The Impact and Implications of the Growth in Residential User-to-User Traffic

Kenjiro Cho
IIJ

Kensuke Fukuda
NII

Hiroshi Esaki
Univ. of Tokyo

Akira Kato
Univ. of Tokyo

ABSTRACT

It has been reported worldwide that peer-to-peer traffic is taking up a significant portion of backbone networks. In particular, it is prominent in Japan because of the high penetration rate of fiber-based broadband access. In this paper, we first report aggregated traffic measurements collected over 21 months from seven ISPs covering 42% of the Japanese backbone traffic. The backbone is dominated by symmetric residential traffic which increased 37% in 2005. We further investigate residential per-customer traffic in one of the ISPs by comparing DSL and fiber users, heavy-hitters and normal users, and geographic traffic matrices. The results reveal that a small segment of users dictate the overall behavior; 4% of heavy-hitters account for 75% of the inbound volume, and the fiber users account for 86% of the inbound volume. About 63% of the total residential volume is user-to-user traffic. The dominant applications exhibit poor locality and communicate with a wide range and number of peers. The distribution of heavy-hitters is heavy-tailed without a clear boundary between heavy-hitters and normal users, which suggests that users start playing with peer-to-peer applications, become heavy-hitters, and eventually shift from DSL to fiber. We provide conclusive empirical evidence from a large and diverse set of commercial backbone data that the emergence of new attractive applications has drastically affected traffic usage and capacity engineering requirements.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network monitoring*

General Terms

Measurement, Management

Keywords

ISP backbone traffic, Residential broadband, Traffic growth

1. INTRODUCTION

Over the past few years, an unprecedented increase in user-to-user traffic has been observed worldwide, particularly in Japan due to its high penetration rate of fiber-based broadband access. Figure 1 depicts the traffic growth in Japanese backbones in terms of aggregated peak traffic at major IXes: JPIX[11], JPNAP[12], and NSPIXP[18].

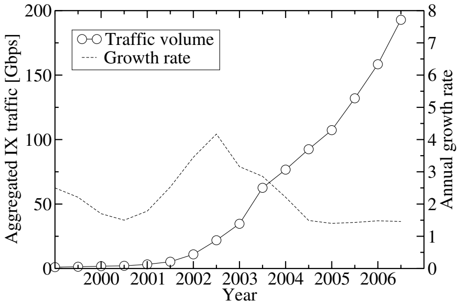

Figure 1: Traffic growth of the aggregated peak rate at the major Japanese IXes

Although a large part of the traffic increase on commercial backbones is often attributed to peer-to-peer traffic, there is little work in literature with statistics detailed enough to prove it. It is also difficult to plan for the future because residential access and its traffic are undergoing a transformation; new innovations in access networking technologies continue to be developed, and new applications as well as new usage of older applications are emerging to take advantage of low-cost high-speed connectivity. Japan leads other countries by far in Fiber-To-The-Home (FTTH) penetration [20], where the number of FTTH subscribers is increasing exponentially while the increase in DSL subscribers is slowing down as shown in Figure 2 [29]. (FTTH includes Fiber-To-The-Building with VDSL for in-building wiring.)

There is a strong concern that, if this trend continues, Internet backbone technologies will not be able to keep up with the rapidly-growing residential traffic. Moreover, commercial ISPs will not be able to invest in backbone networks simply for supporting this low-profit customer segment.

In order to ensure the evolution of the Internet, it is essential that we understand the effects of growing residential traffic, but it is difficult both technically and politically to obtain traffic data from commercial ISPs. Most ISPs collect traffic data which contains sensitive information, and thus seldom make it available to others. In addition, measure-

SIGCOMM'06, September 11–15, 2006, Pisa, Italy.

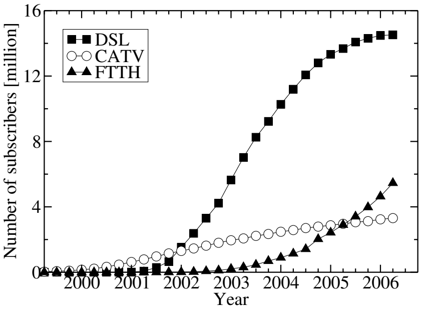

Figure 2: Increase of residential broadband subscribers in Japan: 23.3 million total broadband subscribers, 14.5 million for DSL, 3.3 million for CATV, and 5.5 million for FTTH as of March 2006.

ment methods and policies differ from ISP to ISP so that it is in general not possible to compare a data set with another set obtained from a different ISP.

Seeking a practical way to investigate the impact of residential broadband traffic on commercial backbone networks, we formed a study group with specialists including members from seven major Japanese commercial ISPs in order to identify the macro-level impact of residential broadband traffic on ISP backbones [6]. Our goal is to obtain a clearer grasp of the ratio of residential broadband traffic to other traffic, changes in traffic patterns, and regional differences across ISPs.

We have collected aggregated bandwidth usage logs for several categories of traffic. The results reveal that the backbone is dominated by symmetric residential traffic which increased 37% in 2005. The peak hours have shifted from office hours to evening hours, and a considerable amount of traffic is constantly flowing.

Using these statistics as reference points, we have performed further analyses of residential traffic data provided by one of the ISPs. The results reveal surprisingly diverse behavior of residential traffic.

2. DATA COLLECTION

Our data sets were collected using two different methods. The first set was collected by aggregating interface counters of edge routers from seven ISPs for analysis of residential traffic at a macro-scopic level. The other set was collected by Sampled NetFlow [2] from one of the ISPs for detailed per-customer analysis.

2.1 Data Collection of Aggregated Traffic

We found that most ISPs collect interface counter values of almost all routers in their service networks via SNMP, and archive per-interface traffic logs using MRTG [22] or RRDtool [21]. Thus, it is possible for the ISPs to provide aggregated traffic information if they can classify router interfaces into a common set.

There are several requirements in order to solicit ISPs to divulge traffic information. We need to find a common data set which all the participating ISPs are able to provide with moderate workload and investment. The data set should be coarse enough not to reveal sensitive information about the ISP but be meaningful enough so that the behavior of residential broadband traffic can be analyzed. The data sets should be aggregated with those provided by other ISPs so that the share of each ISP is not revealed.

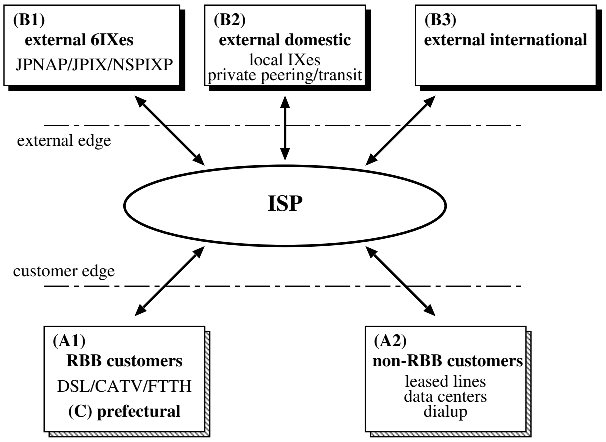

Figure 3: Five traffic groups for data collection at ISP customer and external boundaries

Our focus is on traffic crossing ISP boundaries which can be roughly divided into customer traffic, and external traffic such as peering and transit. For practical purposes, we selected the five traffic groups in Figure 3 for data collection.

(A1) RBB customers represent residential broadband customer lines. This group also includes small business customers using residential broadband access.

(A2) non-RBB customers represent customer lines other than RBB customers, including leased lines, data centers, and dialup lines. This group includes RBB customers behind leased lines, e.g., second or third level ISPs, since ISPs do not distinguish them from other leased lines.

(B1) external 6IXes represent links for 6 major IXes, namely JPIX, JPNAP and NSPIXP in both Tokyo and Osaka in order to compare measurements at these IXes as well as to know the traffic share of our measurement.

(B2) external domestic represents domestic external links other than the 6IXes, including regional IXes, private peering and transit. We used the term "domestic" to indicate that both ends of a link are located in Japan. This group also includes domestic peering with global ASes.

