Traffic Shifting away from P2P File Sharing to Web Services

Internet traffic volumes which had been increasing at a comparatively stable annual rate of approximately 30% until now, declined sharply by almost 20% in the beginning of January 2010. It is said that the reason for this was the amended Copyright Act that makes the download of copyright infringing content illegal. Here we will explore the cause of this drop in traffic by comparing traffic and port usage levels for the week starting May 24, 2010 with data for 2009.

3.1 Introduction

This report is a continuation of the first report in IIR Vol.4, and it summarizes the results of analysis of traffic over the broadband access services operated by IIJ. In the previous report, we reported that domestic and international Internet traffic volumes over the past five years were increasing at an annual rate of about 30%, and were comparatively stable. However, we also stated that it would be difficult to predict the future using previous data because traffic is greatly affected by the behavior of a sub-section of heavy users, and the usage may change due to non-technical elements such as social factors.

In fact, in January 2010 we observed a situation in which traffic volumes experienced minus growth. As shown in Figure 1, broadband traffic volume dropped by almost 20% in early January. There have been fluctuations in traffic volumes up until now, but there has never been a significant decline that continued for over six months like this. It is said that this decline was probably caused by the amended Copyright Act that came into effect in January 2010, making the download of copyright infringing content illegal. However, it was not expected to have such a large impact as no penalty was defined for illegal downloading.

In this report, as with the previous one, we examine the daily traffic volume of users and usage levels by port, and investigate what has actually changed compared to a year ago.

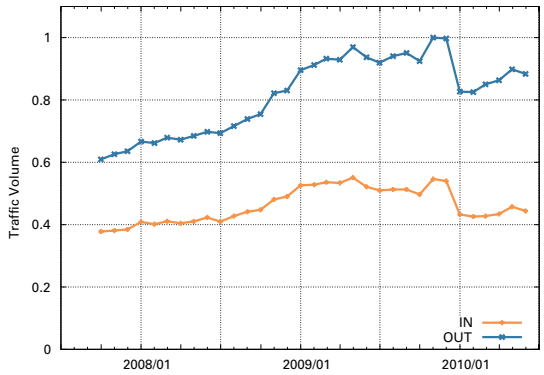

Figure 1: Broadband Traffic Volume Trends for the Past 3 Years (normalized to the OUT value for November 2009)

3.2 About the Data

Similar to the previous report, the survey data utilized here 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 differ between weekdays and weekends, we analyzed a full week of traffic. In this case, we used data for the week spanning May 24 to May 30, 2010. For comparison, we used the data we analyzed in the previous report for the week spanning May 25 to May 31, 2009.

The usage volumes 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 the performance and load of the routers. We estimated overall usage volumes by multiplying observed volumes by the reciprocal of the sampling rate. Because this kind of sampling method was used, there may be slight estimation errors in data for low-volume users. However, for users with reasonable usage levels we were able to obtain statistically meaningful data.

In 2005, approximately the same numbers of fiber-optic and DSL users were observed. However, the migration to fiber-optic connections advanced in the following years, with 85% of the observed users in 2010 now using fiber-optic connections, which represent 92% of the overall volume of traffic. The IN/OUT traffic presented in this report indicates directions from an ISP's perspective. IN represents uploads from users, and OUT represents user downloads.

3.3 Daily Usage Levels for Users

First, we will examine the daily usage volumes for broadband users from several perspectives. Daily usage indicates the average daily usage calculated by dividing each user's data for the period of a week by seven.

Figure 2 shows the complementary cumulative distribution of the daily traffic volume for users. This indicates the percentage of users with daily usage levels greater than the X axis value on a double logarithmic graph, which is an effective way of examining the distribution of heavy users in proportion to the whole. The right side of the graph falls linearly on the graph, showing a long-tailed distribution close to power-law distribution. It can be said that heavy users are distributed throughout the overall group, and are by no means a unique type of user.

Comparing this with the complementary cumulative distribution for 2009 shown in IIR Vol.4, the percentage of heavy users has dropped slightly for IN (upload). For example, the ratio of users that upload 10^8 (100 MB) or more in a day was 8.2% of the total in 2009, but this has declined to 6.5% in 2010. This is only a drop of 1.7% of total users, but when looking at the number of heavy users it represents a decrease of about 20%. Meanwhile, the right-hand edge of the graph is higher in 2010 than the distribution for 2009, indicating that the number of extremely heavy users has increased to the contrary.

