Characterizing the Network Connection Behavior of Residential Broadband Subscribers

Shane Alcock[1], Richard Nelson[1], and David Miles[2]

[1] University of Waikato, Hamilton, New Zealand
[2] Alcatel-Lucent
(salcock, richardn)@cs.waikato.ac.nz
David.Miles@alcatel-lucent.com.au

Abstract. Most current estimates predict that IPv4 address exhaustion will occur before 2012. Migrating to the next generation of IP, IPv6, is a costly and complicated process that many Internet Service Providers are reluctant to undertake. One approach to aid the transition to IPv6 is the use of Service Provider NAT (SP-NAT) whereby multiple subscribers share a single public IPv4 address. This solution introduces a new set of limitations and therefore it is important to investigate the likelihood of customers being adversely affected prior to any implementation. This paper presents a study of the behavior of DSL subscribers for an ISP that has not implemented SP-NAT, examining the number of outgoing and incoming connections that are utilized. This behavior is analyzed within the context of SP-NAT to explore whether the current subscriber behavior would be sustainable under a SP-NAT implementation. The results show that there is a distinct class of users who create and sustain significantly more connections than other broadband subscribers and that the majority of subscribers currently use applications that listen for and accept connections from external hosts.

1 Introduction

IPv4 address exhaustion is arguably the most pressing problem facing the production Internet today. Demand for IP addresses is growing exponentially due to increased global Internet connectivity and the popularity of Internet-capable mobile devices such as mobile phones and PDAs. Most estimates now predict that the remaining IPv4 address space will be exhausted before the end of 2011 [8] [12].

IPv6 [3] was developed as a solution to the problem of IPv4 address exhaustion. Although IPv6 is robust and widely supported in both hardware and software, Internet Service Providers (ISPs) have thus far been reluctant to migrate to the new protocol. One reason for this is that the migration process is expensive in terms of time, money and technical expertise. Another problem is that IPv6 is not backwards-compatible with IPv4. Any networks that are not IPv6-capable will be unreachable for anyone who is using IPv6 alone. The need to future-proof the network is unlikely to be an acceptable explanation for dissatisfied customers who are no longer able to access their favorite websites. As

a result, most ISPs considering migration to IPv6 will still need to retain some form of IPv4 connectivity in the interim.

One technique designed to aid the transition to IPv6 that is attracting attention is Service Provider Network Address Translation (SP-NAT). SP-NAT involves the widespread use of Network Address Translation (NAT) [4] to reduce the number of IPv4 addresses required by an ISP. NAT is already utilized on a per-subscriber basis by many ISPs, allowing each customer to maintain their own private network behind a single public IPv4 address. In this case, the NAT device is typically a modem located on the customer premises. By contrast, SP-NAT utilizes NAT-capable routers located inside the carrier network to share a single IP address amongst multiple subscribers, significantly reducing the number of IPv4 addresses required by the ISP while retaining full IPv4 connectivity.

This approach has some limitations. For instance, NAT devices must maintain internal state for each active connection to ensure a consistent translation for packets belonging to the same session. This places an inherent limit on the number of sessions that are available to the group of subscribers sharing an IP address. Once the session limit is reached, the subscribers will be unable to successfully create new connections. Therefore, it is important for ISPs utilizing SP-NAT to provision NAT devices such that the likelihood of peak session demand exceeding the capabilities of the NAT device is minimized.

In addition, end-to-end connectivity is violated by NAT because external computers cannot easily initiate connections to hosts located behind a NAT device. This is not a problem for most common Internet applications, such as web browsing or email, because the connection is initiated by the client. Applications that do not follow the client-server paradigm, most notably peer-to-peer, can be significantly disadvantaged by NAT because subscribers will often expect to be able to accept incoming connections in addition to initiating outbound sessions [11].

To investigate the impact of these limitations, we use packet traces captured from a major New Zealand ISP to examine the behavior of broadband users within the context of a potential SP-NAT implementation. The subscriber behavior is measured by analyzing the number of outgoing and incoming connections being established by the users. Our results show that there is a distinct class of broadband subscribers who utilize their DSL connection in a manner that could not be feasibly sustained under SP-NAT. We also discover that the majority of subscribers are currently accepting inbound connections from external hosts, probably as a result of the modern proliferation of peer-to-peer applications.