(B3) external international represents international external links with one end point outside of Japan.

(C) prefectural represents RBB links categorized into 47 prefectures in Japan. This group is a subset of (A1), and covers two major residential broadband carriers who provide aggregated links per prefecture to ISPs. Other RBB carriers without prefectural information are not used for this group.

It is impossible to draw a strict line for grouping, e.g., residential/business and domestic/international, on the global Internet, so these groups are chosen by the existing operational practices of the participating ISPs. We re-aggregate each ISP's aggregated logs, and only the resulting aggregated traffic is used in our study so as to not reveal the share of each ISP.

Our main focus is on (A1), *RBB customers*, but we examine the other categories to understand the relative volume of (A1) with respect to other types of traffic as well as to cross-check the correctness of our results. (A2), *non-RBB customers*, is used to obtain the ratio of residential broadband traffic to total customer traffic. The total customer traffic (A) is $(A) = (A1) + (A2)$. (B1), *external 6IXes*, and (B2), *external domestic*, are used to estimate the coverage of the collected data sets. (B3), *external international*, is used to compare domestic traffic with international traffic. The total external traffic (B) is $(B) = (B1) + (B2) + (B3)$. (C), *prefectural*, is used to measure regional differences.

In general, it is meaningless to simply sum up traffic values from multiple ISPs since a packet could cross ISP boundaries multiple times. Customer traffic is, however, summable because a packet crosses customer edges only once in each direction, when entering the source ISP and exiting the destination ISP. The numbers for external traffic are overestimated since a packet could be counted multiple times if it travels across intermediate ISPs other than ingress and egress ISPs. However, the error should be negligible in this particular result since the ISPs in our data sets are peering, and thus, not providing transit to each other.

We collected month-long traffic logs from the participating ISPs. The collected logs have a time resolution of two hours since it was the highest common factor for month-long data. This is because both MRTG and RRDtool aggregate old records into coarser records in order to bound the database size. In MRTG, 2-hour resolution records are maintained for 31 days in order to draw monthly graphs. RRDtool does not have fixed aggregation intervals but it is most likely that RRDtool is configured to maintain 1-hour or 2-hour resolution records needed for monthly graphs. Although the peak rate is often used for operational purposes, only the mean rate is collected since the peak rate is not summable.

We developed a perl script to read a list of MRTG and RRDtool log files, and to aggregate traffic measurements for a given period at a given resolution. It outputs "timestamp, in-rate, out-rate" for each time step. Another script produces a graph using RRDtool. We provided the tools to the ISPs so that each ISP could create aggregated logs by themselves. This allows ISPs not to disclose the internal structure of their network or unneeded details of their traffic.

The highest workload for the ISPs is to classify a large number of per-interface traffic logs and create a log list for each group. For large ISPs, the number of existing per-interface traffic logs can exceed 100,000. To reduce the workload, ISPs are allowed to use the internal interface of a border router instead of a set of external (edge) interfaces if the traffic on the internal interface is an approximation of the sum of the external interfaces. In this case, we instruct the tool to swap "in" and "out" records since the notation in the per-interface logs depicts the perspective of the routers but inbound/outbound records in our data sets signify the ISPs' point of view.

We analyzed month-long traffic logs from seven major Japanese ISPs six times over 21 months; September, October, November in 2004, May and November in 2005, and May in 2006. For data analysis, we focus on the data sets in 2005 to compare them with the per-customer data even though the latest set is from May 2006. We collected the data separately for each month, and checked consistency in each ISP's share, differences from the previous measurements, the coverage of the IX traffic, and others. We also provide the aggregated results to the ISPs so that each ISP can compare and check its own data against the aggregated results. Thus, we are fairly confident about the accuracy of the results. After the initial trials over four months (the first results from August 2004 had errors and have never been published), we decided to collect data only twice a year to reduce the workload of the participating ISPs.

Monthly traffic logs with two-hour resolution allow us to identify major changes in each ISP's traffic. When such changes are found, we contact the ISP to confirm the cause of the change, e.g., a network reconfiguration, an outage, missing SNMP data, or a mis-classification of interface counter logs. Afterwards, if necessary, we ask the ISP for corrected data. We indeed found short periods of missing SNMP data a few times but their impacts to the monthly average were less than 1%. Once, we found a mis-classification and asked for corrected data.

2.2 Data Collection of Per-Customer Traffic

In order to further analyze the behavior of residential traffic, we obtained Sampled NetFlow data from one of the participating ISPs. This ISP has residential broadband customers over DSL and fiber but not over CATV. Data was collected from all edge routers accommodating residential broadband customers. The sampling rate used was 1/2048 so as to not overload the routers. We believe it is enough for analyzing heavy-hitters but there is a certain amount of sampling error, especially for lightweight users. The traffic volume is derived by multiplying the measured volume by the sampling rate.

A week-long data set was collected five times: April, May, October in 2004, February and July in 2005. In this paper, we use only the two sets from February and July 2005 to focus on traffic in 2005.

Data from February 2005 was used to analyze per-customer behavior in Section 4.1 through 4.3 by matching customer IDs with the assigned IP addresses. The ISP provided the inbound/outbound traffic volume of each customer in one hour resolution as well as customer's attributes: the line type (DSL or fiber), and the prefecture.

Data from July 2005 was used to analyze geographic communication patterns from Section 4.4 through 4.5. In our data, one end of a flow is always the residential customer of the ISP but the other end is generally a customer of another ISP. Therefore, it is not possible to classify both ends by the ISP's information alone. For this reason, we used two geo-IP databases, Cyber Area Research Inc's SUTF-POINT and Digital Envoy's Netacuity, to classify both ends of the flows. The former database maps the address blocks of domestic residential customers to prefectures, but it does not cover non-residential addresses such as data-centers and leased-lines. The advantage of using this database is that we can distinguish residential users from other domestic users. The addresses not covered by the former database are classified simply into *domestic* and *international* by the latter database. Thus, *domestic* corresponds mainly to data-centers and leased lines in Japan, but it also includes residential address blocks not listed in the former geo-IP database.

3. ANALYSIS OF AGGREGATED TRAFFIC

The results were obtained by aggregating all traffic logs provided by the seven ISPs. Each ISP provided month-long traffic logs with 2-hour resolution. Both MRTG and RRDtool compute 2-hour boundaries in UTC so that the boundaries fall on odd hours in Japanese Standard Time (UTC+9). Throughout the paper, *inbound* and *outbound* are presented from the ISPs' point of view.

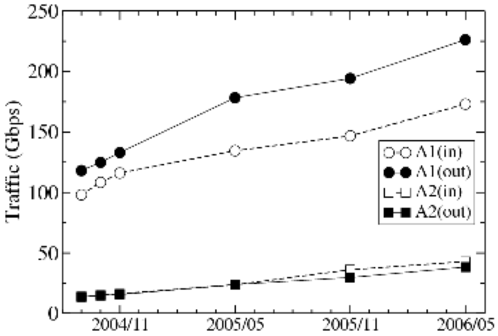

Figure 4: Growth of customer traffic: (A1) RBB customer and (A2) non-RBB customer

3.1 Growth of Traffic

The monthly average rates in bits/second of the traffic groups are shown in Tables 1 through 4. Table 1 shows the average rates of aggregated customer traffic, and the traffic growth is shown in Figure 4. Between November 2004 and November 2005, the growth rate of the RBB customer traffic (A1) was 26% for inbound, 46% for outbound, and 37% for the combined volume. The difference between inbound and outbound slightly widened in the first 6 months. The data for non-RBB customer traffic was obtained only from the four ISPs; it is difficult for the other ISPs to distinguish external links from other links due to historical reasons. Since (A2) from these other ISPs is missing, it is not possible to directly compare (A1) with (A2). Thus, we estimated the ratio of (A1) to (A) using only data from the 4 ISPs with both (A1) and (A2). The estimated ratio (A1)/(A1+A2) was 59% for inbound and 64% for outbound in November 2005.

