Introduction

This document functions as a data description to the public domain data submitted as part of this intervention. It includes a brief explanation of the Measurement Lab, NDT tool, the variables collected during a test and an explanation how the sample was collected.

Measurement Lab Consortium (M-Lab)

The Measurement Lab is an international consortium of companies, researchers and Internet organizations dedicated to creating a global platform to study the Internet. The M-Lab includes open standards to create accurate measurements of the Internet, software tools to perform these measurements and an international infrastructure to these tools as well as data collected. The M-Lab present hosts 14 projects performing tests against over 130 servers located worldwide.[1]

Methodology

All data comes from tests generated using the Network Diagnostic Tool (NDT).[2] NDT is an active, client-side Internet measurement application. Active communicates with the network to specifically measure performance (as opposed to passive Internet measurement that collects data during the everyday network operations) [3]. Client side refers to the test needed to be initiated and run by a local computer, e.g. a home computer, rather than initiated by a server. Once a user initiates NDT, it opens a single TCP connection to measure the upload and download performance.[4] Figure 1 depicts the activity during an NDT test. The single-session attempts to transfers as much data as possible from client to server and client to server in 10 seconds. During this test, NDT also measures the round-trip time delay, congestion windows and packets dropped.

[1] For more details about the Measurement Lab Consortium, see: http://www.measurementlab.net/about

[2] For more technical details about NDT, see: http://software.internet2.edu/ndt/ & https://code.google.com/p/ndt/

[3] M. Murray and kc claffy, “Measuring the Immeasurable: Global Internet Measurement Infrastructure,” in In PAM – A workshop on Passive and Active Measurements, 2001, pp. 159–167.

[4] The single session approach distinguishes NDT from the popular Ookla Speedtest. For a longer discussion, see: http://cira.ca/blog/how-does-ca-internet-performance-test-compare-speedtest. For an extended discussion of different Internet measurement, see: Bauer, S., Clark, D. D., & Lehr, W. H. (2010). Understanding Broadband Speed Measurements. Boston: Massachusetts Institute of Technology. Retrieved from http://mitas.csail.mit.edu/papers/Bauer_Clark_Lehr_Broadband_Speed_Measurements.pdf

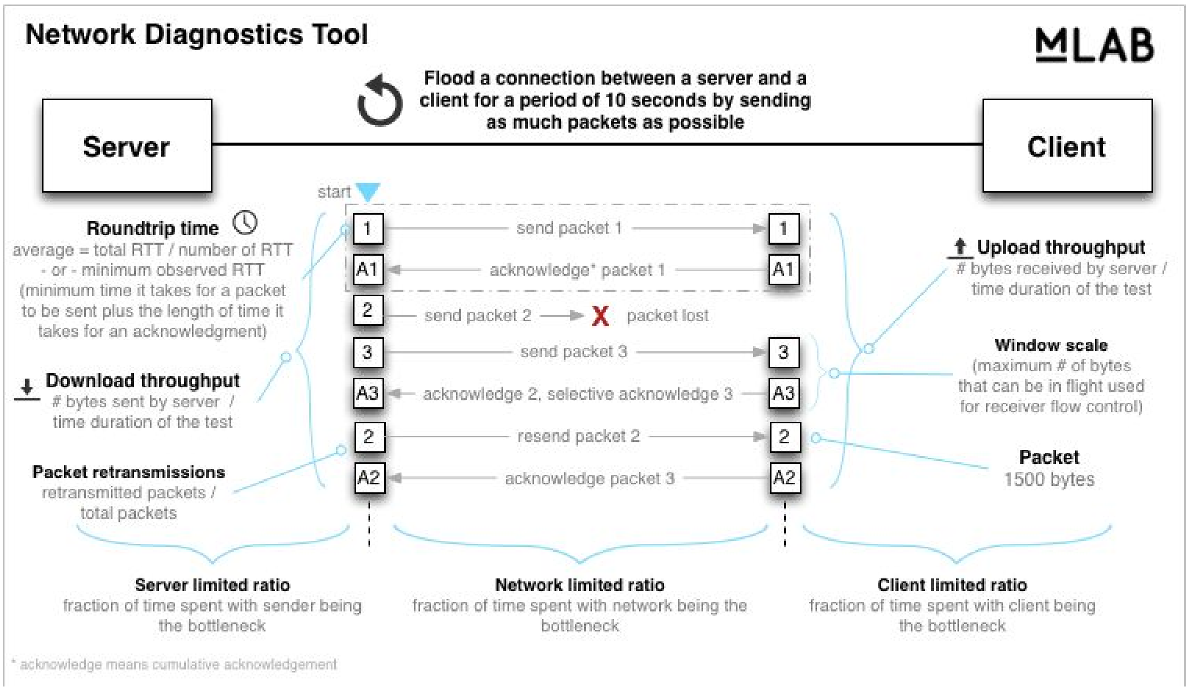
Network Diagnostics Tool
MLAB
Flood a connection between a server and a client for a period of 10 seconds by sending as much packets as possible
Server
Client
start
Roundtrip time
average = total RTT / number of RTT - or - minimum observed RTT (minimum time it takes for a packet to be sent plus the length of time it takes for an acknowledgment)
Download throughput
# bytes sent by server / time duration of the test
Packet retransmissions
retransmitted packets / total packets
1
send packet 1
A1
acknowledge* packet 1
2
send packet 2
X
packet lost
3
send packet 3
A3
acknowledge 2, selective acknowledge 3
2
resend packet 2
A2
acknowledge packet 3
Upload throughput
# bytes received by server / time duration of the test
Window scale
(maximum # of bytes that can be in flight used for receiver flow control)
Packet
1500 bytes
Server limited ratio
fraction of time spent with sender being the bottleneck
Network limited ratio
fraction of time spent with network being the bottleneck
Client limited ratio
fraction of time spent with client being the bottleneck
* acknowledge means cumulative acknowledgement