2 Data Set

The trace set used for this analysis was captured from a New Zealand ISP between February 8 and February 12, 2007. The ISP is one of the largest providers in New Zealand and sells broadband access to subscribers throughout the coun-

try. The traces were captured using a passive monitor placed between an L2TP (Layer 2 Tunnelling Protocol) server and a core router within the ISP's internal network, recording both incoming and outgoing traffic for a subset of the DSL subscribers. The number of subscribers within the subset is not known, but there are over 18000 unique subscriber IP addresses present in the trace set. Only residential DSL customer traffic was captured; no dial-up customer or large-scale corporate traffic was present in any of the traces.

The capture was performed using a single Endace DAG 3.7G hardware capture card [5]. The trace set is entirely contiguous and no sampling was performed during the capture process, ensuring that every packet observed by the monitor was captured and recorded. Full details about this trace set, known as Local ISP B-III, can be found at [13].

3 Outbound Connections

There are two limiting factors that impact the performance of a NAT device: the rate at which new session table entries must be generated and the total size of the session table. A NAT device creates a new session for each connection and both the maximum session creation rate and the size of the session table are finite. The principal concern for any SP-NAT implementation is provisioning the NAT devices efficiently, i.e. how many subscribers can share the same NAT device without a significant degradation in performance?

To address this question, we analyzed the trace set to determine the number of connections being established and maintained by broadband subscribers. As each connection is equivalent to a session, measuring connection quantities should provide sufficient information to generate suitable provisioning estimates for a SP-NAT implementation.

As we did not have access to any information regarding the allocation of IP addresses to subscribers during the measurement period, we assumed that each unique IP address observed from within the address range of the ISP represented an individual subscriber. This is a somewhat naive approach, as subscriber addresses were allocated dynamically from a pool of addresses by the ISP and could have changed whenever a customer rebooted their DSL modem. However, most results presented in this paper are averaged across the number of subscribers observed during a thirty minute period. It is unlikely that the proportion of DSL subscribers switching addresses during that time span was significant enough to adversely affect the results to any notable degree.

Only connections initiated by the subscriber will add an entry to a NAT device session table, so initially only outbound connections were examined. An outbound connection was defined as a connection where the first observed packet was sent from within the ISP network. For TCP connections, the initial packet was also required to be a SYN packet, removing any malformed or invalid TCP connections from the analysis. Unfortunately, the UDP protocol does not feature a similarly obvious indication of connection commencement so we were forced to

assume that the first observed UDP packet always signified the start of a valid UDP connection.

A TCP connection was defined as established if a SYN packet had been observed in both directions. Unestablished TCP connections were expired after four minutes, or twice the maximum segment lifetime, of inactivity [1]. Established TCP connections were expired after two hours and four minutes of idle time, as defined in [7]. Any TCP connection for which a FIN was observed in both directions was expired immediately, as were connections that observed a RST packet. Connections for other protocols, including UDP, were expired after two minutes of inactivity. Finally, an ICMP error message, would result in immediate expiration.

3.1 Results

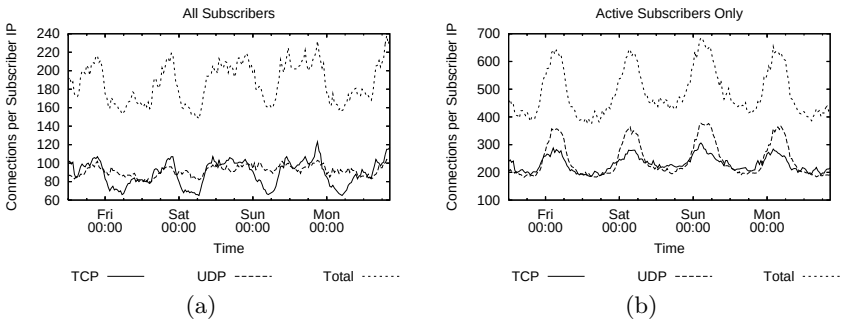

Fig. 1. Mean new connections observed per subscriber.