Figure 3 indicates the deviation in traffic usage levels between users. This graph 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 78% of the total OUT (download) traffic, and 96% of the total IN (upload) traffic. Furthermore, the top 1% of users make up 33% of the total OUT (download) traffic, and 51% of the total IN (upload) traffic.

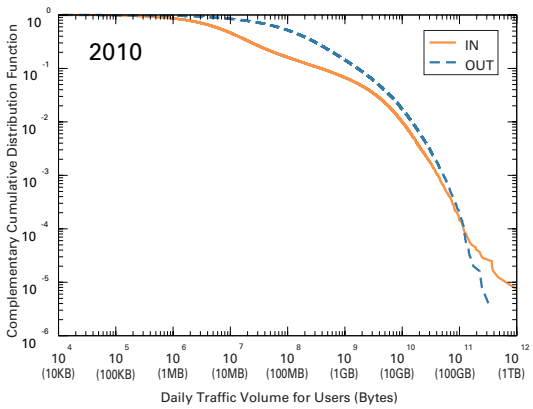

Figure 2: Complementary Cumulative Distribution of the Daily Traffic Volume for Users

Comparing this with the deviation in usage levels for 2009 shown in IIR Vol.4, the deviation for IN (upload) has increased. This is a result of the total number of heavy users decreasing, while the number of extremely heavy users has increased. For OUT (download) there is almost no change, but the deviation for the top 3% has decreased slightly.

However, these kinds of deviation trends are a characteristic of long-tailed distributions, and common in Internet data. For example, even when reexamining deviation after excluding users with the highest usage levels, 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 shows the average daily usage distribution (probability density function) per user. This is divided into IN (upload) and OUT (download), 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. In this survey the traffic volume for users with the highest usage levels climbed to 2 TB, and some users are outside the scope of the graph, but most fall within the scope of 10^{11} (100 GB). A slight spike appears on the left side of the graph, but this is just noise caused by the sampling rate.

The distribution for IN (upload) and OUT (download) 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 is further to the right than the IN distribution, indicating that the download volume is an order of magnitude larger than the upload volume. Average values are affected by the usage levels of heavy users on the right side of the graph, coming to 469 MB for IN (upload) volume and 910 MB for OUT (download) volume. In 2009 these figures were 556 MB and 971 MB respectively, so usage levels have decreased.

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

Comparing the most frequent values for client-type users in 2009 and in 2010, IN (upload) volume rose from 6 MB to 7 MB, and OUT (download) volume rose from 114 MB to 145 MB. This demonstrates that, particularly in the case of downloads, the traffic volume for each user has increased dramatically. On the other hand, there was a slight

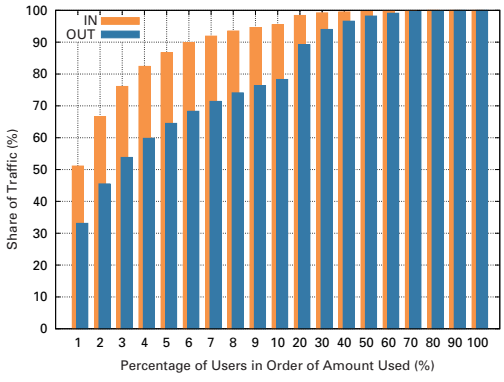

Figure 3: Traffic Usage Deviation Between Users

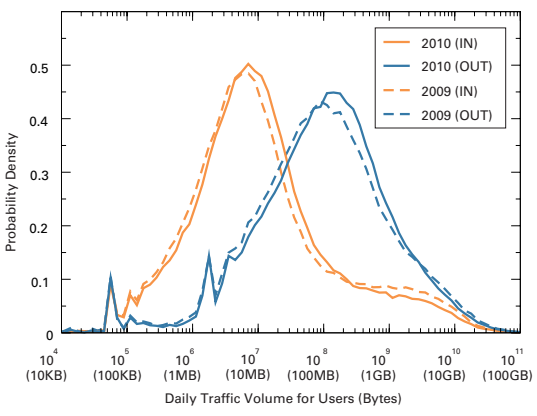

Figure 4: Daily User Traffic Volume Distribution

decrease in the probability density for peer-type users. In other words, it can be said that while usage levels for general users have increased steadily, usage levels for heavy users that account for the majority of volume have decreased, leading to a drop in the total volume of traffic.

Figure 5 plots the IN/OUT volumes for 5,000 randomly sampled users. The X axis shows OUT (download) and the Y axis IN (upload), expressed as a double logarithmic graph. When the IN/OUT traffic volumes for a user are identical, they are plotted on the diagonal line in the graph.