Figure 1 - The NDT Test - Source: Measurement Lab from https://github.com/ndt-project/ndt/wiki

Sample

The following section explains the development of the data sample. The sample was constructed in two phrases. First, a series of queries extracted NDT tests in Canada from the larger public Measurement Lab database. The queries resulted in itemized data of each test conducted in Canada in 2014. The second phase aggregated itemized data by GIS units used by Industry Canada and Statistics Canada. In total, the sample comprises 1,006,250 tests for upload speed and 717,788 tests for download speed.

Phase 1: Measurement Lab Query

The sample selected all data points from NDT tests (labelled as project 0 in the data sample) originating in Canada during 2014. Both upload and download queries include some data validation to remove incomplete tests and test missing key data points. The queries remove any test that lasted less than 9 seconds[2] and more than 60 minutes.[3] Each test also has to have exchanged a minimum of 8192 bytes for inclusion in the sample.[4]

Google BigQuery facilitated access to the total public Measurement Lab data set. The large sample size required dividing the data requested into four queries. Queries requested related to download performance and upload performance then year. The four queries included in the Appendix are:

- upload performance for 2014 from January to July,
- upload performance for 2014 from August to December,
- download performance for 2014 from January to July, and
- download performance for 2014 from August to December.

An expert in the Measurement Lab data, Christopher Ritzo at the Open Technology Institute at the Washington, DC think tank, New America, wrote all four queries.

The queries resulted in four CSVs contained in Item 2 of the intervention. Upon review of the data, congestion metrics overall and latency tests in the upload results proved inconsistent. The submitted CSVs have removed these variables.

Phase 2: Geographic Analysis

The Government of Canada employs a number of geographic units to locate statistical information. The intervention selected the dissemination areas used by Statistics Canada and the hexagons developed by Industry Canada as part of the Connecting Canadians 150 program. Statistics Canada has a large catalog of GIS units and the intervention elected to use the Census Dissemination Areas[5]. Each area contains roughly 400 to 700 people.

The intervention relied on the free software QGIS to locate tests within these systems. QGIS parsed itemized data and mapped the geo-located latitude and longitude per test. The join by location function in QGIS aggregated tests by hexagon and dissemination area. The submitted map includes layers for both itemized data, aggregated data, Industry Canada hexagons and Statistics Canada dissemination blocks.

