Analysis of Long Duration Traces

Richard Nelson, Daniel Lawson, Perry Lorier
University of Waikato
Private Bag 3015
Hamilton, New Zealand
{richardn,dlawson,perry}@cs.waikato.ac.nz

ABSTRACT

This paper introduces a new set of long duration captures of Internet traffic headers. The capture is being performed on a continuous on-going basis and is approaching a year in duration. Based on the current extent of the archive some typical analyses are presented, covering protocol mix, network trip times and TCP flag analysis.

1. INTRODUCTION

Collection of network traffic data is typically only performed for short periods of time. There are several reasons for this, including the large volume of data that is generated, the difficulty of maintaining accurate clock synchronization for long periods of time and the purpose of the collection. Providing these difficulties can be overcome however, collection of data over a long time duration can provide data sets of unique value for a wide variety of network analyzes.

The usefulness of long duration network traces has been shown by the popularity of the Waikato Internet Traffic Storage (WITS [7]) data sets. The largest of these, the Auckland-IV set, provides data from a 45 day period in 2001 and contains 3.157 billion IP Headers. Despite its size (65 GB) it has been used by many researchers for data analysis and has been cited in many publications, eg [3, 4, 5, 6].

Based on the success of the WITS archive the Waikato University Network Research Group (WAND) has embarked on a new effort to build an archive of network traces of significantly longer durations and greater size.

The aim of this effort is to build an archive that is of significant size providing useful new data in a generic format for many types of analysis. The data set is continuously updated. A key reason for doing this is to provide long term data that can be analyzed to find trends, either pro-actively to find new trends or retrospectively to show the emerging pattern of established trends.

2. WAIKATO CAPTURE POINT

The new WAND capture effort is based at our university campus with a permanent dedicated installation designed to allow long-term continuous network trace capture. Access to perform the capture is under an agreement with the University of Waikato Information Technology Services (ITS) department that provides strict privacy for user data.

The Waikato Capture Point is located at the boundary of the University of Waikato network, outside the firewall. The uplink is a 10Mbps connection, which is shaped by the tel-co into a 5Mbps national (NZ) circuit and a 5Mbps international circuit upstream from the capture point. A header-only capture is performed and the headers are sanitized before being delivered back to the WAND server room via a private network, where it is stored to disk. The sanitation process involves verifying the original IP checksum and replacing with a boolean indicating a valid or an invalid checksum, and anonymizing the IP addresses using Crypto-PAN [1]. The capture point uses an Endace DAG3.5E [2] capture card with GPS time synchronization. The card provides a signal that the link layer packet was received correctly. Processing based on the IP address is used to determine the direction of the packet. The capture point has a rated capacity five times the peak bidirectional bandwidth limit of the uplink. Additionally, the DAG hardware is able to identify whether any packets are lost through buffer overflow. So far, this has not occurred in nearly a year of continuous operation. The only possibility of packet loss that we know of is if the capture point is disconnected from the switch for some reason. It is located in a secure network facility and we do not believe this has happened during the capture so far.

2.1 Details of Data Traces

Continuous capture started on December 6 2003 and has run without interruption until the time of writing (approximately 320 days). The intention is to continue this capture indefinitely allowing for occasional restarts for maintenance and upgrades. The trace files are stored on disk although there may come a point where older data has to be deleted due to lack of storage.

The trace files are collected in Endace Record Format (ERF) and are stored as one file per day using gzip compression. As the collection is still ongoing, statistics on the total trace are only approximate. So far we have 680 GB of compressed traces (averaging about 2GB/day) storing headers for 25 billion packets (80 million / day). These packets represent 9.8 TB of transmitted data (32 GB/day).

Despite quoting daily mean values, in fact very few days are 'average' as there are strong weekly cycles and strong variations with the level of student activity on campus, primarily based on the point within the current semester.

As can be seen in the protocol analysis in section 3.1 file-sharing and peer to peer traffic do not make up a large part of the traffic observed. While this will obviously differ from a typical consumer ISPs traffic model, the traces captured the University of Waikato may fit closer with a typical large corporate network with rules on

acceptable network use.

