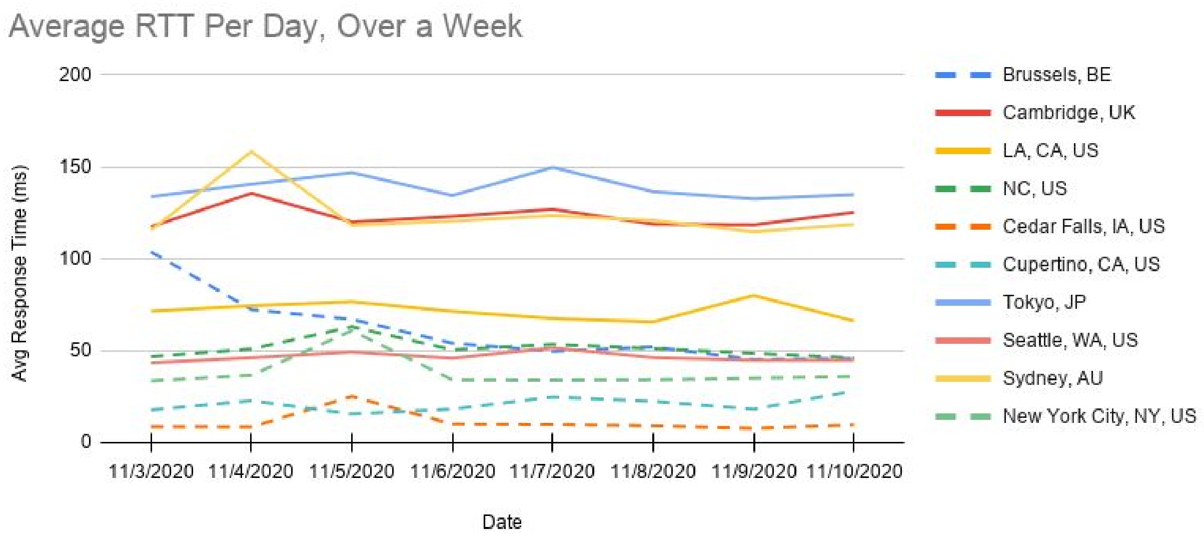


Figure 4
Bar chart of hop minimum, average, and maximum hop distances for each website