First, we measured the quantity of new connections observed every thirty minutes for each subscriber. The result of calculating the mean across the total number of unique subscriber IP addresses is depicted in Figure 1(a). As might be expected, the afternoon and evening appear to be the busiest times with a decrease in activity between midnight and 6 a.m.. Most of this variation is due to fluctuations in the rate at which TCP connections are created whereas UDP seems to remain static throughout the day.

For each half-hour period, subscribers that did not register any active connections were removed from the mean calculations to produce the results shown in Figure 1(b). The diurnal variation becomes more distinct but the peaks now occur between midnight and 6 a.m. when only the most prolific users were likely to be active. Also, a similar diurnal pattern appears for UDP, suggesting that the overnight users were engaged in activities that require significant quantities of UDP traffic such as online gaming, file sharing or media streaming.

Arguably the most important result revealed by these graphs is the sheer quantity of new connections being created by subscribers. On average, an active

subscriber was generating between 400 and 600 new outbound connections every half-hour. This number was much higher than we had initially expected especially as, in our experience, consumer-premises DSL routers are typically poor at handling such high quantities of translations.

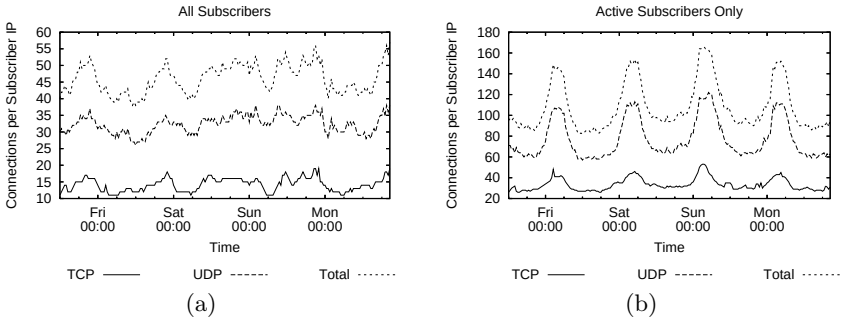

Fig. 2. Mean peak concurrent active connections per subscriber.

To determine the number of session table entries that each subscriber is likely to require, we measured the peak number of concurrently active connections observed for each subscriber IP address every thirty minutes. The mean value calculated across all subscribers for each half-hour period is shown in Figure 2(a). These results suggest that the average DSL subscriber regularly exceeds 40 concurrent sessions, regardless of the time of day. Most of the concurrent connections appear to be UDP, suggesting that the activities that are normally associated with conventional Internet users, such as web browsing and email, were not responsible for most of the simultaneous sessions.

Removing the data for the inactive subscriber IP addresses from each measurement period produces the results shown in Figure 2(b). As before, the mean value increases significantly and there is a very distinct diurnal pattern with peaks occurring during the early morning hours. Once again, this can be attributed to the heaviest users being the only active subscribers during those particular hours.

All of the averages presented thus far were much higher than we had anticipated, suggesting that outlying measurements were having a notable effect on the results. To investigate this, we used the same set of measurements to calculate the median values shown in Figure 3. There are several notable differences in these results. Firstly, the median values are significantly lower than the means that were presented earlier. Secondly, the diurnal trends are an almost exact mirror-image of the pattern observed for the means, with very little activity observed during the early morning hours. Finally, the medians for UDP are particularly low (less than ten new connections and no more than two concurrently active connections every half-hour) and there is little variation in the UDP measurements throughout the entire measurement period.

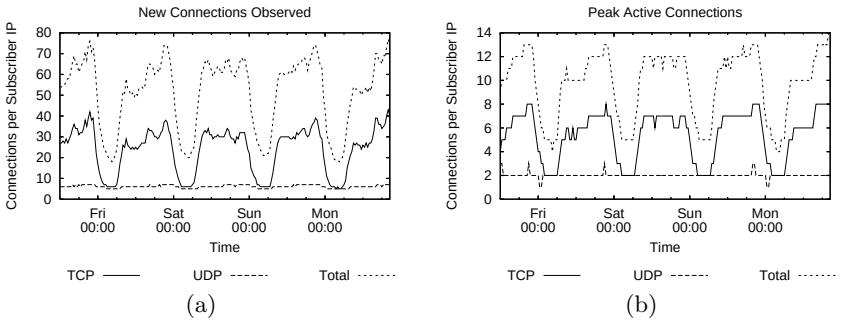

Fig. 3. Median connection counts per active subscriber.

