Measuring Broadband New Zealand

Data Dictionary, Winter Report, August 2021

In 2018, the Commerce Commission appointed SamKnows to measure New Zealand's internet performance. The programme, called Measuring Broadband New Zealand, gives internet users in New Zealand access to the SamKnows platform to measure the quality of their fixed-line internet. The aim of the programme is to increase transparency about actual in-home broadband performance and provide consumers with independent information about internet performance across different providers, plans, and technologies, to help them choose the best broadband for their homes. It will also encourage providers to improve and compete on their performance. This report provides an overview of the findings from data collected between 1st May and 31st May 2021.

Raw Data

Alongside the report ComCom also release the raw data and summary unit information used to produce the Winter 2021 report.

Two levels of data are included in this publication: **Raw Measurement Data** and **Per-Whitebox Summary Data**. More information on what is included in these files is outlined below.

Raw Measurement Data

This is the measurement data in its raw, unaggregated form.

The raw data is available in the '/raw_data' directory, and a data dictionary describing the fields is included later in this document.

Per-Whitebox Summary Data

The measurements in the raw data are aggregated by Whitebox ID (also known as unit_id) as part of the data analysis process. The per-Whitebox data is far smaller, and therefore more accessible to third parties, than the raw data. It also includes additional derived fields which are used later in the analysis.

This summary data is calculated from the raw data using the statistical analysis tool R. The eventual aim is to release the R script used to create the charts along with the raw data so that interested parties can recreate the results. Due to the fact that certain metadata fields are excluded in the raw data, results in the report cannot be fully replicated with this data release.

A data dictionary describing the fields is included later in this document.

File listing

File	Description
./raw_data	
raw_download_tests.csv	Download speed test data
raw_upload_tests.csv	Upload speed test data
raw_webpage_tests.csv	Webpage loading test data
raw_youtube_tests.csv	YouTube test data
raw_outage_tests.csv	Outage test data
raw_latency_tests.csv	Latency and packet loss data
raw_netflix_tests.csv	Netflix data
raw_gaming_tests.csv	Gaming data excluding Fortnite
raw_fortnite_tests.csv	Fortnite data
raw_social_media_tests.csv	Social Media data
raw_video_conferencing_tests.csv	Video conferencing data
./output	
chart_data.csv	Data behind the graphs which appear in the Winter report
unit_summary_statistics_download_upload_latency.csv	One line per Whitebox with download, upload and latency results per target country
unit_summary_statistics_gaming.csv	One line per Whitebox per game
unit_summary_statistics_netflix.csv	One line per Whitebox with Netflix results
unit_summary_statistics_social_media.csv	One line per Whitebox per service per media type
unit_summary_statistics_video_conferencing.csv	One line per Whitebox per video conferencing service
unit_summary_statistics_webpage.csv	One line per Whitebox per webpage
unit_summary_statistics_youtube.csv	One line per Whitebox
unit_summary_statistics_outage.csv	One line per Whitebox per target country

Raw Data

raw_download_tests.csv (Download speed)

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
download_mbps	decimal	Test speed in Mbps.
did_test_complete_successfully	boolean	Did the speed test complete successfully?
target_server_country	string	The country in which the test server is located.

raw_upload_tests.csv (Upload speed)

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
upload_mbps	decimal	Test speed in Mbps.
did_test_complete_successfully	boolean	Did the speed test complete successfully?
target_server_country	string	The country in which the test server is located.

raw_latency_tests.csv (Latency & Packet Loss)

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
latency_ms	decimal	The time for a round trip from Whitebox -> Server -> Whitebox.
num_successes	int	Number of packets which made a successful round trip.
num_failures	int	Number of packets which failed to make a round trip.
packet_loss_pct	decimal	Ratio of packets which did not make a successful round trip: failures divided by (successes + failures).
target_server_country	string	The country in which the test server is located.

Raw Data

raw_outage_tests.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
duration_sec	decimal	Duration in seconds of the disconnection
target_server_country	string	The country in which the test server is located.

raw_netflix_tests.csv (Netflix)

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the server assigned by Netflix to stream content.
bitrate_mbps	decimal	The bitrate that can be reliably streamed without stalls (in Mbps).
video_quality	string	Either SD, HD or UHD quality streamed relating to bitrate.
download_mbps	decimal	The download speed when downloading content from Netflix (in Mbps).
latency_ms	decimal	The time taken to establish a TCP connection with Netflix (in milliseconds). Used as a proxy for the latency between Whitebox and Netflix server.
did_test_encounter_stall	boolean	Did the test encounter a stall event?
did_test_complete_successfully	boolean	Did the test complete successfully?
is_during_reporting_period	boolean	Is the test during the reporting period?