Waikato University holds a /16 network internally, along with a handful of other /24 networks. Usage is admittedly sparse, with approximately 7000 active IPs internally. We are currently performing analysis of backscatter traffic on our address space, and have observed about 600 internal IPs actively participating in traffic across our uplink. In most cases, web and mail traffic go through the campus web proxy and mail servers, although there are some machines which generate traffic directly.

Due to the anonymization of addresses and removal of all user data there is no legal/ethical impediment to sharing of this data with other researchers. However the volume of data and the Internet charging regime in New Zealand mean that there is a significant practical difficulty in publishing the data. A couple of initiatives are just starting that may provide a solution to this. In the meantime researchers with an interest in the data should contact the authors to negotiate access.

3. DATA ANALYSIS

As the data has not been collected for any specific study, but to provide a generic resource. In this section we show some typical graphs from the data collected. Graphs do not distinguish between directionality of traffic.

3.1 Protocol Analysis

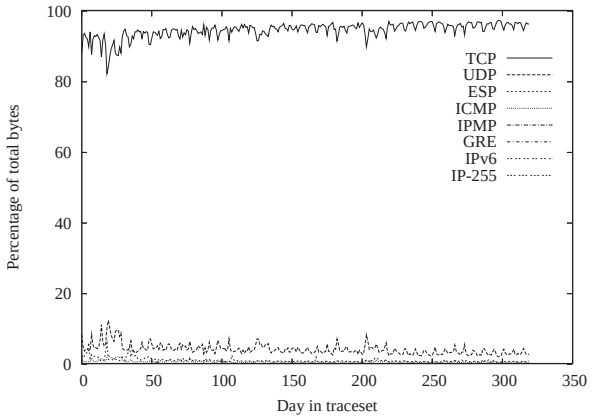


Figure 1: Percentage of the top IP transport protocols observed by byte count per day

The bulk of the traffic traversing the University of Waikato's uplink is TCP, as shown in Figure 3.1. These graphs show, for each day of the trace, the percentage of the total traffic for each IP protocol observed. Displaying this data as a proportion of the total traffic slightly smooths out the strong weekly variation observed. One artifacts of this is that when the total traffic observed is very low any constant traffic observed has a greater effect. This is especially observed in the first few weeks of the trace which is the Christmas and southern hemisphere summer vacation period, where the percentage of TCP traffic dips and the other protocols, particularly ICMP, increase.

Figure 2 splits the former graph up for improved readability. We observe a negligible amount of IPv6 encapsulation and GRE traffic, and we start observing some IPMP traffic in May, corresponding with the Waikato University AMP monitor joining the full HPC AMP mesh. (Figure 2(d)).

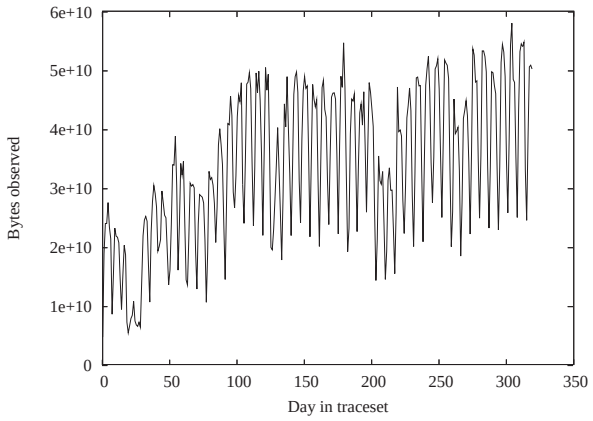


Figure 3: Total bytes observed per day

Figure 3 displays the raw values for the total byte count measured. As well as the strong weekly variations, there are large drops in use over the Christmas / New Year holiday period (around day 25), and subsequent drops in usage over the teaching recesses in April, July and September (around days 130, 210 and 270, respectively). The apparent growth in the University's Internet traffic over the capture period is exaggerated due to low usage during the New Zealand summer vacation period at the beginning of the trace.

Figure 4 shows the absolute distribution of the top 5 TCP ports observed across the entire dataset by packet and byte count. The two graphs show very similar patterns, including the strong weekly variation due to weekends, and the overall fall-off consistent with holiday and student recess periods.