Table 1: Monthly average rates of aggregated customer traffic

		(A1)customer-RBB (7 ISPs)		(A2)customer-non-RBB (4 ISPs)	
		inbound	outbound	inbound	outbound
2004	Sep	98.1G	111.8G	14.0G	13.6G
	Oct	108.3G	124.9G	15.0G	14.9G
	Nov	116.0G	133.0G	16.2G	15.6G
2005	May	134.5G	178.3G	23.7G	23.9G
	Nov	146.7G	194.2G	36.1G	29.7G
2006	May	173.0G	226.2G	42.9G	38.3G

Table 2 summarizes the average rates of aggregated external traffic, and the traffic growth is shown in Figure 5. It is observed that the total volume of external domestic traffic (B2), mainly private peering, exceeds the volume for the six major IXes (B1). From this result, it would be misleading to simply rely on data from IXes to estimate and understand nation-wide traffic, because a considerable amount of traffic is exchanged by private peering. At the same time, it is possible that the ratio of private peering is overestimated for the rest of the Japanese ISPs because private peering is usually exercised only between large ISPs. The ratio of international traffic to the total external traffic was 30% for inbound and 26% for outbound in November 2005.

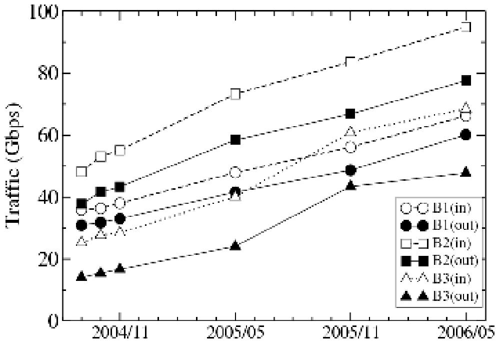

Figure 5: Growth of external traffic: (B1) 6 major IXes, (B2) other domestic and (B3) international

Table 2: Monthly average rates of aggregated external traffic

		(B1)ext-6ix (7 ISPs)		(B2)ext-dom (7 ISPs)		(B3)ext-intl (7 ISPs)	
		in	out	in	out	in	out
2004	Sep	35.9G	30.9G	48.2G	37.8G	25.3G	14.1G
	Oct	36.3G	31.8G	53.1G	41.6G	27.7G	15.4G
	Nov	38.0G	33.0G	55.1G	43.3G	28.5G	16.7G
2005	May	47.9G	41.6G	73.3G	58.4G	40.1G	24.1G
	Nov	54.0G	48.1G	80.9G	68.1G	57.1G	39.8G
2006	May	66.2G	60.1G	94.9G	77.6G	68.5G	47.8G

Table 3 shows a relationship between the total customer traffic (A) and the total external traffic (B). If we assume all inbound traffic from other ISPs is destined to customers, the inbound traffic volume for the total external traffic (B) should be close to the outbound traffic volume for the total customer traffic (A). Similarly, the outbound traffic volume (B) should be close to the inbound traffic volume (A). However, non-RBB customer data is provided by only 4 ISPs. If we interpolate the missing ISPs in the non-RBB customer traffic using the ratio from the four reporting ISPs, the total inbound and outbound customer traffic for November 2005 is estimated to be 248.4Gbps and 304.4Gbps, respectively. These figures are higher than those for the total external traffic, and the difference is considered to be customer traffic whose source and destination belong to the same ISP.

Table 3: Monthly average rates of total customer traffic and total external traffic

		(A)customer(A1+A2)		(B)external(B1+B2+B3)	
		inbound	outbound	inbound	outbound
2004	Sep	112.1G	125.4G	109.4G	82.8G
	Oct	123.3G	139.8G	117.1G	88.8G
	Nov	132.2G	148.6G	121.6G	93.0G
2005	May	158.2G	202.2G	161.3G	124.1G
	Nov	182.8G	223.9G	192.0G	156.0G
2006	May	215.9G	264.5G	229.6G	185.5G

Last, we examined the relationship between our IX traffic data (B1) and the total input rate of the six major IXes, as obtained directly from these IXes [29]. In comparison with the published total incoming traffic of these IXes, our data consistently represents about 42% of the total traffic as shown in Table 4. If we assume this ratio to be the traffic share of the seven ISPs, the total amount of residential broadband traffic in Japan in November 2005 can roughly be estimated to be 353Gbps for inbound and 468Gbps for outbound. We are not aware of any data from other countries against which to compare these numbers.

Table 4: IX traffic observed at ISPs and IXes

		(B1)ext-6ix outbound	6 major IXes inbound	ratio (%)
2004	Sep	30.9G	74.5G	41.5
	Oct	31.8G	77.1G	41.2
	Nov	33.0G	80.3G	41.1
2005	May	41.6G	99.1G	42.0
	Nov	48.1G	115.9G	41.5
2006	May	60.1G	139.2G	43.2

3.2 Customer Traffic

Figure 6 shows weekly customer traffic. For weekly data analysis, we took the averages of the same weekdays in a month. We excluded holidays from the weekly analysis since holiday traffic patterns are closer to those of weekends.

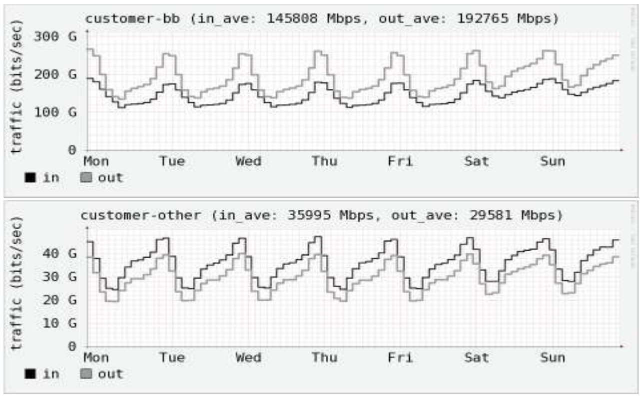

Figure 6: Aggregated weekly traffic of RBB customers (top) and non-RBB customers (bottom) in November 2005. Darker vertical lines indicate the start of the day (0:00 am in local-time).

The top graph shows RBB customers, (A1), consisting of DSL/FTTH/CATV residential users. The residential broadband customer traffic already exceeds 260Gbps in evening hours. The inbound and outbound traffic volumes are almost equal, and about 120Gbps is constantly flowing in both directions, probably due to peer-to-peer applications which generate traffic independent of daily user activities. The diurnal pattern indicates that home user traffic is dominant, i.e., the traffic increases in the evening, and the peak hours are from 21:00 to 23:00. Weekends can be identified by larger daytime traffic although the peak rates are similar to weekdays'. The outbound traffic to customers is slightly larger than the inbound, even though it is often assumed that home users' downstream traffic is much larger than upstream. We believe that peer-to-peer applications contribute significantly to the upstream traffic as we will see in Section 4. Figure 7 compares the RBB customer inbound traffic in November 2004 and November 2005. The overall increase appears to be derived from the growth of the constantly flowing traffic.

The bottom graph in Figure 6 shows the weekly traffic of non-RBB customers (A2). Since this group also includes leased lines used to accommodate second or third level ISPs, the traffic pattern still appears to be dominated by residential traffic, which is indicated by the peak hours and the differences between weekdays and weekends. However, we also observe office hour traffic (from 8:00 to 18:00) in the daytime on weekdays but traditional office customer traffic is smaller than residential customer traffic. The traffic patterns common to both graphs in Figure 6 are different from well-known academic or business usage patterns in which the peak is found during office hours [4, 5, 25].

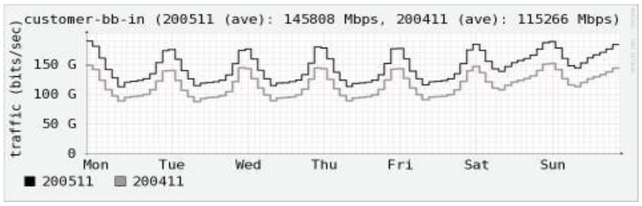

Figure 7: Growth of inbound traffic of RBB customers between November 2004 and November 2005

3.3 External Traffic

The external traffic groups are used to understand the total traffic volume in backbone networks. The top graph in Figure 8 shows traffic to and from the six major IXes (B1). It is apparent that the traffic behavior is strongly affected by residential traffic.