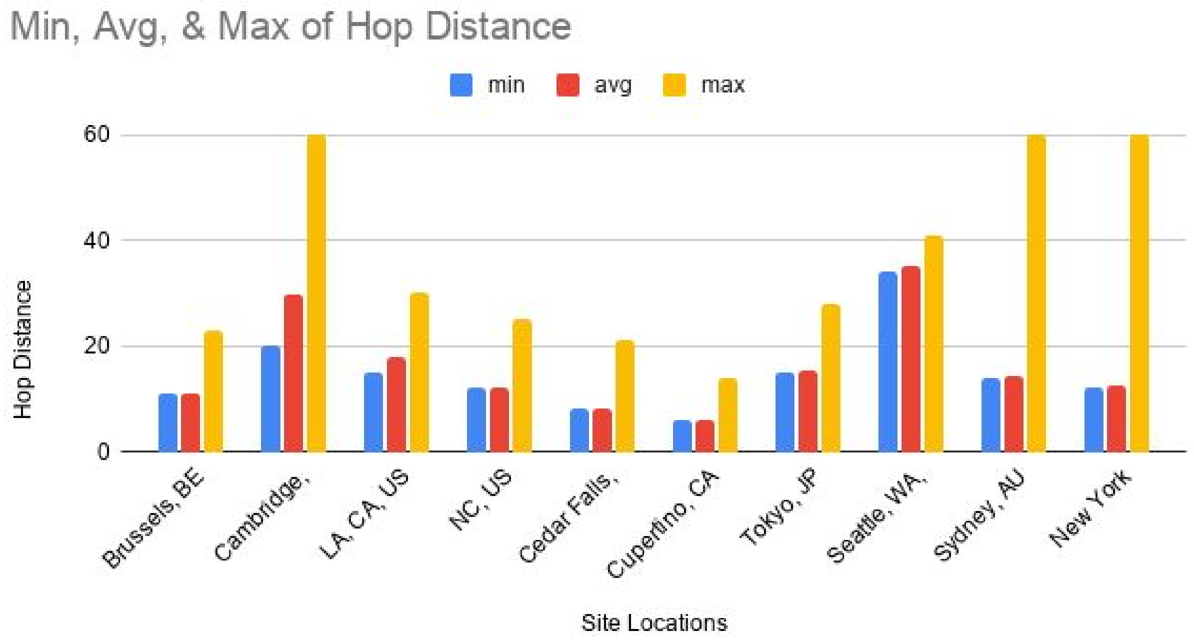
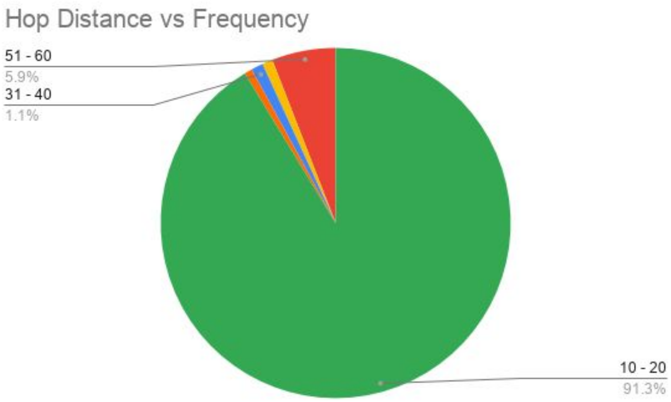


Figure 5
Pie chart depicting percentage of routes that had routes reaching 60 hops
(only includes 3 sites that successfully reached 60 hops | In total, the percentage of routes that reached 50-60 is less than 6%)
(Sites included: Cambridge, UK | Sydney, AU | New York, NY, US)



Discussion

From the results of this experiment, a couple of key ideas can be taken. For one, the difference in packet sizes had no considerable effect on the latency between each website. We hypothesis that the amount of data is simply not large enough to result in a higher latency. Secondly, the time at which a connection to a website is established does not seem to make a difference in latency as well. This went against our original hypothesis that a website would be busier during that website's peak times. However, the latency does not show those results. Lastly, as the distance between site locations grew, so did the latency. In addition, the hop distance also grew with geographical distance.

The main question is, how does one reduce response time for a website of a rather far geographical distance? We accidentally discovered an example of a solution to this problem when vetting websites to experiment. The website 'politico.eu' has a registered location of Brussels, Belgium. Yet, when looking at the servers that host this website, something unexpected occurred. The website appeared to be located within the US, specifically California. This website uses a CDN (Content Distribution Network) to allow for quicker access within the US. Instead of having one website hosted in one location, this website is hosted on multiple servers in many different locations. This ultimately resulted in partly skewing the results from the experiment. However, it was the only website we found that took part in a CDN, so we decided to leave it in the experiment.

Possible Extensions

Since the main aspect of this experiment was to find correlations in latency and geographic distances, another important measurement that could be included would be the amount of dropped packets for each attempted connection. Additionally, translating the IPv4 hop addresses to geographical locations would give more insight into where data goes on the way to a website.

Conclusion

With the addition of Ping and Traceroute tools, key information to describe a network can be collected. Over the course of a week, we've found that packet size doesn't considerably affect latency. Within the list of 10 websites, no peak usage times were found to be reflected within latency. However, as distance grew regarding physical website locations, so did the latency. This convinces us to believe that geographical distance is the main factor in response time. This points a star on the back physical distance. One way to shortcut this issue is to have a website hosted on many different servers located in many different locations, otherwise known as a CDN (Content Distribution Network).