This shows the reason for the increase in TCP domination at the protocol level which is the continuing increase in web traffic. As well as port 80 (HTTP) and 443 (HTTPS), we see a reasonable volume of traffic on port 2048. The University of Waikato has a proxy server running on this port to service online journal access. The noticeable increase in port 25 (SMTP) traffic at around day 200, lasting for a reasonable period of time, corresponds with an ourbreak of the Zafi.B email worm.

Figure 5 shows the top 5 TCP protocol byte counts in greater detail. The packet counts have not been expanded in this fashion for brevity.

It should be noted that there is a very low level of peer-to-peer traffic present. We believe this is predominantly due to the charging regime. To download a CD worth (600MB) of data would cost a student around NZ$70 and even to download 50 mins of MP3 music ($\sim$ 50MB) would cost NZ$6, whereas to buy a blank CD-ROM for copying this costs about NZ$1.

UDP traffic across the link, as shown in Figure 6, is largely dominated by DNS traffic. The pattern of requests does not closely follow the pattern shown by TCP traffic however (Figure 4), although artifacts such as the Christmas holiday period are still demonstrated in here. There is a consistent level of NTP traffic throughout the entire year.

A few interesting ports appear. We see a consistent volume of UDP port 1701 traffic, which is the registered port for L2TP VPN connections. Ports 1026 and 1027 show a sudden increase in usage at about day 150. Both of these ports are associated with Windows Messenger Service, and so are probably indicating 'Windows Messenger Service Spam', although there is some indication that UDP port 1027 is used for legitimate ICQ traffic. The correlation