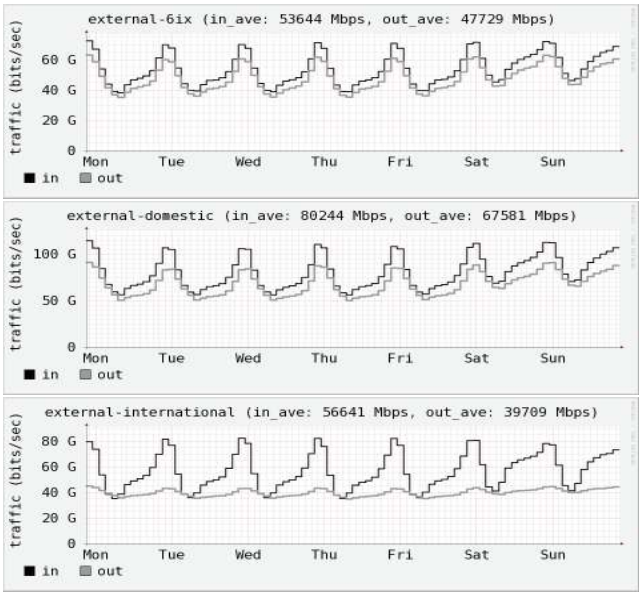

Figure 8: Aggregated weekly traffic for the 6 major IXes (top), other domestic (middle) and international (bottom) in November 2005

The middle graph in Figure 8 shows the external domestic traffic (B2) including regional IXes, private peering and transit but not including traffic for the six major IXes. The traffic pattern is very similar to the top graph.

The bottom graph in Figure 8 shows international traffic (B3). The inbound traffic is much larger than the outbound, and the traffic pattern is clearly different from the domestic traffic. The peak hours are still in the evening, but outbound traffic volume is virtually flat when compared to the inbound volume, suggesting that the traditional behavior of

Japanese users downloading content from overseas is still a non-negligible part of international traffic. At the same time, the constant part is about 70% of the average inbound rate, which could be due to machine-generated traffic (e.g., peer-to-peer file-sharing).

3.4 Prefectural Traffic

In order to investigate regional differences, i.e., between metropolitan and rural areas, we collected regional traffic data of the 47 prefectures. Figure 9 illustrates aggregated traffic of one metropolitan prefecture (top) and of one rural prefecture (bottom). Both graphs exhibit similar temporal patterns such as peak positions and weekday/weekend behavior. In addition, about 70% of the average traffic is constant regardless of the traffic volume. These characteristics are common to other prefectures. One noticeable difference is that metropolitan prefectures experience larger volumes of office hour traffic, probably due to larger business usage.

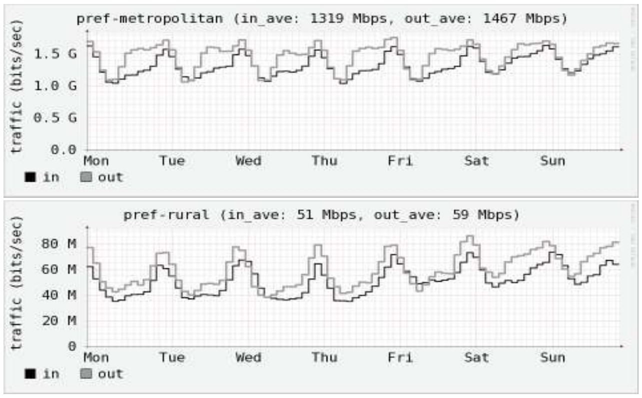

Figure 9: Example prefectural traffic: a metropolitan prefecture (top) and a rural prefecture (bottom)

Figure 10 plots traffic volumes and populations for the 47 prefectures. We found that a prefecture's traffic is roughly proportional to the population of the prefecture. The result indicates that there is no clear regional concentration of heavy-hitters of the Internet. That is, the probability of finding a heavy-hitter in a given population is constant and the distribution of aggregated traffic volume directly depends on the population. A possible reason is the availability of fairly universal access services in Japan; 100Mbps fiber access is available in most areas.

4. ANALYSIS OF PER-CUSTOMER TRAFFIC

This section analyzes Sampled NetFlow data from one of the ISPs. By comparing the aggregated traffic graphs in the previous section with the ISP's corresponding graphs, we can say that the traffic characteristics are consistent. Thus, although the data sets are from only one ISP, the results are likely to represent Japanese residential traffic.

Table 5 shows the number of unique active users identified by customer IDs in the February data set. As we explain later, users are classified into two groups by average daily inbound traffic, one of more than 2.5 GB/day and the other of less than 2.5GB/day. The total number of active users of DSL is slightly higher than fiber, but there are more heavy-hitters among fiber users.

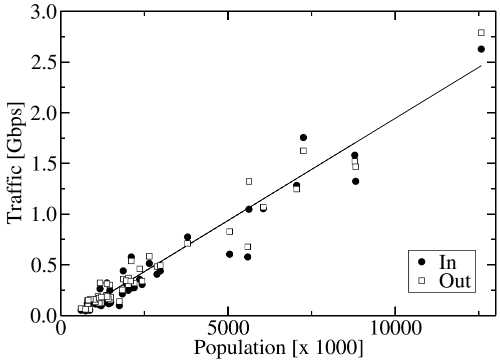

Figure 10: Prefectural traffic volumes are roughly linear to populations

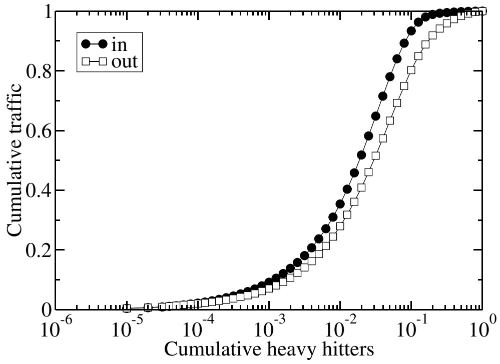

Figure 11: Cumulative distribution of traffic volume of heavy-hitters in decreasing order of volume

Table 5: Ratio of fiber and DSL active users in the February 2005 data set

	ratio (%)	$\geq 2.5GB/day$ (%)	$< 2.5GB/day$ (%)
total	100	4.46	95.54
fiber	46.45	3.66	42.79
DSL	53.55	0.80	52.75

4.1 Distribution of Heavy-hitters

Figure 11 shows the cumulative distribution of the total traffic volume of heavy-hitters in decreasing order of volume. The distribution is computed independently for inbound and outbound traffic. The graph reveals a skewed traffic distribution among users; the top N% of heavy-hitters use X% of the total traffic. For example, the top 4% use 75% of the total inbound traffic, and 60% of the outbound. In other words, a small group of heavy-hitters represent a significant part of the total traffic.

Figure 12 shows the (complementary) cumulative distribution of daily traffic per user on a log-log scale, and compares the total users (top) with the fiber users (middle) and the DSL users (bottom). The daily traffic volume is the average for the week, and the distribution is computed independently for inbound and outbound traffic.

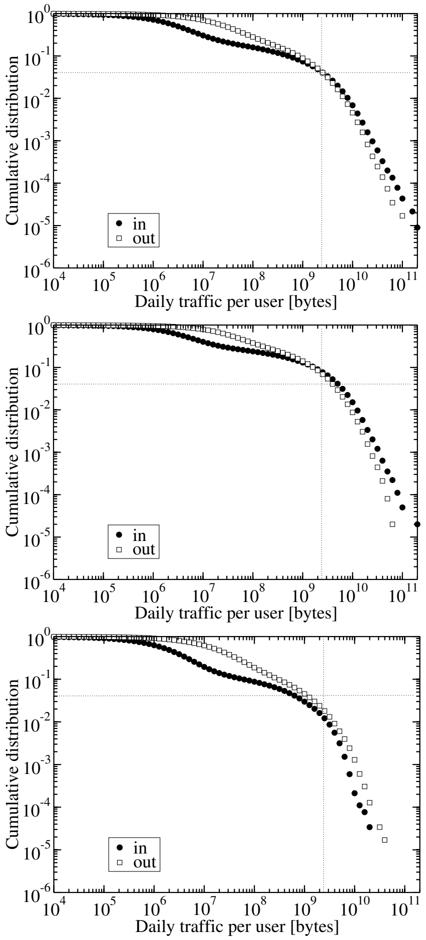

Figure 12: Cumulative distribution of daily traffic per user: total users (top), fiber users (middle) and DSL users (bottom). The lines are drawn at 2.5GB/day and the top 4% heavy-hitters, the knee of the total users' slope.