These results illustrate the extent to which the mean values were skewed by the behavior of a minority of subscribers. It also suggests that this minority, the heavy users, utilize residential broadband in an entirely different manner to the other subscribers. This prompted us to ask two questions: what proportion of subscribers fall into this heavy user category and how many connections are the heavy users really creating and sustaining?

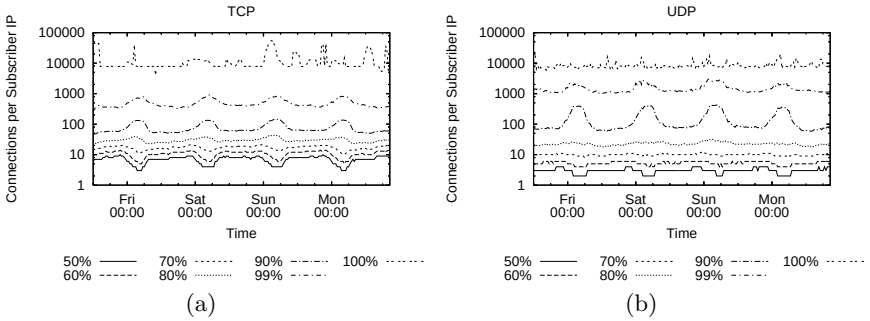

Fig. 4. Percentile analysis of peak active sessions for active subscribers.

To address these questions, we deconstructed the peak active connection counts into percentiles. The results are presented in Figure 4. The behavior that we have associated with heavy users, such as increased use of UDP applications, does not become apparent until the 90th percentile, suggesting that between 10 and 20 percent of subscribers could be classed as heavy users. In addition, the top 1 percent of users appear to be in an entirely different class altogether, sustaining an order of magnitude more active connections than any other single subscriber. The absolute highest value observed was 44788 simultaneously active TCP connections. Maximum UDP utilization is also significant, with 10000 concurrent sessions being surpassed by a single subscriber on multiple occasions.

We have not yet directly investigated the applications that were being used to generate such vast quantities of connections but we suspect that the cause will be some form of malicious behavior, probably the result of an inadvertent malware infection. To distinguish this new category of users from the heavy users, we shall subsequently refer to them as malicious users, regardless of their intentions.

The malicious behavior that we have observed is unlikely to be sustainable under SP-NAT. It is not feasible to provision NAT devices to accommodate an arbitrary subscriber that is maintaining over 10000 concurrent connections. Therefore, we expect that SP-NAT implementors will take action to restrict the amount of outgoing connections that a single subscriber can create and maintain to simplify NAT device provisioning. This will have the additional benefit of reducing the impact of the malicious users on the ISP network.

4 Inbound Connections

As mentioned earlier, NAT violates the IP connectivity model by preventing external devices from directly connecting to any hosts behind a NAT device. This presents a significant obstacle for anyone who wishes to operate an Internet service, such as a web server, or participate in peer-to-peer exchanges. Some applications attempt to circumvent this by utilizing NAT traversal techniques, such as STUN [10]. However, a lack of standardization for NAT device behavior ensures that no single technique will succeed in every situation [6].

An alternative approach is to configure a permanent session, called a port forward, on the NAT device to listen for inbound connections on a specified port number. This solution will be successful in the vast majority of cases, but each port forward will consume a session that would otherwise be available for outgoing connections. Also, a port number may only be forwarded once per NAT device, e.g. two subscribers sharing the same NAT device under SP-NAT cannot both operate web servers that listen on TCP port 80.