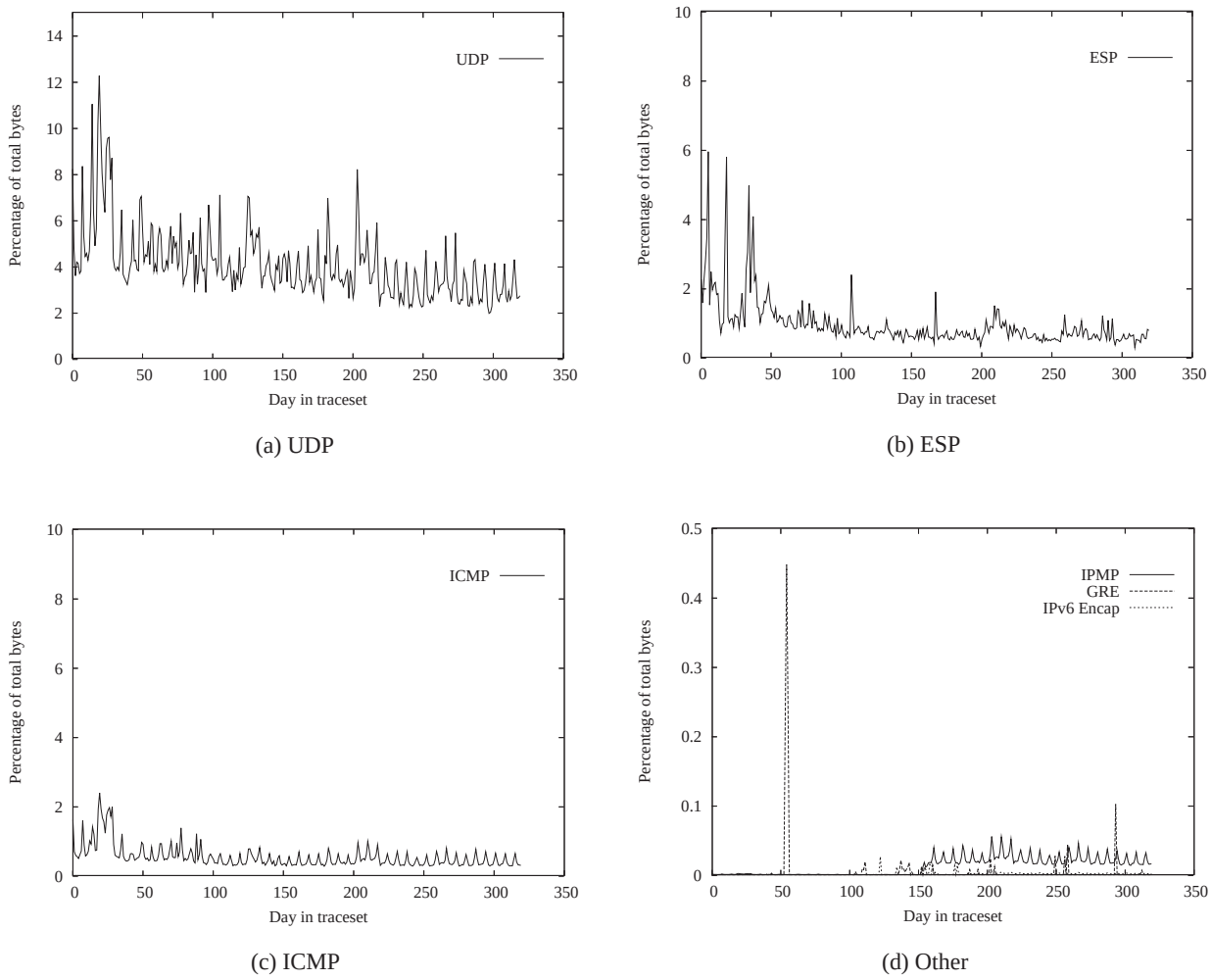


Figure 2: Percentage of individual IP transport protocols observed by byte count per day

between the increase in traffic on both ports implies that they are both related to the Windows Messenger Service Spam. . There are many background probes on UDP port 137, the Netbios Name Service port. Ports 6970 and 27960 are used for RealAudio streaming, and a popular online game, respectively.

Figure 7 shows the top 5 UDP protocol byte counts in greater detail. The packet counts have not been expanded in this fashion for brevity.

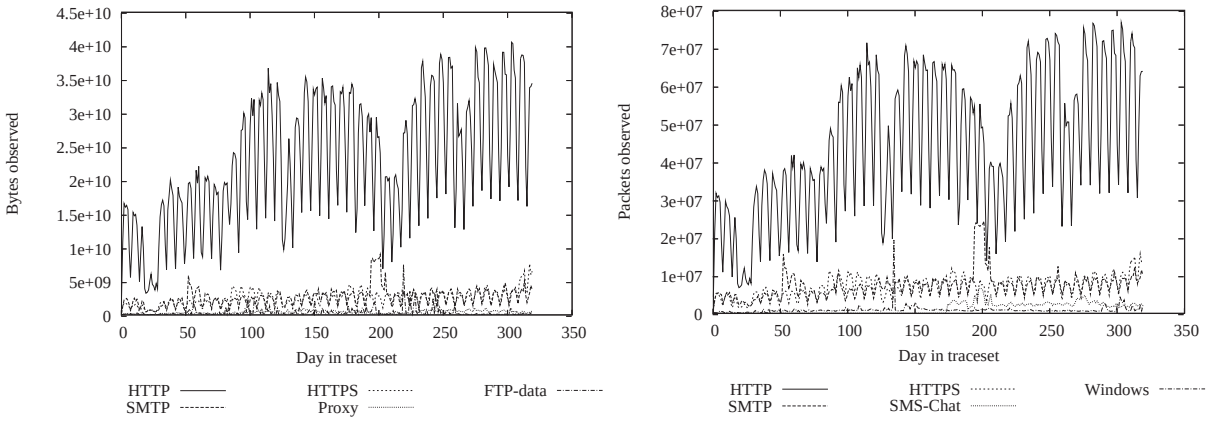
3.2 Round Trip Times Analysis

Due to the volume of data and the close proximity of the observation point to one end of all the observed connections a simpler analysis of the return trip times from the observation point was conducted, based using the TCP timestamp option as a unique key. Trip times were calculated using the DAG timestamps. This means that each TCP connection generates two sets of observations, one from the capture point into the University network and one from the capture point into the global Internet. Further recent analysis shows that between 50 and 60 percent of TCP connections have the timestamp option.

The density plot of trip times observed each minute for one day (May 24, 2004) are shown in Figure 8. This shows the structure of the Internet as seen from New Zealand with distinct banding separated by gaps. The gaps mostly represent the oceanic hops. The bottom set of bands in the figure are the return trip time to the University machines observed on incoming connections. The dark line is mainly the University servers including the web proxy and mail server. The lines below this are machines on the same Ethernet segment. The lighter band above this are connections to New Zealand based machines. These are a small proportion of the total and on the log scale have the widest distribution of trip times. Above this is the hop across the Pacific Ocean to the United States. Any points that appear within this gap are due to either very long queuing delays on NZ based connections or connections to countries that have shorter direct connections such as Australia (although many, but not all connections to Australia are routed via the US).