The distributions is heavy-tailed but there is a knee in the slope, at the top 4% of heavy-hitters using more than 2.5GB/day (or 230kbits/sec) for the total users, and at the top 10% using more than 2.5GB/day for the fiber users. It is less clear for the DSL users, but a knee can be seen at around the top 2% using more than 2.5GB/day. The distribution also shows that outbound traffic is larger for the majority of the users on the left side of the knee but it does not hold for heavy-hitters on the right side of the knee.

The distribution has a different slope for those who upload more than 2.5GB/day so we use this figure to statistically distinguish heavy-hitters from the rest of the users. We classify users who upload more than an average of 2.5GB/day to be in the *heavy-hitter* group, and those who upload less than 2.5GB/day to be in the *normal user* group. The *normal user* group should be interpreted as users other than the most influential heavy-hitters. Note that the difference is only in the slope of the distribution, and the boundary between the two groups is not clear. In other words, users are distributed statistically over a wide traffic volume range, even up to the most extreme heavy-hitters. A concave curve is not found in the plots denoting that there is no significant gap in traffic usage among users.

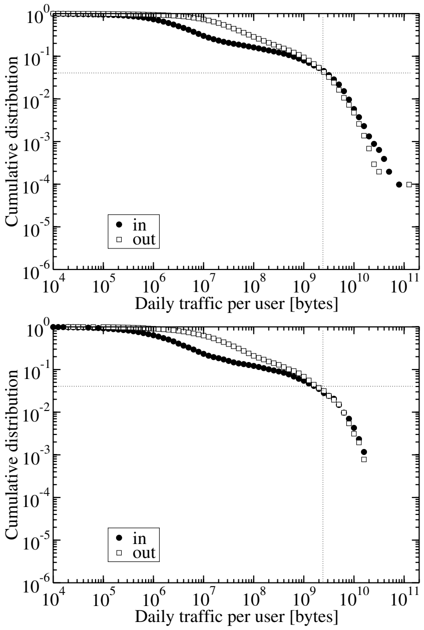

Figure 13: Cumulative distribution of daily traffic per user: a metropolitan prefecture (top) and a rural prefecture (bottom)

As for prefectural differences, the distributions look similar across different prefectures as shown in Figure 13 which compares one metropolitan prefecture (top) with one rural prefecture (bottom). One difference is the tail length due to the difference in the number of users. Another difference is that the distribution of the metropolitan prefecture is closer to that of the total users, and the distribution of the rural prefecture is closer to that of the DSL users. The results indicate that the distribution of heavy-hitters is similar across different regions with slight differences in the ratio of heavy-hitter population due to the ratio of fiber users with a larger heavy-hitter population.

4.2 Correlation of Inbound and Outbound Volumes

The correlation between inbound and outbound volumes for each user is shown as log-log scatter plots in Figure 14. These are taken from a metropolitan prefecture and plot about 4300 points for fiber and about 5400 for DSL but the characteristics are common to all the prefectures.

There is a positive correlation as expected, and the highest density cluster is below and parallel to the unity line where outbound volume (downstreaming for users) is about ten times larger than that of inbound. In a higher volume region, a different cluster appears to exist around the unity line. The slope of the cluster seems to be slightly larger than 1, which explains the inversion of inbound and outbound traffic volumes in Figure 12. It can be also observed that, across the entire traffic volume range, the inbound/outbound traffic ratio varies greatly, up to 4 orders of magnitude.

Both fiber and DSL plots show similar distributions but, as expected, the high-volume cluster is larger in the fiber

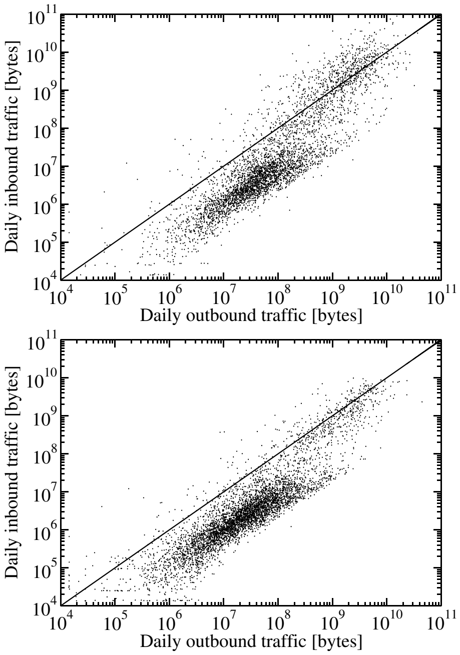

Figure 14: Correlation of inbound and outbound traffic volumes per user in one metropolitan prefecture: fiber (top) and DSL (bottom)

plot, especially above the unity line. A plausible interpretation of excess upstream traffic of the fiber heavy-hitters is that available bandwidth in fiber access is used to compensate for the shortage of upstream bandwidth of DSL heavy-hitters. It is also noticeable that there are much more low-volume users in the DSL plot. However, the boundary of the two clusters is not very clear. There seems to be no clear qualitative difference in the behaviors of fiber and DSL users except the percentage of heavy-hitters.

4.3 Temporal Behavior

Figure 15 and Figure 16 compare the temporal behaviors of the fiber users and the DSL users. The volume is normalized to the peak value of the total traffic size so as to not reveal the absolute traffic volume of the ISP. The graphs are shown in the same scale to compare fiber and DSL volumes.

The plots show that the inbound and outbound volumes are almost equal for fiber traffic but the inbound is 61% larger for heavy-hitters and the outbound is 166% larger for the normal users. The total is counterbalanced by the two groups. In the DSL traffic, the outbound volume is 83% larger for the total users, only 11% larger for the heavy-hitters and 179% larger for the normal users. The total reflects the offset of the normal users.

The inbound traffic of the fiber heavy-hitters is much larger than the outbound traffic, and has large daily fluctuations. On the other hand, the inbound traffic of DSL heavy-hitters is saturated. As a result, the fiber traffic accounts for 86% of the total inbound volume and 80% of the total residential volume, and the behavior of the total traffic is heavily influenced by the fiber heavy-hitters.

Figure 17 compares the temporal change in the number of active users in fiber and DSL. Again, the active user numbers are normalized to the peak value of the total active users. The number of active users is fairly constant for the heavy-hitters, especially for DSL. The constant portion seems to be users running automated data-transfer software. The increase of active users in the morning is larger than that of traffic volume but the increase in the evening is smaller, which suggests that bandwidth use is more intense, i.e., higher bandwidth demand per user, in the evening.