To investigate the effects of these limitations, our analysis was extended to include inbound connections, i.e. connections from external hosts to subscribers. A connection was defined as inbound if the first packet observed had originated from outside the ISP address space. In addition, we required that an outgoing response packet had to be observed to eliminate unsolicited traffic such as port scans. As with outbound connections, the initial packet for TCP connections was required to be a SYN. There was no restriction placed on UDP connections due to the lack of flags indicating connection state. This created some problems with the UDP measurements whenever a previously expired outbound UDP session would "reappear" after a long delay. If the observed packet was destined for the subscriber, the inbound connection counter would be erroneously incremented by our analysis.

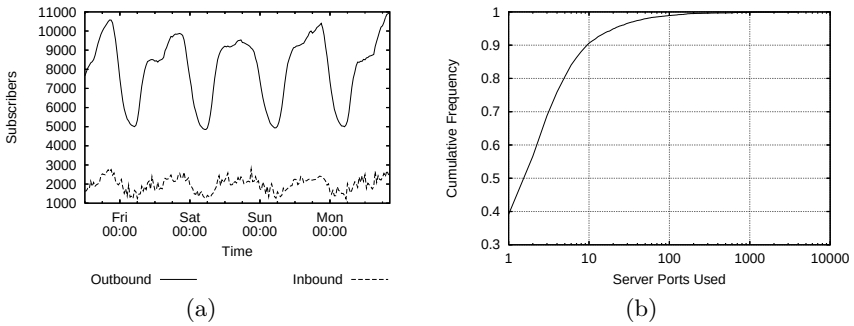

Fig. 5. Analysis of inbound connection behavior. The leftmost graph shows the number of active subscribers observed during each 30 minute period. The graph on the right depicts the distribution of TCP server port usage.

4.1 Results

The number of subscribers that accepted or sustained at least one incoming connection from an external host during each thirty minute period is shown in Figure 5(a). As a point of reference, the number of subscribers that observed an active outbound connection has also been included in the graph. During the busiest hours, over 2500 subscribers appeared to be involved with an inbound connection. Given that many of the subscribers would have been behind a NAT-enabled DSL modem, this is a much higher number than we had anticipated. This could be due to improvements in NAT traversal algorithms used by popular applications. Another possibility is that subscribers are assigning port forwards on their home modem to support the services they wish to run.

Upon further investigation, we found that 43.7% of all observed subscriber IP addresses accepted at least one inbound TCP connection throughout the entire measurement period. When inbound UDP connections were also considered, the proportion of subscribers operating a service rose to 67.1%. As mentioned earlier, differentiating between the start of a new UDP connection and the reappearance of an expired session had proven difficult. Therefore, it is possible that not all of the inbound UDP connections were genuine. Nonetheless, the TCP ratio alone is exceptionally high, suggesting that subscribers accepting incoming connections is normal, rather than exceptional, behavior.

Figure 5(b) shows how TCP server port usage is distributed amongst subscribers. Subscribers that did not accept any incoming TCP connections at all were removed from this analysis. Less than half of the subscribers that were operating a TCP service used just the single server port whereas 20 percent required more than five unique ports. The distribution is very heavy-tailed with a small number of subscribers being observed using in excess of 1000 different server ports.

Further analysis revealed that port utilization was primarily spread across a variety of high-numbered ports. There was no single dominant port being used,

with the most frequently used port accounting for less than five percent of all inbound connections. However, three of the four most popular server ports were default ports for peer-to-peer applications (BitTorrent, eMule and Gnutella). Given that it is common practice for peer-to-peer users to change the default port within the application to avoid traffic shaping, it is likely that peer-to-peer services are even more popular than this simple port-based analysis would suggest [9].

5 Related Work

Due to privacy concerns and commercial sensitivity, it has become difficult to acquire relevant and suitable traffic data that can be used to investigate the behavior of residential Internet users. As a result, there is little existing literature on the subject. The most relevant work that we have encountered is [2] which studied residential broadband utilization in Japan. Instead of examining connection counts, this study used aggregated traffic logs and sampled NetFlow records to analyze raw traffic volumes. The authors were also able to include fiber optic users in their analysis. However, the existence of heavy users as a significant proportion of the subscriber population was also noticed and remarked upon by the authors.