Raw Data

raw_fortnite_tests.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
region	string	The region of the Fortnite server, as defined by the game author
datacenter	string	The datacenter code (e.g. “SYD") of the Fortnite server, as defined by the game author
selected_server	string	The IP address of the server that we measured against
hop_count	int	The number of hops to the server
latency_ms	decimal	The time for a round trip from Whitebox -> Server -> Whitebox
did_test_complete_successfully	boolean	TRUE if any packets made a successful round trip, FALSE if not.
num_successes	int	Number of packets which made a successful round trip.
num_failures	int	Number of packets which failed to make a successful round trip.

raw_gaming_tests.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
target	string	Target hostname or IP address
ping_latency_ms	decimal	The time for a round trip from Whitebox -> Server -> Whitebox
num_successes	int	The number of pings successful sent and received.
num_failures	int	The number of pings which failed.
packet_loss_pct	decimal	The proportion of pings which failed.
target_server_grouping	string	The gaming service being tested (e.g League of Legends)
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
is_during_reporting_period	boolean	Is the test during the reporting period?

Raw Data

raw_video_conferencing_tests.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
service	string	The video conferencing provider being tested.
region	string	The region of the server, as defined by the service provider.
latency_ms	decimal	The time taken to establish a connection with video conferencing services (in milliseconds).
did_test_complete_successfully	boolean	TRUE if any packets made a successful round trip, FALSE if not.
num_successes	int	Number of packets which made a successful round trip.
num_failures	int	Number of packets which failed to make a successful round trip.

raw_youtube_tests.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
video_hostname	string	The hostname of the YouTube server which hosts the video.
video_bitrate_mbps	decimal	The bitrate at which the video content was encoded.
video_quality	string	Either SD, HD or UHD, YouTube's classification of video quality.
download_mbps	decimal	The download speed when downloading content from YouTube (in Mbps).
latency_ms	decimal	The time taken to establish a TCP connection with YouTube's video server.
did_test_encouter_stall	boolean	Did the test encounter a stall event?
did_test_complete_successfully	boolean	TRUE if any packets made a successful round trip, FALSE if not.

Raw Data

raw_webpage_tests.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	The URL of the target webpage.
fetch_time_sec	decimal	The time taken to fetch the target webpage and all associated objects.
time_to_first_byte_sec	decimal	The latency between starting the test and receiving the first byte of data from the web server.

raw_social_media_tests.csv (Social Media)

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
service	string	Currently one of Facebook-app, Facebook-messenger, Instagram-app, Instagram-messenger, Snapchat, Whatsapp, Twitter.
media	string	Currently one of Image, Text, Video or Audio.
direction	string	One of "Downlink" or "Uplink". Indicates whether we're measuring receiving or uploading content.
hop_count	int	The number of hops to the server.
latency_ms	decimal	The time for a round trip from Whitebox -> Server -> Whitebox
did_test_complete_successfully	boolean	TRUE if any packets made a successful round trip, FALSE if not.
num_successes	int	Number of packets which made a successful round trip.
num_failures	int	Number of packets which failed to make a successful round trip.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
is_during_reporting_period	boolean	Is the test during the reporting period?

Summary Data

unit_summary_statistics_download_upload_latency.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
target_server_country	string	The country in which the test server is located.
trimmed_mean_download_mbps_24h	decimal	The 1% trimmed mean (average of the middle 98% of data) of download_mbps. Results where download_samples_24h is less than 5 are removed from the final dataset.
trimmed_mean_download_mbps_peak	decimal	The 1% trimmed mean (average of the middle 98% of data) of download_mbps - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE. Results where download_samples_peak is less than 5 are removed from the final dataset.
download_samples_24h	int	The number of download tests (count of rows in raw_download_tests.csv).
download_samples_peak	int	The number of download tests (count of rows in raw_download_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.
trimmed_mean_upload_mbps_24h	decimal	The 1% trimmed mean (average of the middle 98% of data) of upload_mbps. Results where upload_samples_24h is less than 5 are removed from the final dataset.
trimmed_mean_upload_mbps_peak	decimal	The 1% trimmed mean (average of the middle 98% of data) of upload_mbps - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE. Results where upload_samples_peak is less than 5 are removed from the final dataset.
upload_samples_24h	int	The number of upload tests (count of rows in raw_upload_tests.csv).
upload_samples_peak	int	The number of upload tests (count of rows in raw_upload_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.
trimmed_mean_latency_ms_24h	decimal	The 1% trimmed mean (average of the middle 98% of data) of latency_ms. Results where latency_samples_24h is less than 5 are removed from the final dataset.
trimmed_mean_latency_ms_peak	decimal	The 1% trimmed mean (average of the middle 98% of data) of latency_ms - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE. Results where latency_samples_peak is less than 5 are removed from the final dataset.
latency_samples_24h	int	The number of latency tests (count of rows in raw_latency_tests.csv).
latency_samples_peak	int	The number of latency tests (count of rows in raw_latency_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.