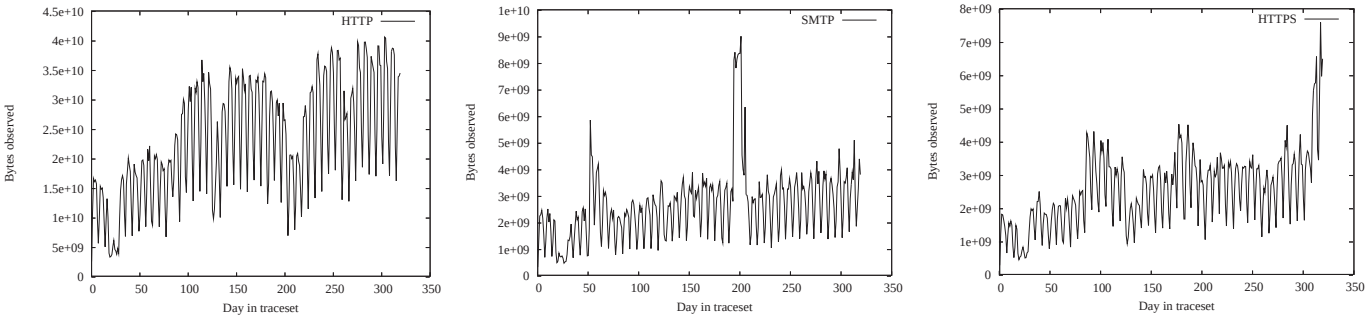
The top bands represent the structure beyond the US East Coast with the US West Coast appearing and then other oceanic hops either to Europe or Asia. This structure is more blurred due to the complexity of the structure of the global Internet and are fairly compressed due to the log scale.