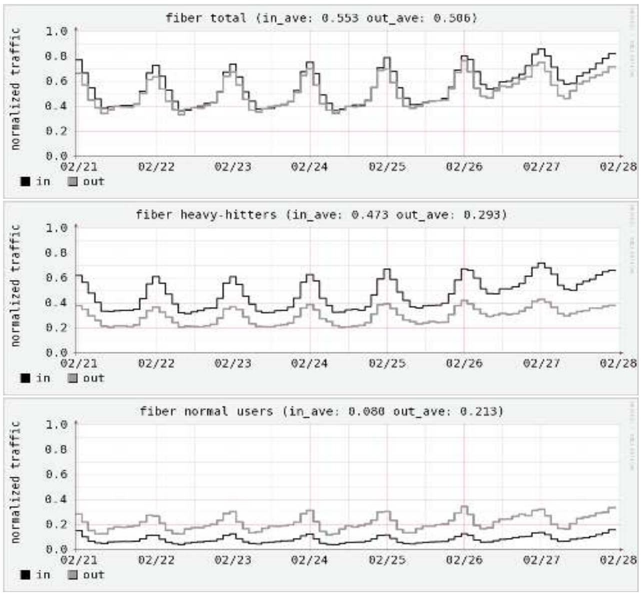

Figure 15: Fiber weekly traffic: total fiber users (top), heavy-hitters (middle) and normal users (bottom)

4.4 Protocol and Port Usage

Table 6: Protocol breakdown: TCP dynamic ports account for 83% of the total traffic

protocol	port	name	(%)	port	name	(%)
TCP	*		**97.43**			
	(< 1024		13.99)	81	-	0.15
	80	http	9.32	25	smtp	0.14
	20	ftp-data	0.93	119	nntp	0.13
	554	rtsp	0.38	21	ftp	0.11
	443	https	0.30	22	ssh	0.09
	110	pop3	0.17		others	2.27
	(>= 1024		83.44)	1935	macromedia-fsc	0.20
	6699	winmx	1.40	1755	ms-streaming	0.20
	6346	gnutella	0.92	2265	-	0.13
	7743	winny	0.48	1234	-	0.12
	6881	bittorrent	0.25	4662	edonkey	0.12
	6348	gnutella	0.21		others	79.41
UDP	*		**1.38**	6257	winmx-	0.06
	6346	gnutella	0.39		others	0.93
ESP			**1.09**			
GRE			**0.07**			
ICMP			**0.01**			
others			**0.02**			

Table 6 shows the ranking of protocols and ports. To rank port numbers in TCP and UDP, we took the smaller of the source and destination ports for a flow. TCP ports are further divided into well-known ports that are smaller than 1024, and dynamic ports that are equal to or larger than

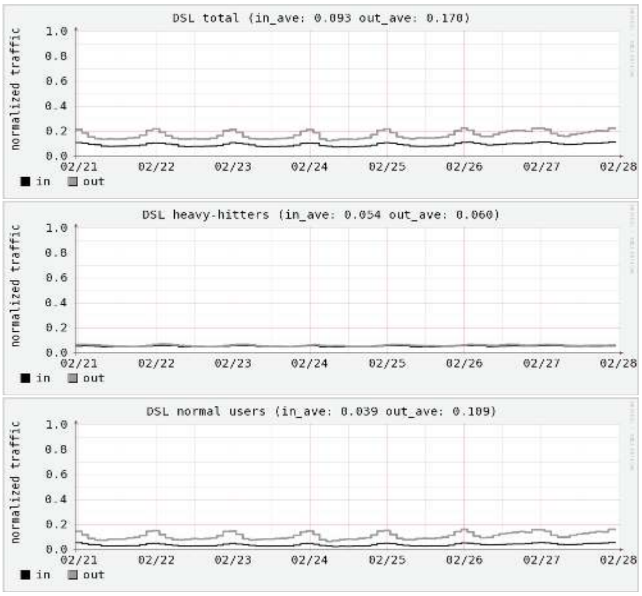

Figure 16: DSL weekly traffic: total DSL users (top), heavy-hitters (middle) and normal users (bottom)

1024. We do not distinguish registered ports from dynamic ports since many implementations use the registered port range from 1024 through 49151 for dynamic ports.

Port 80 (http) accounts only for 9% of the total traffic. TCP dynamic ports account for 83% but the usage of each port is small, probably because the most popular peer-to-peer file-sharing software in Japan, WINNY [13], uses arbitrary ports. The largest one, port 6699, is only 1.4%. It is evident that it is no longer possible to make use of port numbers for identifying applications.

4.5 Geographic Traffic Matrices

To investigate geographic communication patterns among residential users, we classify traffic using the geo-IP databases.

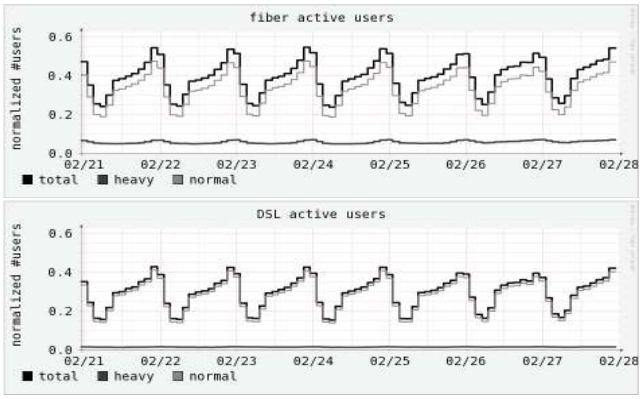

Figure 17: Normalized number of active users in fiber (top) and DSL (bottom): total fiber active users, heavy-hitters and normal users

Table 7 shows the traffic matrix among residential users (RBB), domestic data-centers and leased-lines (DOM), and international addresses (INTL). Note that the data covers only the ISP's residential customer traffic so that the results depict only the behavior of residential users. Residential user-to-user traffic accounts for 63% of the total residential traffic. This is a conservative estimate since the international group also includes residential users.

Table 7: Traffic matrix of the July data set

src\dst	ALL	RBB	DOM	INTL
ALL	100.0	84.8	11.1	4.1
RBB	77.0	63.3	9.8	3.9
DOM	18.0	16.7	1.1	0.2
INTL	5.0	4.8	0.2	0.0

Table 8: Traffic matrices further divided into heavy-hitters and normal users

	heavy-hitters				normal users			
src\dst	ALL	RBB	DOM	INTL	ALL	RBB	DOM	INTL
ALL	69.7	57.8	8.6	3.3	30.3	27.0	2.6	0.7
RBB	59.4	48.4	7.8	3.2	17.6	14.9	2.0	0.7
DOM	8.7	7.9	0.7	0.1	9.3	8.8	0.5	0.0
INTL	1.6	1.5	0.1	0.0	3.4	3.3	0.1	0.0

A surprisingly large portion, about 90%, is domestic communication where both ends are either domestic residential users or other domestic addresses. One possible explanation is language and cultural barriers; the majority of content is in the Japanese language and/or is popular only with the Japanese. However, there are many Japanese worldwide who may access content in Japan, and Japanese content such as animation is popular with non-Japanese as well. Another plausible explanation is that domestic fiber users are connected so well in terms of bandwidth and latency that super-nodes in peer-to-peer networks are interconnected mainly among domestic heavy-hitters.

A small degree of mis-classification is found in the table; 1.5% among DOM and INTL. Since the data is taken from residential traffic, and non-residential flow entries, e.g., management flows for routers, were filtered by the ISP in advance, the traffic not including RBB should be zero. The disparity is caused by new residential address blocks not listed in the geo-IP database. Although it was possible to fix the database using the information from the ISP, we did not do so since errors of the same kind are expected in address blocks of other ISPs at a similar error rate.

Table 7 is further divided into heavy-hitters and normal users in Table 8 where flows with a heavy-hitter's address at either end are classified to the heavy-hitter traffic and the rest of the flows are classified into the normal user traffic. The ratio of user-to-user traffic is 69% (48.4/69.7) for heavy-hitters and 49% (14.9/30.3) for normal users. The ratio of download traffic from DOM or INTL to RBB is much larger for the normal users.

To show the geographic distribution of domestic user-to-user traffic, a prefectural traffic matrix is shown in Figure 18 in which the prefectures are ordered by geographic locations for source (row) and destination (column). In order to observe differences among prefectures, the traffic volumes are normalized to the source prefecture so that the sum of the columns for each row becomes 100%.

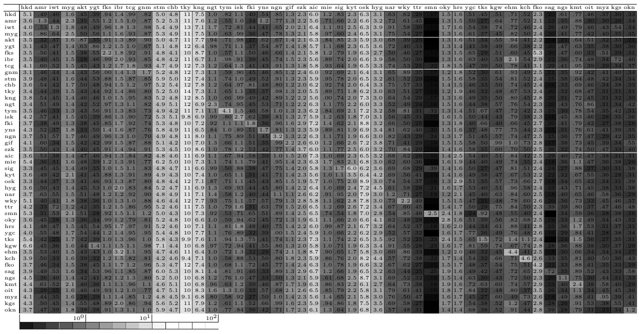

Figure 18: Traffic matrix of 47 prefectures normalized to the source prefecture. The columns have similar values indicating that the distributions of destinations are similar among different prefectures.

