Measuring
Broadband
New Zealand

Measuring Broadband New Zealand

Raw data and data dictionary

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 SamKnows Whiteboxes 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. The first report provides an overview of the initial findings from the data collected during the early stages of the project.

Raw data

Alongside the report ComCom also release the raw data and summary units information used to produce the Spring report. The format of the raw data files has been updated for this report in order to make the information contained clearer and easier to use. The methodology behind how the raw data is provided has not changed.

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. However, only measurements that were used within the report have been included in this raw data package. Additionally, metadata fields which were not used in the report have been excluded (e.g. RSP name and product in specific instances).

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 (e.g. the fraction of YouTube videos that were delivered at HD or better quality).

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 (e.g. the advertised speeds for each RSP's Fibre Max product is not shared, as this could allow analysis on Fibre Max by RSP which is not currently reported on), the R script is not able to run.

The per-Whitebox summary data is available in the './output' directory, and 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_latency_tests.csv	Latency and packet loss data.
raw_netflix_tests.csv	Netflix data.
raw_youtube_tests.csv	YouTube data.
./output	
report_charts_public_tables.csv	Data behind the graphs which appear in the Spring report.
unit_stats_download_upload_latency_public_tables.csv	One line per Whitebox per target server country.
unit_stats_youtube_netflix_public_tables.csv	One line per Whitebox.

Data dictionary

raw_download_tests.csv (Download speed)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual unit.
dtime	datetime	The time of the test (local time).
ddate	date	The date of the test.
target	string	Hostname of the test server.
download_mbps	decimal	Test speed in Mbps.
successes	int	Number of successes (always 1 or 0 for this test).
failures	int	Number of failures (always 1 or 0 for this test).
target_server_country	string	The country in which the test server is located.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
is_to_server_in_nz	boolean	Is the test server located in New Zealand?
during_which_rwc_2019_game	string	The Rugby World Cup game which was being broadcast when the test was run. Blank if not during any game.

raw_upload_tests.csv (Upload speed)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual unit.
dtime	datetime	The time of the test (local time).
ddate	date	The date of the test.
target	string	Hostname of the test server.
upload_mbps	decimal	Test speed in Mbps.
successes	int	Number of successes (always 1 or 0 for this test).
failures	int	Number of failures (always 1 or 0 for this test).
target_server_country	string	The country in which the test server is located.
is_during_peak_hour	string	Is the test in peak hour (7-11pm Mon - Fri)?
is_to_server_in_nz	string	Is the test server located in New Zealand?
during_which_rwc_2019_game	string	The Rugby World Cup game which was being broadcast when the test was run. Blank if not during any game.

raw_latency_tests.csv (Latency & Packet Loss)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (local time).
ddate	date	The date of the test.
target	string	Hostname of the test server.
latency_ms	decimal	The time for a round trip from Whitebox -> Server -> Whitebox.
successes	int	Number of packets which made a successful round trip.
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.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
is_to_server_in_nz	boolean	Is the test server located in New Zealand?
during_which_rwc_2019_game	string	The Rugby World Cup game which was being broadcast when the test was run. Blank if not during any game.

Data dictionary

raw_netflix_tests.csv (Netflix)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (local time).
ddate	date	The date of the test.
target	string	Hostname of the server assigned by Netflix to stream content.
netflix_bitrate_mbps	decimal	The bitrate that can be reliably streamed without stalls (in Mbps).
netflix_download_mbps	decimal	The download speed when downloading content from Netflix (in Mbps).
netflix_latency_ms	decimal	The time taken to establish a connection with Netflix (in milliseconds). Used as a proxy for the latency between Whitebox and Netflix server.
stall_events	int	The number of times the test stalled at this bitrate.
successes	int	1 if the test runs for the full duration (may have stalls, though) i.e. it was not aborted and stepped down.
failures	int	1 if the test was aborted for some reason
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
is_to_server_in_nz	boolean	Is the test server located in New Zealand?
during_which_rwc_2019_game	string	The Rugby World Cup game which was being broadcast when the test was run. Blank if not during any game.

