2 Broadband Traffic

2.1 Introduction

In this whitepaper we will analyze traffic for the broadband access services that IIJ operates, and present our findings.

It has been reported that growth in Internet traffic levels over the past five years has been comparatively stable both in Japan and internationally (see references 3, 4, and 5, p.23). The total volume of broadband traffic in Japan is increasing at an annual rate of 30%, and this represents approximately 60% of the overall backbone traffic in Japan. The majority of individual Internet users have broadband access so that identifying broadband traffic trends is also important for understanding overall traffic. (See references 1 and 2, p.23).

This whitepaper examines recent broadband traffic trends based on the daily traffic volume of users and usage levels by port. Heavy users of communication applications such as P2P file sharing still account for a dominant portion of traffic volume, but traffic attributed to these users has not increased significantly. On the other hand, the volume of traffic attributed to general users is steadily increasing, due to a surge in video content and content-rich websites.

2.2 About the Data

The survey data utilized in this whitepaper was collected using Sampled NetFlow from the routers accommodating fiber-optic and DSL broadband customers of our personal and enterprise broadband access services. Because broadband traffic trends vary between weekdays and weekends, we analyzed a full week of traffic, in this case the period from May 25 to May 31, 2009. For comparison, we used the period from February 21 to February 27, 2005. In 2005, video sharing services such as YouTube and Nico Nico Douga had yet to appear.

The usage levels of each user were obtained by matching the IP address assigned to each user with the IP addresses observed. We collected statistical information by sampling packets using NetFlow. The sampling rate was set to either 1/1024, 1/2048, 1/4096, or 1/8192, depending on router performance and load. We estimated overall usage levels by multiplying observed usage levels by the reciprocal of the sampling rate. Depending on the sampling rate, there may be slight estimation errors in data for low-volume users, but for users with reasonable usage levels we were able to obtain statistically meaningful data.

Approximately the same numbers of fiber-optic and DSL users were observed as in 2005. However, the migration to fiber-optic connections advanced in 2009, with 84% of the observed users now using fiber-optic connections, which represent 90% of the overall volume of traffic.

The IN/OUT traffic presented in this whitepaper indicates directions from an ISP's perspective, with IN representing uploads from users, and OUT representing user downloads.

2.3 Daily Usage Levels for Users

First, let us examine the daily usage levels for broadband users from several perspectives. Daily usage indicates the average daily usage for each user over the period of a week.

Figure 1 indicates the average daily usage distribution (probability density function) per user, divided into uploads (IN) and downloads (OUT), with user traffic volume on the X axis, and probability density of users on the Y axis. The X axis indicates volumes between 10^4 (10 KB) and 10^{11} (100 GB) using a logarithmic scale. Some users are outside the scope of the graph, with usage for the highest volume users climbing to over 200 GB, but most fall within the scope of 10^{11} (100 GB). A slight spike appears on the left side of the

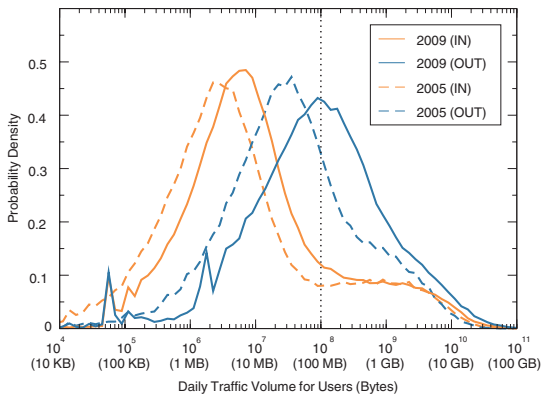

Figure 1: Daily User Traffic Volume Distribution

© 2009 Internet Initiative Japan Inc.

2009 graph, but this is just noise caused by the coarser sampling rate according to increased traffic.

The distribution for IN and OUT shows almost log-normal distribution, which is the normal distribution in a semi-log graph. A linear graph would show a long-tailed distribution, with the peak close to the left end and a slow decay towards the right. The OUT distribution slips further to the right than the IN distribution, indicating that the download volume is an order of magnitude larger than the upload volume. As average values are pulled up by the heavy users on the right side of the graph, the average IN volume was 430 MB in 2005 and rose to 556 MB in 2009. The average OUT volume was 447 MB in 2005 and rose to 971 MB in 2009.