All the rows in the matrix have similar distributions, that is, users access similar destinations regardless of the user location. It is also confirmed that traffic volume is roughly proportional to prefectural population. The traffic local to the prefecture is on the diagonal line from the upper left to the bottom right, and is only 2-3% of the total volume for all the prefectures. On the other hand, we cannot identify any increase in traffic to neighbor prefectures. A similar result was found when the distribution is normalized to the destination prefecture. The results suggest that Internet traffic has very poor locality, in contrast to telephone communication where users tend to talk to nearby neighbors. However, this phenomenon might just be the behavior of dominant applications rather than the fundamental nature of Internet communication.

In order to distinguish application types in user-to-user traffic, we investigated the number of peers for each user. Before the experiment, we expected to observe two application types: a small number of peers for video-streaming and downloading from servers, and a large number of peers for peer-to-peer file-sharing.

To observe the number of peers by unique IP address, it is necessary to exclude peers with small traffic volumes since the tail of the distribution is long due to small transactions such as DNS lookups and web browsing. Thus, each user's peers are sorted inversely by volume, and then, the number of peers exceeding the 50th-percentile of the user's traffic volume is counted. We call this number, the *dominant peer count*. Our data set consisted of traffic from one day taken on July 5th, 2005. The inbound traffic for those users who used more than 1GB were extracted for analysis. The destination peer types are classified by the geo-IP databases into residential *users*, *domestic* and others. Then, each user is classified by the largest destination type into the user-to-user group, the user-to-domestic group and others.

Figure 19 shows the (complementary) cumulative distribution of dominant peer counts for user-to-user and user-to-domestic for comparison. For the majority of the users in the upper left region in the graph, the user-to-user group has a much larger number of peers than the user-to-domestic group; 80% of the user-to-user group at the horizontal line have less than 18 dominant peers while 80% of the user-to-domestic group have only less than 4.7 dominant peers. For those who have many peers, the difference becomes smaller since the domestic group also includes file-sharing users in organizations connected by leased-lines.

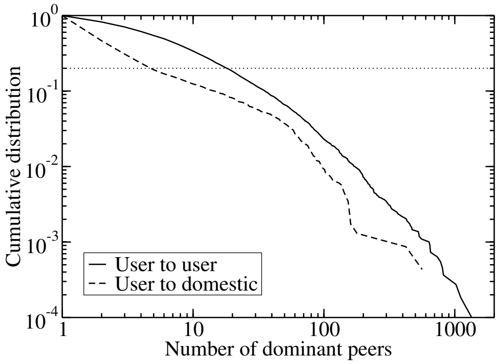

Figure 19: CDF of dominant peer counts for user-to-user and user-to-domestic

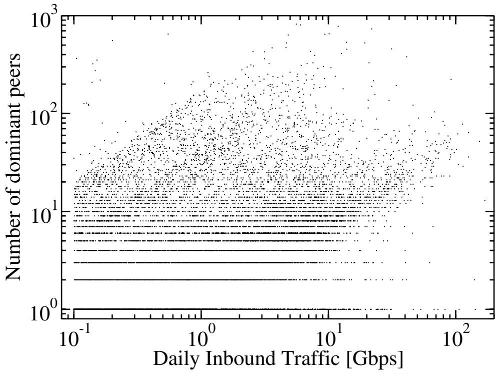

Figure 20: Correlation between dominant peer counts and traffic volumes in user-to-user traffic

Figure 20 shows the correlation between dominant peer counts and the traffic volumes for the user-to-user group on a log-log scale. There are a surprisingly wide range of peer numbers regardless of the traffic volume. Some users communicate with more than 1000 peers but those outliers above the graph range seem to perform scanning. A large number of users communicate with only one or a few peers even in the high-volume region but similarly many users communicate with 10-100 peers. If file-sharing applications have a typical transfer size for each peer, we should be able to observe positive correlation between peer numbers and traffic volumes. Although a positive correlation coefficient (0.28) is observed, it spans a wide range of traffic volume, and the extreme heavy-hitters with a few peers do not follow the correlation. This suggests that high-volume traffic is generated not only by peer-to-peer file-sharing but also by other applications such as content-downloading from a single server. A plausible explanation for the large variance between peer numbers and traffic volumes is that many users use both file-sharing and downloading at different ratios.

On the other hand, the correlation plot for the user-to-domestic group (omitted due to the space constraints) has a slight negative correlation coefficient (-0.12). That is, heavy-hitters in this group tend to communicate with a small number of peers when compared with the user-to-user group.

5. RELATED WORK

The properties of residential broadband traffic differ considerably from those of academic or office traffic often seen in literature. The peak hours have shifted from office hours to evening hours, and the constantly flowing portion of daily traffic fluctuations is much larger than those found in earlier reports [4, 5, 25].

There is little solid work in literature that tries to estimate the growth rate of Internet traffic. Although old studies [23, 3] exist, it has become harder and harder after the privatization of the Internet in mid 90s. Recently, Odlyzko analyzed various aspects of traffic growth, and reported the growth rate of 100% per year for the U.S. in 2003 [19]. The IX traffic in Figure 1 shows that the Internet traffic in Japan is still steadily growing although the growth rate has slowed down after 2002 and has been stable at about 50% per year for the last few years. Statistics in governmental reports of Hong Kong and Australia also show similar trends in their traffic growth [9, 1]. Probably, it is partly because broadband deployment has reached most technically conscious users, and partly because the IT industry is still recovering from the post dot-com bubble economy.

Our results are consistent with earlier measurements of peer-to-peer traffic; peer-to-peer traffic is dominant in commercial backbones [24], highly variable and skewed among participating nodes [26, 27, 8], and exhibits behavior considerably different from traditional web traffic [7]. However, measurement techniques relying on known port numbers to identify applications can no longer be applied since peer-to-peer traffic is shifting from known to arbitrary ports [14].

Among various peer-to-peer measurements, a study of France Telecom's ADSL networks [24] is similar to our per-customer analysis in monitoring access lines and comparing traffic volumes among data sets over a year. However, their focus is on file-sharing applications and the monitoring method relies on known port numbers. The results showing two distinct user segments are considerably different from ours, probably due to the fiber user ratio and differences in popular application software from our measurements.

Many access technologies employ asymmetric line speed for inbound and outbound based on the assumption that content-downloading is dominant for average users. However, this assumption does not hold in our measurements. Although many studies report the asymmetric nature of peer-to-peer traffic [27, 26, 24], it is clear from our comparison between fiber and DSL users that the bandwidth demands of applications and users are not asymmetric, and the deployment of symmetric access will change traffic patterns.

It is known that, in general, peer-to-peer traffic exhibits very poor geographic locality [16, 15]. The intersection between query sets across different regions is also very small [16] partly due to language and cultural barriers [10, 24]. However, modern peer-to-peer networks are characterized by the small-world model (i.e., small diameter and highly clustered topologies) [28], which could lead to heterogeneous behavior in different geographic regions [16]. Our result of poor locality in user-to-user traffic is consistent with others, though the granularity of the analysis is only at the prefectural level.

This paper focuses on user-to-user traffic rather than peer-to-peer file-sharing. Our results show that file-sharing is not the only dominant application in user-to-user traffic. Our work is, to the best of our knowledge, the first involving the collection of long-term measurements from multiple ISPs to estimate nation-wide traffic volume, and the first investigating user-to-user traffic in both fiber and DSL access lines.

6. IMPLICATIONS

Our per-customer measurements reveal the behavior of residential traffic in depth. At first, we noticed a large skew in traffic usage: the top 4% of heavy-hitters account for 75% and 60% of inbound and outbound traffic, respectively. Fiber traffic accounts for 86% and 75% of inbound and outbound traffic, respectively. We tend to attribute the skews to the divide between a handful of heavy-hitters and the rest of the users. Our in-depth analysis, however, shows the existence of diverse and widespread heavy-hitters who appear to be casual users rather than more dedicated users. In addition, the total traffic behavior seems to reflect the balance of the diversity.