Although this graph is only for one day, the plot for one year

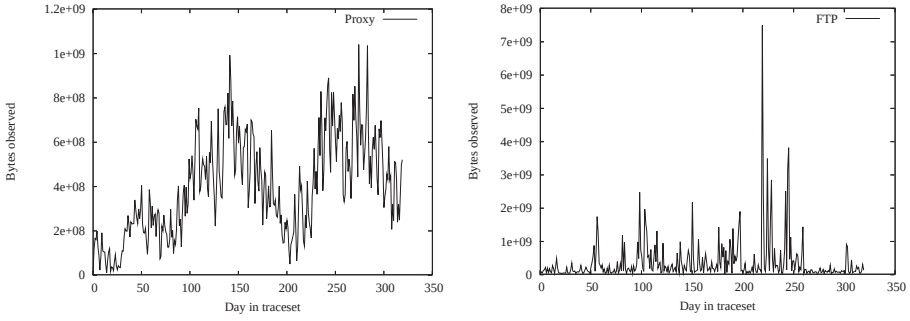


(a) bytes per day (b) packets per day

Figure 4: Top 5 TCP ports observed per day for the capture duration



(a) HTTP (b) SMTP (c) HTTPS



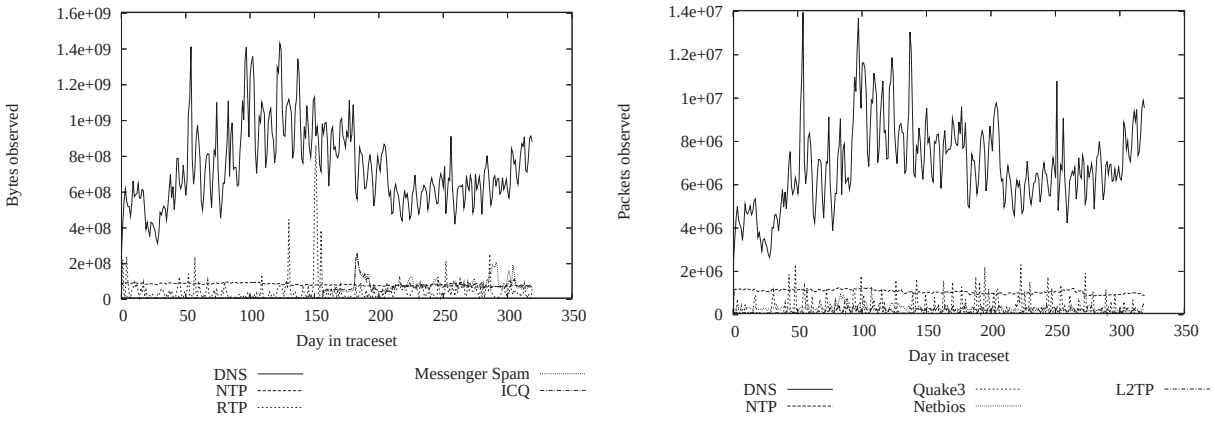
(d) Proxy (e) FTP

Figure 5: Individual plots of the top 5 TCP protocol bytecounts, by day

is almost identical. Routing changes, although known to be quite frequent make almost no apparent impact, except in small measure to the NZ band. This is mostly because of the fixed size of the Pacific Ocean and the small number of cables that reach NZ.

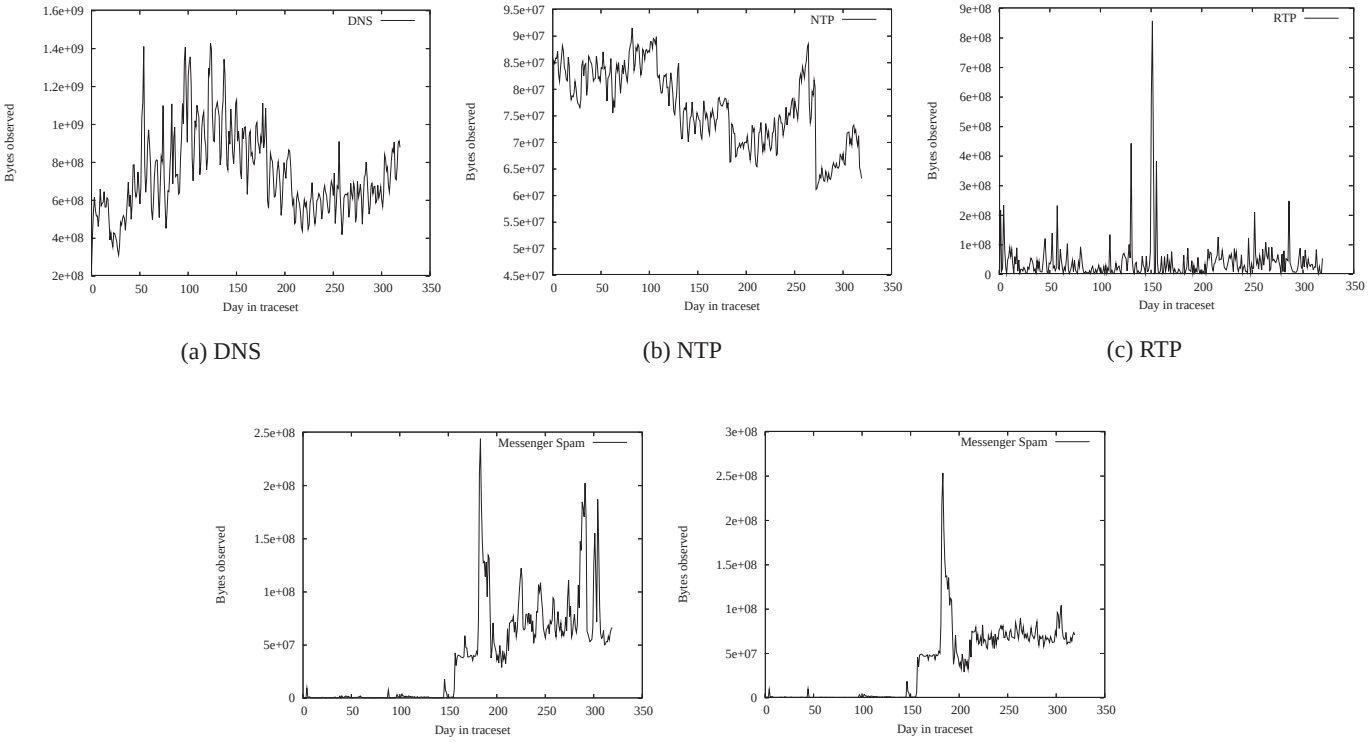
A log scale histogram of the same day's observations are given in Figure 9. The trip time bins are exponentially sized to give a bin size proportional to its value across the entire range measured. The large peak of University traffic can be seen at around 1 ms and below. The small broad peak of NZ traffic is between 10 and 50ms with the international bands occurring above 100ms.