Looking at the right end of the IN distribution, you will notice another small peak in the distribution. In fact, a similar peak can be seen on the OUT side, overlapping with the main distribution. These distributions have IN and OUT volumes at about the same position, indicating heavy users with symmetrical IN/OUT volumes. For convenience, we will call the asymmetrical IN/OUT distribution that makes up the vast majority "client-type users," and the distribution of heavy users with symmetrical IN/OUT volumes making up a minority on the right side "peer-type users."

Comparing the most frequent distribution value for client-type users in 2005 and in 2009, the IN volume rose from 3.5 MB to 6 MB, and the OUT volume rose from 32 MB to 114 MB. This demonstrates that, particularly in the case of downloads, the traffic volume for each user has increased dramatically. In contrast, there was no significant change in the most frequent distribution value for peer-type users, which approached 2 GB in both 2005 and 2009. In other words, while usage levels for general users have increased greatly, usage levels for heavy users have remained constant.

While not shown in the figure, looking into similar distributions for both fiber-optic and DSL connections, distribution points for client-type and peer-type users are about the same for a given year, but for fiber-optic connections the ratio of peer-type users is greater. This means that while there is no difference in the typical usage levels for each distribution, there is a larger ratio of heavy users in fiber-optic connections. The figure of 2 GB per day as the most frequent distribution value for peer-type users is equivalent to 185 kbps when converted to bits/second.

Figure 2 shows daily traffic volume for users in complementary cumulative distribution form. This indicates the percentage of total users with usage levels lower than the X axis value on the Y axis using a logarithmic scale, which is an effective way of examining the distribution of heavy users. The right side of the graph falls linearly, showing a long-tailed distribution close to power-law distribution. It can be stated that heavy users are distributed statistically, and are by no means a unique type of user.

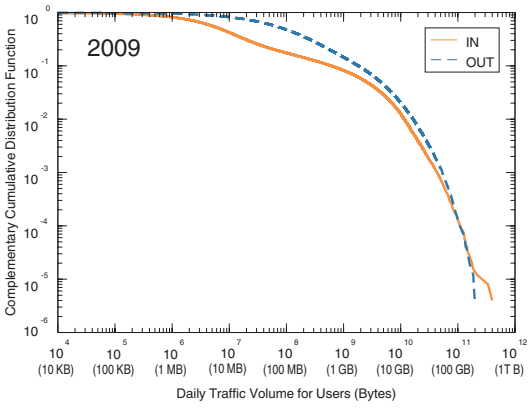

Figure 2: Complementary Cumulative Distribution of the Daily Traffic Volume for Users

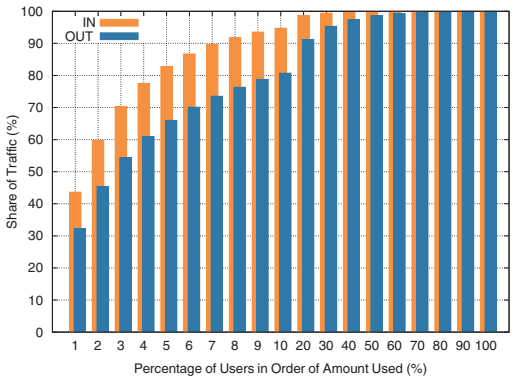

Figure 3: Traffic Usage Deviation Between Users

© 2009 Internet Initiative Japan Inc.

Figure 3 indicates the deviation in traffic usage levels between users. It shows that users with the top X% of usage levels account for Y% of the total traffic volume. There is a great deal of deviation in usage levels, and as a result traffic volume for a small portion of users accounts for the majority of the overall traffic. For example, the top 10% of users make up 80% of the total OUT traffic, and 95% of the total IN traffic. Furthermore, the top 1% of users make up 30% of the total OUT traffic, and 40% of the total IN traffic. Although minor fluctuations occur depending on the behavior of top users, this deviation is mostly unchanged from 2005. This is a characteristic of long-tailed distributions, and a trend that matches Internet data. For example, even when deviation after removing peer-type users is examined, almost the same deviation is observed. Deviations like this are not at all uncommon outside the Internet as well, and are known to appear often in large-scale, complex statistics such as the frequency of word usage and the distribution of wealth.

At a glance, you may get the impression that traffic deviations between users are polarized between those who are heavy users and those who are not. However, the distribution of usage levels follows power-law, demonstrating that a diverse range of users exist.