6 Conclusion

This paper presents the results of a study of residential broadband subscriber behavior. The study was conducted with the aim of investigating the potential impact of Service Provider NAT. This was done by examining the number of connections utilized by subscribers and analyzing the results within the context of the known limitations of SP-NAT. In doing so, we made two notable discoveries.

The first was that there were three distinct behavioral classes of broadband customers: typical users, heavy users and (possibly inadvertently) malicious users. Although connection volumes were enough to differentiate between the categories, we also noticed other behavioral patterns that were specific to each user class. For instance, typical users tended to only be active during the evening hours and primarily used TCP applications whereas heavy users utilized UDP applications to a greater extent and were active throughout the entire day.

When evaluating SP-NAT, the most important class of users to consider are the malicious users. Malicious users constituted no more than one percent of subscribers but utilized tens of thousands of simultaneous connections. Designing a SP-NAT implementation to accommodate such behavior is not a feasible option. To ensure that NAT devices can be provisioned sensibly, we expect that SP-NAT implementors will instead attempt to impose limits on the number of connections a subscriber can create and sustain.

We also discovered that over 40% of subscribers use their residential broadband to operate at least one TCP service. Our results suggest that the principal

cause of this is the proliferation of peer-to-peer applications. Despite recent improvements in NAT traversal techniques, we expect that port forwarding will be required to continue to fully support the current subscriber behavior. This will also have implications for NAT device provisioning estimates as each port forward will permanently consume a session on the NAT device.

In the future, we aim to extend this work through further investigation into the behavior of individual subscribers, especially the malicious users. We are also considering the value of other metrics that could be used to profile subscriber behavior, such as connection duration. Detailed analysis aimed at determining the applications being utilized by subscribers is another possibility, as the port-based metrics used in this study appeared to be inadequate. Finally, we are also interested in repeating the entire analysis using trace sets captured from other ISPs.

References

1. Braden, R.: RFC 1122 - Requirements for Internet Hosts - Communication Layers (1989).
2. Cho, K., Fukuda, K., Esaki, H., Kato, A.: The Impact and Implications of the Growth in Residential User-to-User Traffic. In SIGCOMM'06, pages 207-218, Pisa, Italy (Sept. 2006).
3. Deering, S., Hinden, R.: RFC 2460 - Internet Protocol, Version 6 (IPv6) Specification (1998).
4. Egevang, K., Francis, P.: RFC 1631 - The IP Network Address Translator (NAT) (1994).
5. Endace, `http://www.endace.com`, (Accessed on 16 September 2008).
6. Guha, S., Francis, P.: Characterization and Measurement of TCP Traversal through NATs and Firewalls. In Proceedings of the Internet Measurement Conference, Berkeley, CA (October 2005).
7. Guha, S. (Ed), Biswas, K., Ford, B., Sivakumar, S., Srisuresh, P.: NAT Behavioral Requirements for TCP (Internet-Draft), `http://www.ietf.org/internet-drafts/draft-ietf-behave-tcp-08.txt` (Accessed on 16 September 2008).
8. Huston, G., IPv4 Address Report, `http://www.potaroo.net/tools/ipv4/index.html` (Accessed on 10 September 2008).
9. Karagiannis, T., Broido, A., Brownlee, N., claffy, kc., Faloutsos, M.: Is P2P Dying or Just Hiding?, In Globecom 2004 (2004).
10. Rosenberg, J., Weinberger, J., Huitema, C., Mahy, R.: RFC 3489: STUN - Simple Traversal of User Datagram Protocol (UDP) Through Network Address Translators (NATs) (2003).
11. Srisuresh, P., Ford, B., Kegel, D.: RFC 5128 - State of Peer-to-Peer (P2P) Communication across Network Address Translators (NATs) (2008).
12. van Beijnum, I.: IPv4 Address Consumption. In: The Internet Protocol Journal - Volume 10, No. 3 (September 2007).
13. WAND Network Research Group: WITS: Local ISP B-III, `http://www.wand.net.nz/wits/localisp/b/3/` (Accessed on 16 September 2008).