Summary Data

unit_summary_statistics_netflix.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
netflix_uhd_fraction	decimal	The percentage of successful Netflix tests where netflix_bitrate_mbps was greater than 6 and successes was equal to 1.
netflix_hd_fraction	decimal	The percentage of successful Netflix tests where netflix_bitrate_mbps was between 2.35 and 6, and successes was equal to 1.
netflix_sd_fraction	decimal	The percentage of successful Netflix tests where netflix_bitrate_mbps was greater than less than 2.35, and successes was equal to 1.
mean_netflix_download_mbps	decimal	The (untrimmed) mean of netflix_download_mbps - results where there were less than 5 tests were excluded from the final dataset.
netflix_samples	int	The number of Netflix tests.
check_percentages_add_to_1	boolean	TRUE/FALSE to check that percentages add to 1, with a tolerance of 0.01
max_concurrent_uhd_streams	int	The greatest integer smaller than mean_netflix_download_mbps divided by 15.6 (i.e. floor(mean_netflix_download_mbps / 15.6). This field is used to as an estimate of the number of simultaneous users who could stream Netflix in UHD; because this field is estimated based on download speed rather than bitrate, the threshold is different to that used in netflix_uhd_fraction.

unit_summary_statistics_social_media.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
service	string	Currently one of Facebook-app, Facebook-messanger, Instagram-app, Instagram-messenger, Snapchat, Whatsapp, Twitter
media	string	Currently one of Image, Text, Video or Audio.
direction	string	One of "Downlink" or "Uplink". Indicates whether we're measuring receiving or uploading content.
trimmed_mean_latency_ms	decimal	The 1% trimmed mean (average of the middle 98% of data) of latency_ms. Results where social_media_samples is less than 5 are removed from the final data set.
median_hop_count	decimal	The median number of hops to the server
trimmed_mean_hop_count	decimal	The 1% trimmed mean (average of the middle 98% of data) of hop_count. Results where social_media_samples is less than 5 are removed from the final dataset.
social_media_samples	int	The number of successful tests.

unit_summary_statistics_gaming.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
game	string	The name of the game tested.
target	string	The target server tested against.
trimmed_mean_latency_ms	decimal	The 1% trimmed mean (average of the middle 98% of data) of latency_ms. Results where gaming_samples is less than 5 are removed from the final dataset.
gaming_samples	int	The number of successful tests.

Raw Data

unit_summary_statistics_outage.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
target_server_country	string	Country in which the test server is located.
latency_test_duration_hours	decimal	Duration in hours of the latency tests run.
n_outages	decimal	The total number of outages.
outages_per_hour_of_testing	decimal	The number of outages measured per hour.

unit_summary_statistics_video_conferencing.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
service	string	The name of the video conferencing tested.
trimmed_mean_latency_ms	decimal	The 1% trimmed mean (average of the middle 98% of data) of hop_count. Results where video_conferencing_samples is less than 5 are removed from the final dataset.
video_conferencing_samples	int	The number of successful tests.

unit_summary_webpage.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
target	string	The URL of the target webpage.
trimmed_mean_fetch_time_sec	decimal	The 1% trimmed mean (average of the middle 98% of data) of fetch_time_sec. Results where webpage_samples are fewer than 5 are removed from the final dataset.
trimmed_mean_time_to_first_byte_sec	decimal	The 1% trimmed mean (average of the middle 98% of data) of time_to_first_byte. Results where webpage_samples are fewer than 5 are removed from the final dataset.
webpage_samples_24h	int	The number of successful tests.

Raw Data

unit_summary_statistics_youtube.csv

Field Name	Type	Description
id	int	Unique identifier for an individual Whitebox.
youtube_uhd_fraction	decimal	The percentage of successful tests where UHD was streamed reliably.
youtube_hd_fraction	decimal	The percentage of successful tests where HD was streamed reliably.
youtube_sd_fraction	decimal	The percentage of successful tests where SD was streamed reliably.
mean_youtube_download_mbps	decimal	The (untrimmed) mean of youtube_download_mbps - results where there were less than 5 tests were excluded from the final dataset.
youtube_samples	int	The number of YouTube tests.