Figure 4 plots the individual IN/OUT usage levels for 5,000 randomly sampled users in 2005 and 2009. The X axis shows OUT (download volume) and the Y axis IN (upload volume), both using a logarithmic scale. When the IN/OUT volumes for a user are identical, they are plotted on the diagonal line.

Two clusters can be observed. The cluster below the diagonal line and spread out parallel to it is client-type general users with download volumes an order of magnitude higher than their upload volumes. The other cluster is peer-type heavy users spread out around the diagonal line in the upper right. However, the boundary between the two clusters is ambiguous. This is because client-type general users also use peer-type applications such as Skype, and peer-type heavy users also use download-based applications on the web. In other words, many users use both types of applications in varying ratios. There are also significant differences in the usage levels and IN/OUT ratio for each user, pointing to the existence of diverse forms of usage.

By comparing 2005 and 2009, we can see that the center of the client-type cluster is moving towards the upper right, and the peer-type cluster is spreading out and becoming less dense.

2.4 Usage by Port

Next, we will look at a breakdown of traffic from the perspective of usage levels by port. Recently, it has been difficult to identify applications by port number. Many P2P applications use dynamic ports on both ends, and a large number of client/server applications use port 80 assigned for HTTP to avoid firewalls. To broadly categorize, when both parties use a dynamic port higher

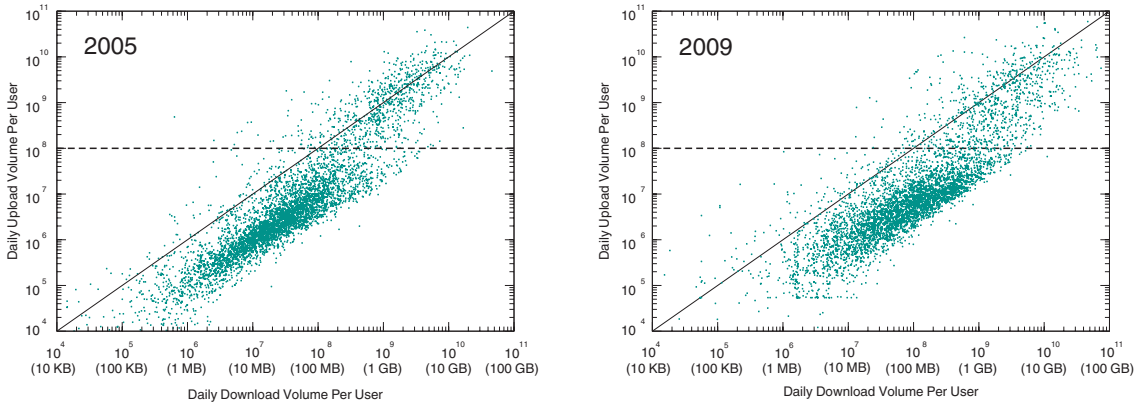

Figure 4: IN/OUT Usage for Each User, 2005 (left) and 2009 (right)

© 2009 Internet Initiative Japan Inc.

than port 1024, there is a high possibility of it being a P2P application, and when one party uses a well-known port lower than port 1024, there is a high possibility of it being a client/server application. In light of this, we will look at usage levels for TCP and UDP connections by taking the lower port number of the source and destination ports.

As overall traffic is dominated by peer-type heavy user traffic, to examine trends for client-type general users, we have taken the rough approach of extracting data for users with a daily upload volume of less than 100 MB, and treating them as client-type users. This corresponds to the intermediate point between the two IN distributions in Figure 1, and users below the horizontal line at the IN = 100 MB point in Figure 4.

Figure 5 shows an overview of port usage, comparing all users and client-type users for 2005 and 2009. Table 1 shows detailed numeric values for this figure.

Over 95% of traffic is TCP based. Looking at the overall picture, the majority of traffic is through TCP dynamic ports, with both parties using dynamic ports for 78% of the total traffic in 2009. Specific ports in the dynamic port range make up a small percentage, comprising 1.1% of the total traffic at most. Use of port 80 has increased from 9% in 2005 to 14% in 2009.

When data is limited to client-type users, port 80 is even more common, increasing from 51% in 2005 to 67% in 2009. Conversely, the ratio of dynamic ports has decreased from 36% to 18%. The second most common port was port 554. This is the port assigned to the Real-Time Streaming Protocol (RTSP), and is related to the increase in video content.