raw_youtube_tests.csv (YouTube)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	time	The time of the test (local time).
ddate	date	The date of the test.
video_hostname	int	Hostname of the YouTube server which served the video.
youtube_video_bitrate_mbps	decimal	The bitrate that can be reliably streamed without stalls (in Mbps).
youtube_download_mbps	decimal	The video download speed in mbps. Note that YouTube rate-limit server side.
youtube_latency_ms	decimal	The time taken to establish a connection with YouTube (in milliseconds). Used as a proxy for the latency between Whitebox and YouTube server.
stall_events	int	The number of times the test stalled at this bitrate.
successes	int	1 if the test runs for the full duration (may have stalls, though) i.e. it was not aborted and stepped down.
failures	int	1 if the test was aborted for some reason.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
is_to_server_in_nz	boolean	Is the test server located in New Zealand?
during_which_rwc_2019_game	string	The Rugby World Cup game which was being broadcast when the test was run. Blank if not during any game.

Data dictionary

unit_summary_statistics_download_upload_latency.csv

Field Name	Type	Description
unit_id	int	Unique identifier for an individual unit.
technology	string	The broadband technology the Whitebox was assigned to
geographical_area	string	The geographical area the Whitebox was located in
target_server_country	string	The country in which the test server is located. Each unit's results are reported separately for each test server country.
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.
mean_download_mbps_during_abs_vs_boks	decimal	The (untrimmed) mean of download_mbps - only considering tests where during_which_rwc_2019_game = "New Zealand v South Africa".
mean_download_mbps_during_peak_hour_rwc_2019_games	decimal	The (untrimmed) mean of download_mbps - only considering tests where during_which_rwc_2019_game is not blank and is_during_peak_hour is TRUE.
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.
mean_upload_mbps_during_abs_vs_boks	decimal	The (untrimmed) mean of upload_mbps - only considering tests where during_which_rwc_2019_game = "New Zealand v South Africa".
mean_upload_mbps_during_peak_hour_rwc_2019_games	decimal	The (untrimmed) mean of upload_mbps - only considering tests where during_which_rwc_2019_game is not blank and is_during_peak_hour is TRUE.
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.
mean_latency_ms_during_abs_vs_boks	decimal	The (untrimmed) mean of latency_ms - only considering tests where during_which_rwc_2019_game = "New Zealand v South Africa".
mean_latency_ms_during_peak_hour_rwc_2019_games	decimal	The (untrimmed) mean of latency_ms - only considering tests where during_which_rwc_2019_game is not blank and is_during_peak_hour is TRUE.
latency_samples_24h	int	The number of upload tests (count of rows in raw_latency_tests.csv).
latency_samples_peak	int	The number of upload tests (count of rows in raw_latency_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.

Data dictionary

unit_summary_statistics_netflix_youtube.csv

Field Name	Type	Description
unit_id	int	Unique identifier for an individual unit.
technology	string	The broadband technology the Whitebox was assigned to
geographical_area	string	The geographical area the Whitebox was located in
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.
youtube_uhd_fraction	decimal	The percentage of successful YouTube tests where youtube_bitrate_mbps was greater than 6.8 and successes was equal to 1.
youtube_hd_fraction	decimal	The percentage of successful YouTube tests where youtube_bitrate_mbps was between 2.5 and 6.8, and successes was equal to 1.
youtube_sd_fraction	decimal	The percentage of successful YouTube tests where youtube_bitrate_mbps was greater than less than 2.5, and successes was equal to 1.
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.
youtube_successes_plus_failures	int	Each 'test' starts by trying to stream at UHD; if it is not possible to stream at UHD then the test reports 'failure' and tries to stream at HD. If it is not possible to stream at SD, the test reports failure and tries to stream at SD. This field keeps track of the total number of tests at a more granular level than used in the report.
youtube_hd_or_better_fraction	decimal	youtube_uhd_fraction + youtube_hd_fraction. Since not all YouTube videos are available in UHD, the report considers the fraction of tests which could stream at 'HD or Better'.