3.3 TCP Analysis



(a) bytes per day (b) packets per day

Figure 6: Top 5 UDP ports observed per day for the capture duration



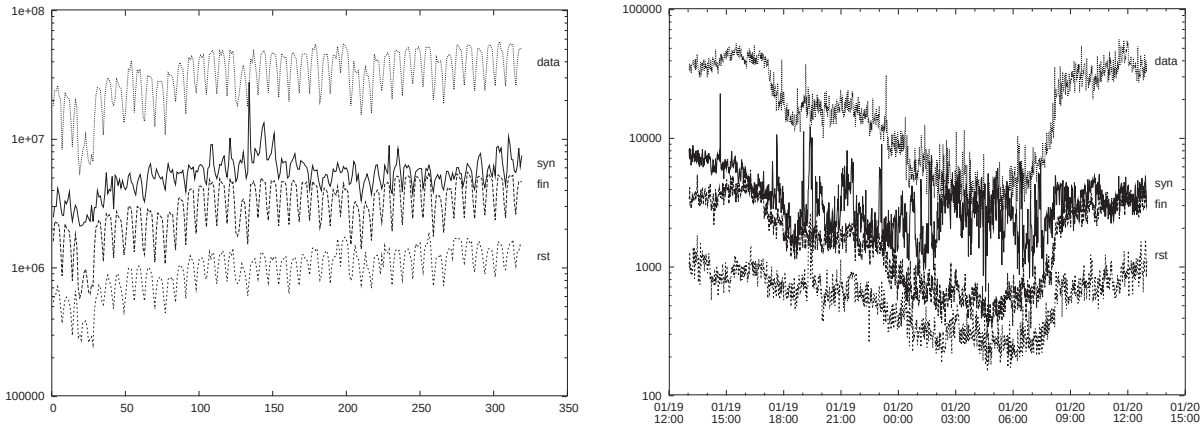
(a) DNS (b) NTP (c) RTP

(d) UDP 1026 (e) UDP 1027

Figure 7: Individual plots of the top 5 UDP protocol bytecounts, by day

Tracking the use of TCP flags tracks attempts to setup TCP connections and the number of connections that are reset or closed. Figure 10(a) plots the daily use over the entire dataset of the TCP flags SYN, FIN and RST as well as the number of segments seen with non-zero payload (DATA). The trend in total SYNs seen does not follow that of total data in that it decreases somewhat in the second half of the year, moving much closer to the number of FINs. This may indicate a reduced level of network scanning, but whether this is due to improved security or increased sophistication by those scanning is unknown.

A per-minute trace for a day (January 19, 2004) is shown in Figure 10(b). This is a typical pattern and the significant feature is



(a) Flags per day for one year

(b) Flags per minute for one day

Figure 10: TCP flags.