Two clusters can be observed. The cluster below the diagonal line and spread out parallel to it is client-type users with download volumes an order of magnitude higher than their upload volumes. The other cluster is peer-type users spread out around the diagonal line in the upper right. However, the boundary between these 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. These trends showed almost no change compared to those for 2009.

3.4 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 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, here 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 (upload) distributions in Figure 4, and users below the horizontal line at the IN = 100 MB point in Figure 5.

Figure 6 shows an overview of port usage, comparing all users and client-type users for 2009 and 2010. Table 1 shows detailed numeric values for this figure.

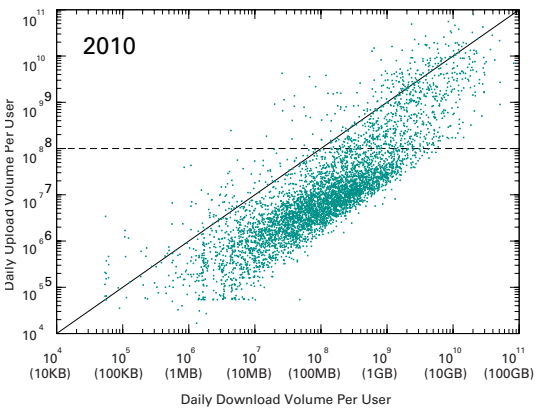

Figure 5: IN/OUT Usage for Each User

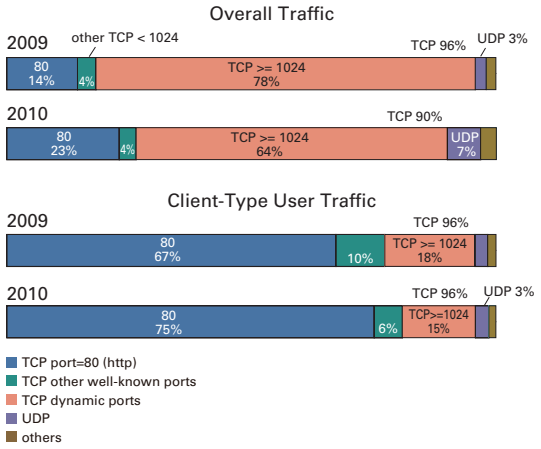

Figure 6: Usage Level Overview by Port

Over 90% of traffic for 2010 is TCP based. As for overall traffic, although the ratio of TCP dynamic ports higher than 1024 was 78% of the total volume in 2009, this has decreased sharply to 64% in 2010. Specific ports in the dynamic port range make up a small percentage, comprising 1% of the total traffic at most. Meanwhile, the use of port 80 has increased from 14% in 2009 to 23% in 2010. In other words, considering the decrease in traffic for 2010, we can surmise that communications between dynamic ports have dropped by approximately 25%, and about a third of this traffic has migrated to port 80.

When focusing on client-type users, use of port 80 that accounted for 67% in 2009 has increased to 75% in 2010. On the other hand, the ratio of dynamic ports has decreased from 18% to 15%.

From this data, we can see that TCP traffic over port 80 is still 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 is fair to say that client/server communications are increasing.

Figure 7 shows trends in TCP port usage over a week for overall traffic in both 2009 and 2010. This shows 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 ports are predominant among overall traffic, with peaks between 11:00 P.M. and 1:00 A.M. Traffic also increases in the daytime on Saturday and Sunday, reflecting times when the Internet is used at home. Comparing the 2009 and 2010 data, we can see that the percentage of port 80 traffic is increasing.

Figure 8, similarly to Figure 7, shows weekly trends in TCP port usage by client-type users. The ratio of part 80 traffic has also increased for 2010 here. Peak times are about two hours earlier than those indicated in the overall traffic usage in Figure 7, occurring between 9:00 P.M. and 11:00 P.M. Additionally, use from the morning on Saturdays and Sundays is increasing.

protocol port	2009		2010	
	total (%)	client type	total (%)	client type
TCP *	95.80	95.73	90.09	95.82
(<1024)	18.23	77.31	26.46	80.87
80 (http)	14.46	67.30	23.00	75.12
554 (rtsp)	1.48	6.89	1.15	2.45
443 (https)	0.64	1.91	0.98	2.28
20 (ftp-data)	0.19	0.17	0.18	0.07
(>=1024)	77.57	18.42	63.63	14.95
1935 (rtmp)	0.36	1.51	1.04	2.91
6346 (gnutella)	1.10	0.60	0.86	0.33
6699 (winmx)	0.70	0.24	0.65	0.17
8084	0.00	0.00	0.61	0.00
UDP	2.24	2.60	6.79	2.76
ESP	1.87	1.55	2.91	1.30
GRE	0.07	0.08	0.14	0.06
IP-IP	0.01	0.00	0.04	0.01
ICMP	0.02	0.05	0.02	0.04