From this data, we can see that TCP traffic over port 80 is on the rise. Port 80 traffic is also used for data such as video content and software updates, so we cannot identify the type of content this is attributed to, but it demonstrates the fact that client/server communications are increasing.

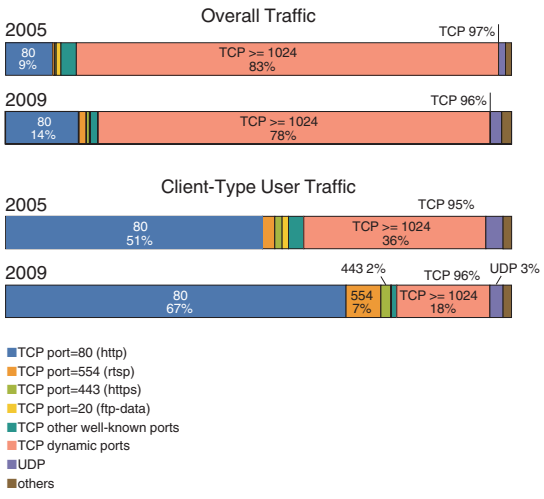

Figure 5: Usage Level Overview by Port

protocol port	2005		2009	
	total (%)	client type	total (%)	client type
TCP *	97.43	94.93	95.80	95.73
(<1024)	13.99	58.93	18.23	77.31
80 (http)	9.32	50.78	14.46	67.30
554 (rtsp)	0.38	2.44	1.48	6.89
443 (https)	0.30	1.45	0.64	1.91
20 (ftp-data)	0.93	1.25	0.19	0.17
(>=1024)	83.44	36.00	77.57	18.42
6346 (gnutella)	0.92	0.84	1.10	0.60
6699 (winmx)	1.40	1.14	0.70	0.24
1935 (rtmp)	0.20	0.81	0.36	1.51
7743 (winny)	0.48	0.15	0.25	0.03
UDP *	1.38	3.41	2.24	2.60
53 (dns)	0.03	0.14	0.03	0.07
ESP	1.09	1.35	1.87	1.55
GRE	0.07	0.12	0.07	0.08
ICMP	0.01	0.05	0.02	0.05

Table 1: Usage Level Details by Port

© 2009 Internet Initiative Japan Inc.

Figure 6 compares trends in TCP port usage over a week for overall traffic between 2005 and 2009. This shows trends for TCP port usage divided into three categories: port 80, other well-known ports, and dynamic ports. We cannot disclose absolute amounts of traffic, so we have presented data normalized by the total peak traffic volume. Dynamic port traffic is predominant overall, with peaks between 11:00 P.M. and 1:00 A.M., and traffic increases in the daytime on Saturday and Sunday, reflecting times when the Internet is used at home.

Similarly, Figure 7 shows trends in TCP port usage over a week for client-type users. This data indicates that although port 80 usage was slightly higher than dynamic port usage in 2005, in 2009 port 80 is the predominant form of traffic. Peak times are slightly earlier, occurring between 9:00 P.M. and 11:00 P.M., and use from the morning on Saturdays and Sundays has increased.

Comparing trends for total traffic and client-type traffic, we can see that there is a difference in the drop-off in traffic after midnight. Port 80 traffic drops off suddenly after midnight, and drops to its minimum level at around 4:00 A.M. In contrast, dynamic port traffic flow slopes off gradually in the early hours of the morning, dropping to its minimum level at around 8:00 A.M. We speculate that this is influenced by files being uploaded manually in the evening and distributed overnight via P2P file sharing applications that use dynamic ports, and by users who stop applications after they have finished downloading the files that they seek.

2.5 Conclusion

As we have observed, peer-type traffic such as P2P file sharing still dominates traffic from a volume perspective, but it has not increased significantly since 2005. One possible reason for this is that users have shifted from P2P file sharing applications to services such as video sharing sites that are easier to use and more popular. This may also be influenced by the fact that P2P file sharing mechanisms have been revised not to use excessive bandwidth, as a result of the rapidly increased traffic volumes from P2P file sharing being identified as a problem. Changes in user awareness due to ISPs introducing countermeasures against excessive usage such as bandwidth cap may have also contributed.