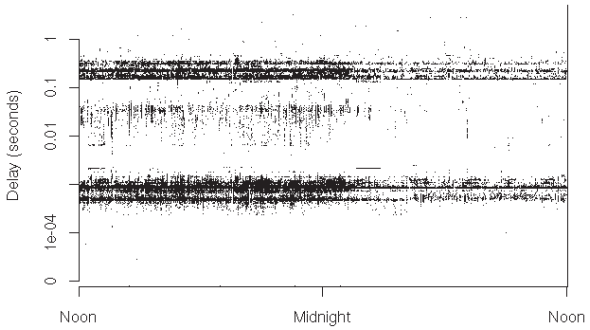


Figure 8: Return trip times for one day from the observervation point

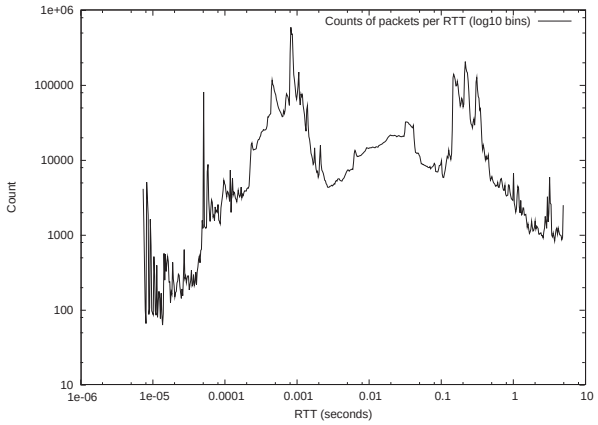


Figure 9: Histogram of return trip times for one day from the observation point

the behavior of the SYN traffic, particularly overnight. During the period between midnight and 9:00 the volume of data reduces as do the FINs and RSTs seen however the volume of SYNs seen do not. This indicate that the volume of incoming invalid SYNs increases during this period. The cause of this is not clear, possibly it is linked to the number of ”zombie” machines that are turned on in other time zones. Due to the anonymization of addresses we are not able to investigate the source of these SYNs.

The other feature of the SYN data are the spikes that occur between 19:00 and 22:00. During the first spike the volume of RSTs seen also spikes whereas it doesn't during the second. The volume of FINs does not spike at all.

Figure 11 shows the rate of SYN packets seen during one day (October 12, 2004). The spikes on this day were caught by chance but it is not particularly unusual. Two different types of event are visible, the shorter duration higher peak events seen around 18:00 and the long duration event at 05:00.

One other aspect of TCP flag analysis of note is that the PSH and URG flags are effectively redundant. The PSH flag because it is used on nearly all TCP data packets and the URG flag because it is used on almost none.

4. CONCLUSIONS

This new data collection effort by the WAND group has achieved long term continuous trace collection. It is due to achieve a year's collection, hopefully without a break on December 6 2004. This provides a dataset unique in length and significant in volume. It is able to shared, but practical difficulties need to be addressed

The analyses in this paper show typical measurements of Internet traffic characteristics using the whole dataset. We are further developing our tools to make analyzes such as these routine, continually updated and publicly available.

5. REFERENCES

[1] Crypto PAN.
http://www.cc.gatech.edu/computing/Telecomm/cryptopan/.

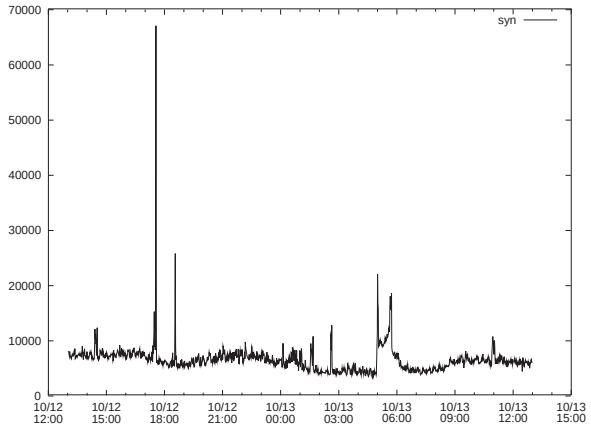


Figure 11: Per minute rate of TCP SYN packets seen over one day.

[2] DAG Group. http://dag.cs.waikato.ac.nz, October 1999. Computer Science Department, University of Waikato, New Zealand.

[3] N. Hohn, D. Veitch, and P. Abry. Does fractal scaling at the ip level depend on tcp flow arrival processes. In *ACM SIGCOMM Internet Measurement Workshop (IMW-2002.)*, 2002.

[4] N. Hohn, D. Veitch, and P. Abry. The impact of the flow arrival process in internet traffic. In *IEEE International Conference on Acoustics, Speach and Signal Processing ICASSP*, Hong Kong, Apr. 2003.

[5] G. Lazarou, X. Xia, and V. S.Frost. Internet Traffic Modeling Using the Index of Variability . In *IASTED International Conference Modelling and Simulation*, Palm Springs, California, Feb. 2003.

[6] S. Uhlig. Conservative cascades: an invariant of internet traffic. In *IEEE International Symposium on Signal Processing and Information Technology*, Darmstadt, Germany, Dec. 2003.

[7] Waikato Internet Trace Storage. http://wand.cs.waikato.ac.nz/wand/wits/index.html.