Table 1: Usage Level Details by Port

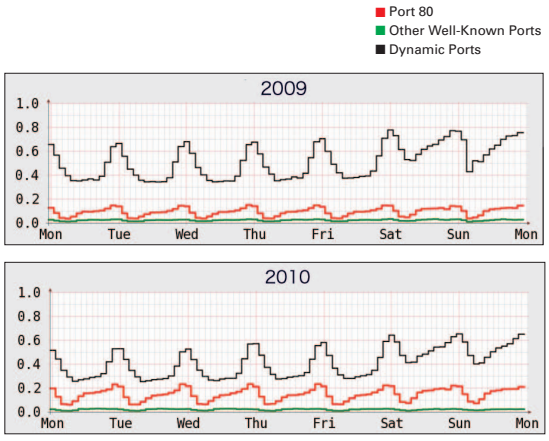

Figure 7: Weekly TCP Port Usage Trends

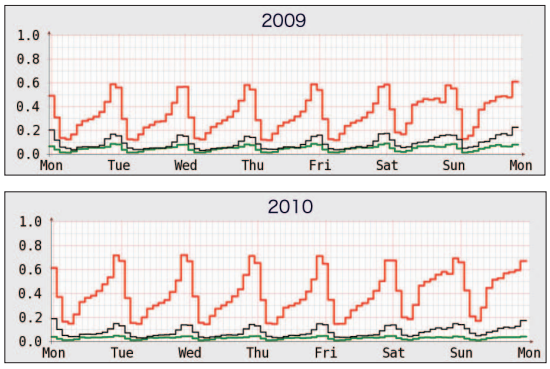

Figure 8: Weekly TCP Port Usage Trends for Client-Type Users

3.5 Conclusion

In our previous report, we identified that although peer-type traffic such as P2P file sharing still dominates traffic from a volume perspective, it has not increased much since 2005. We also stated that users are tending to shift away from P2P file sharing applications to services such as video sharing sites that are easier to use and more attractive. Meanwhile, we also showed that usage levels for general users are steadily increasing due to rich video content and web 2.0 content.

These trends remained unchanged in this survey, but it could be said that changes in the heavy user segment were much larger than in the past. However, this does not mean that heavy users and communications between dynamic ports have simply declined. The percentage of heavy users fell by about 20%, but on the other hand traffic volumes for extremely heavy users are increasing. Additionally, with regard to communications by port, traffic volumes between dynamic ports decreased by about 25%, but about a third of this volume migrated to port 80. This points to the fact that the migration of general users from P2P file sharing applications to web services that was noted in our last report is spreading to heavy users.

Consequently, regarding the decrease in traffic volume and the impact of the amended Copyright Act, rather than saying the decline in P2P file sharing is due to the amended Copyright Act, I believe we should consider that the existing trend of migration from P2P file sharing applications to web services has progressed further as a result of the amended Copyright Act. To use an analogy, landslides occurring in an earthquake may actually be caused because the ground was already unstable, with the earthquake merely acting as a trigger.

Traffic is shifting from P2P file sharing applications to web services all over the world (see references 2, 3, and 4). There have also been reports of a traffic decrease in other countries due to new regulations, legislation regarding copyright, or the arrest of infringers, similar to the amended Copyright Act case in this report. However, behind individual cases there are overall changes in the social recognition of copyright infringement risks, or in application usage as alternatives for P2P file sharing mature.

The decrease in traffic in early 2010 that we covered in this report is a phenomenon unique to Japan. In Japan, fiber-optic access is widespread and there is excess bandwidth, so the shares of heavy user traffic and P2P file sharing traffic are higher than other countries. As a result, changes in the behavior of heavy users have a significant effect on overall traffic.

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

References
1: Kenjiro Cho. Broadband Traffic: Increasing Traffic for General Users. Internet Infrastructure Review. Vol.4. pp 18-23. August 2009.
2: G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. IMC2009. Chicago, IL, November 2009.
3: C. Labovitz, D. McPherson, and S. Iekel-Johnson. 2009 Internet Observatory Report. NANOG47. Dearborn, MI. October, 2009.
4: Cisco. Visual Networking Index — Forecast and Methodology. 2009-2014. June 2010.

Author:
Kenjiro Cho
IIJ Innovation Institute Inc. Research Laboratory