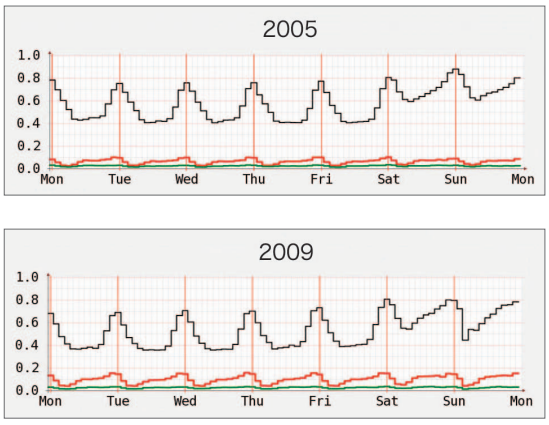

Figure 6: Weekly TCP Port Usage Trends for Overall Traffic 2005 (Top) 2009 (Bottom)

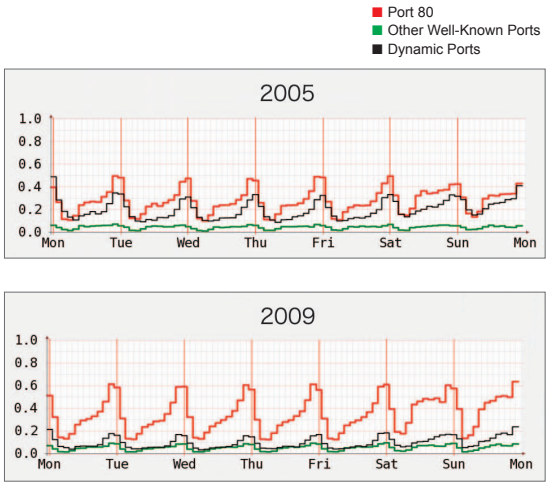

Figure 7: Weekly TCP Port Usage Trends for Client-Type Users 2005 (Top) 2009 (Bottom)

© 2009 Internet Initiative Japan Inc.

Meanwhile, usage levels for general users are steadily increasing due to rich video content and web 2.0 content. In addition to video content, there are also an increasing number of websites that automatically provide a variety of constantly changing information, or preload content in the background which a user may view next, without any explicit user action such as mouse clicks. This also leads to an increase in traffic volume.

Traffic increased dramatically around 2004 with the appearance of P2P file sharing, and this was expected to put strain on network capacity. Over the five years since then, traffic has been increasing at a stable annual rate of approximately 30%. Meanwhile, it is said that capacity for networks such as backbones has been increasing at an annual rate of approximately 50% (see reference 6). Because of this, it is believed that there is currently a surplus of bandwidth capacity on a macro level.

However, it is difficult to predict future Internet traffic based on past data. This is because the behavior of a small number of heavy users has a considerable impact, and when there is a change in this behavior, predictions can be wildly inaccurate. The way that users utilize the Internet is also influenced greatly by not only technological factors, but also by economic, social, and political ones. Additionally, just as the appearance of the web and P2P file sharing have caused an upheaval in traffic volumes, there is always the possibility of the appearance of new technology drastically changing the way the Internet is used. Considering that until now traffic volumes have undergone significant change in cycles of five or ten years, in a sense recent traffic growth has almost been too stable. Before too long we may find ourselves facing another upheaval.

IIJ monitors traffic levels on an ongoing basis so we can respond swiftly to changes in forms of Internet usage. We will continue to publish whitepapers such as this periodically.

Author:
Kenjiro Cho
Deputy Research Director, IIJ Innovation Institute Inc. Research Laboratory. Dr. Cho's research topics include the analysis of complex network dynamics in order to turn the Internet into a simpler, more flexible and more dependable communication infrastructure, QoS communications, and operating system support for networking. He is a board member of the WIDE Project, and an adjunct professor at Japan Advanced Institute of Science and Technology.

References

1: K. Cho, K. Fukuda, H. Esaki, and A. Kato.
The impact and implications of the growth in residential user-to-user traffic.
In ACM SIGCOMM2006, Pisa, Italy, Aug. 2006.
2: K. Cho, K. Fukuda, H. Esaki, and A. Kato.
Observing Slow Crustal Movement in Residential User Traffic.
In ACM CoNEXT2008, Madrid, Spain, Dec. 2008.
3: Cisco. Visual Networking Index - Forecast and Methodology, 2007-2012. June 2008.
4: Cisco. Approaching the zettabyte era. June 2008.
5: A.M. Odlyzko. Minnesota Internet traffic studies.
http://www.dtc.umn.edu/mints/home.html.
6: TeleGeography Research. Global Internet Geography. 2008.

© 2009 Internet Initiative Japan Inc.