Itemized Data Variables

NDT variables may be distinguished between upload and download performance. The following section details the information collected per test for both upload and download performance. The section defines first names each variable means in plain language, its name in the data and a brief description.[6]

Identification Variables

Time of test (day_timestamp)

The date and time of the test.

Local address (web100_log_entry.connection_spec.local_ip)

The IP address of the computer or client that initiated the test.

Server address (web100_log_entry.connection_spec.remote_ip)

The IP address of the Measurement Lab testing server that facilitated the test.

Local computer name (connection_spec.client_hostname)

An optional field if the client’s computer has a name on the network.

[5] Details about dissemination areas is available at: https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm

[6] A longer description of the data variables can be found at: https://cloud.google.com/bigquery/docs/dataset-mlab

Geo-location Variables

NDT appends geographic information to each test using the MaxMind[7] database that matches IP address to country and city as well as their approximate latitude and longitude.

These variables are:

Name of country (connection_spec.client_geolocation.country_name)

City name (connection_spec.client_geolocation.city)

Postal Code (connection_spec.client_geolocation.postal_code)

Latitude (connection_spec.client_geolocation.latitude)

Longitude (connection_spec.client_geolocation.longitude)

Download Performance

Download Speed (downloadThroughput)[8]

Measured in megabits per second. Download Speed is the best known Internet performance metric. It indicates how quickly a home computer can download data from a server. NDT measures download speed by counting the number of octets (or 8 bits) sent from server to client in a single TCP session. It calculates the average megabits sent per second.

Packet Loss (packetRetransmitRate)[9]

The ratio of packets retransmitted against the total number of packets sent during the download test. Packets might be lost during transmission due to congestion, weak signals or faculty hardware. While the metric does not detect the cause of the losses, a high packet loss rate indicates an overall poor network performance during the test.

Latency (avgRTT)[10]

Average time in milliseconds for a client to get a response from the server. Commonly known as round trip time, latency indicates the responsiveness of a network of client requests. High latency degrades user experience as network interactions might seem delayed or unresponsive.

Minimum Latency (web100_log_entry.snap.MinRTT)

[7] Maxmind is a leading database matching IP addresses to geographic

[8] Calculated in the query as 8 * web100_log_entry.snap.HCThruOctetsAcked/ (web100_log_entry.snap.SndLimTimeRwin + web100_log_entry.snap.SndLimTimeCwnd + web100_log_entry.snap.SndLimTimeSnd)

[9] Calculated in the query as web100_log_entry.snap.SegsRetrans/web100_log_entry.snap.DataSegsOut

[10] Calculated in the query as web100_log_entry.snap.SumRTT/web100_log_entry.snap.CountRTT

Lowest round trip time in milliseconds observed during the test.

Upload Performance

Upload Performance (uploadThroughput)

Measured in megabits per second.. NDT measures upload speed by counting the number of octets (or 8 bits) sent from client to server in a single TCP session. It calculates the average megabits sent per second.

Aggregated Data Variables

The intervention includes aggregated data of the itemized test results. All results have been aggregated by upload and download speed and by Industry Canada's Hexagon GIS and Statistics Canada's Dissemination Areas. Each aggregated data sample includes the mean, median, maximum and minimum calculations for the itemized data. These aggregations were calculated using QGIS. The Industry Canada hexagons also include some additions labels related to the Connecting Canadians program.

Aggregated Download Variables

Column	Variable
meandownlo	Mean of Download Speed
mindownloa	Minimum of Download Speed
maxdownloa	Minimum of Download Speed
mediandown	Median of Download Speed
meanpacket	Mean of Packet Loss
minpacketR	Minimum of Packet Loss
maxpacketR	Minimum of Packet Loss
medianpack	Median of Packet Loss
meanavgRTT	Mean of Latency
minavgRTTn	Minimum of Latency
maxavgRTTn	Minimum of Latency
medianavgR	Median of Latency
count	Total number of tests in region

Aggregated Upload Variables

Column	Variable
meanupload	Mean of Upload Speed
minuploadT	Minimum of Upload Speed
maxuploadT	Minimum of Upload Speed
medianuplo	Median of Upload Speed
count	Total number of tests in region