For example, the large skew in per-customer traffic seems to be caused by a small number of heavy-hitters but, in fact, the distribution of per-customer traffic is heavy-tailed and it is difficult to draw a line between heavy-hitters and the rest of the users. The large skew in traffic volume between fiber and DSL is not caused by qualitative differences in the behaviors of fiber and DSL users but simply by the larger percentage of heavy-hitters among fiber users. The domination of user-to-user traffic in residential traffic seemingly points to peer-to-peer file-sharing but it is apparently a mixture of file-sharing and content-downloading. All the results indicate that the perceived divides are actually caused by diversity, or at least the divides are blurred by diversity. At the same time, the overall user behavior reflects the balance of this diversity, but it is sometimes dictated by the most influential group.

We can no longer view heavy-hitters as exceptional extremes since there are too many of them, and they are statistically distributed over a wide traffic volume range. It is more natural to think they are casual users who start playing with new applications such as video-downloading and peer-to-peer file-sharing, become heavy-hitters, and eventually shift from DSL to fiber. Other users subscribe to fiber first, and then, look for applications to use the abundant bandwidth. These casual users do not pay attention to underlying technologies and are capricious; their behavior would be easily affected by social, economical or political factors. The implication is that, if a new attractive application emerges, a drastic rise could occur in traffic usage. By the same token, traffic could decrease. For example, current peer-to-peer applications exhibit poor locality but locality-aware applications would make better use of backbone bandwidth. Current peer-to-peer file-sharing algorithms are designed to fill the narrow upstream bandwidth of DSL and are too aggressive for fiber access where it would not take too long to download contents on demand.

In fact, the current total traffic volume is heavily impacted by extreme heavy-hitters so that a slight change in the algorithms or charging policies could have a significant impact to backbone traffic. This situation can be regarded as a tragedy of the commons [17] since the cost-effective nature of the Internet architecture relies on statistical multiplexing assuming bandwidth sharing among users. In fact, ISPs are tempted to avoid congestion by suppressing traffic from extreme heavy-hitters, and it might be even desirable from a view point of fair share of the infrastructural costs. At the same time, it is just a stopgap since users as a whole are shifting towards high-volume usage.

As for the generality of our measurements, several aspects are specific to Japanese traffic. One is the high penetration of fiber access. It seems to take some time for other countries to deploy fiber access; even Korea that has the highest broadband penetration ratio does not have widespread fiber access [20]. Japan can be regarded as a model of widespread symmetric residential broadband access. Another aspect is fairly closed domestic traffic. The current situation is partly due to language and cultural barriers and partly due to rich connectivity within the country. The former could be common to other non-English speaking countries to some extent, and the latter can be seen simply as the geographic concentration of bandwidth-rich users.

7. CONCLUSION

The widespread deployment of residential broadband access has tremendous implications on our lives. Although its effects on the Internet infrastructure are difficult to predict, it is essential for researchers and industry to prepare to accommodate innovations brought by empowered end users. Extensive effort to establish protected data sharing mechanisms with commercial Japanese Internet backbone providers has allowed us to achieve an unprecedented empirical analysis of a significant segment of the Japanese residential broadband traffic.

The growth of residential broadband traffic has already contributed to a significant increase in commercial backbone traffic. In our study, residential broadband traffic accounts for two thirds of the ISP backbone traffic and is increasing at 37% per year, which will force significant reevaluation of the pricing and cost structures of the ISP industry.

We have further studied residential per-customer traffic in one of the ISPs, and investigated differences between DSL and fiber users, heavy-hitters and normal users, and in geographic traffic matrices. We found that a small segment of users dictates the overall behavior; 4% of heavy-hitters account for 75% of the inbound volume. The fiber users account for 86% of the inbound volume. About 63% of the residential traffic volume is user-to-user traffic that exhibits diverse behavior. The distribution of heavy-hitters is heavy-tailed without a clear boundary between heavy-hitters and the rest of users.

For future work, we will continue collecting aggregated traffic logs from participating ISPs. We are also planning to do per-customer traffic analysis from other ISPs, and hope to compare our results with measurements from non-Japanese ISPs.

Acknowledgments

We are grateful to the following ISPs for their support and assistance with our data collection; IIJ, Japan Telecom, K-Opticom, KDDI, NTT Communications, POWEREDCOM, and SOFTBANK BB. We would like to thank the Ministry of Internal Affairs and Communications of Japan for their support in coordinating our study. We would also like to thank Atsushi Shionozaki of Sony CSL, Randy Bush of IIJ, k claffy of CAIDA, and the anonymous reviewers for their valuable input.

8. REFERENCES

[1] Australian Bureau of Statistics: Internet Activity, Australia, Mar 2005. http://www.abs.gov.au/.

[2] Cisco Sampled NetFlow. http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newf%t/120limit/120s/120s11/12s_sanf.htm.

[3] K. Claffy, G. Polyzos, and H. Braun. Traffic characteristics of the T1 NSFNET backbone. In *INFOCOM'93*, pages 885–892, San Francisco, CA, Mar. 1993.

[4] A. Feldmann, A. G. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving traffic demands for operational IP networks: methodology and experience. In *SIGCOMM*, pages 257–270, Stockholm, Sweden, Aug. 2000.

[5] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot. Packet-Level Traffic Measurements from the Sprint IP Backbone. *IEEE Network*, 17(6):6–16, Nov. 2003.

[6] K. Fukuda, K. Cho, and H. Esaki. The impact of residential broadband traffic on Japanese ISP backbones. *SIGCOMM CCR*, 35(1):15–21, Jan. 2005.

[7] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *SOSP-19*, pages 314–329, Bolton Landing, NY, Oct. 2003.

[8] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of bittorrent-like systems. In *IMC2005*, pages 35–48, Berkeley, CA, Oct. 2005.

[9] Hong Kong OFTA: Statistics of Internet Traffic Volume. http://www.ofta.gov.hk/en/tele-lic/operator-licensees/opr-isp/s2.html.

[10] M. Izal, G. Urvoy-Keller, E. W. Biersack, P. A. Felber, A. A. Hamra, and L. Garceś-Erice. Dissecting bittorrent: Five months in a torrent's lifetime. In *PAM2004 (LNCS3015)*, pages 1–11, Antibes Juan-les-Pins, France, Apr. 2004.

[11] Japan Internet Exchange Co., Ltd. http://www.jpix.co.jp.

[12] Multifeed JPNAP service. http://www.jpnap.net.

[13] I. Kaneko. *The Technology of Winny (In Japanese)*. ASCII, Tokyo, Japan, 2005.

[14] T. Karagiannis, A. Broido, N. Brownlee, kc claffy, and M. Faloutsos. Is p2p dying or just hiding? In *Globecom 2004*, pages 1532–1538, Dallas, TX, Dec. 2004.

[15] T. Karagiannis, P. Rodriguez, and D. Papagiannaki. Should Internet service providers fear peer-assisted content distribution? In *IMC2005*, pages 63–76, Berkeley, CA, Oct. 2005.

[16] A. Klemm, C. Lindemann, M. K. Vernon, and O. P. Waldhorst. Characterizing the query behavior in peer-to-peer file sharing systems. In *IMC2004*, pages 55–67, Sicily, Italy, Oct. 2004.

[17] L. Lessig. *The future of ideas*. Random House, 2001.

[18] NSPIXP. http://nspixp.wide.ad.jp.

[19] A. M. Odlyzko. Internet traffic growth: Sources and implications. In *Optical Transmission Systems and Equipment for WDM Networking II*, pages 1–15, San Jose, CA, Jan. 2003.

[20] OECD Broadband Statistics, December 2005. http://oecd.org/sti/ict/broadband.

[21] T. Oetiker. RRDtool: Round Robin Database Tool. http://ee-staff.ethz.ch/~oetiker/webtools/rrdtool/.

[22] T. Oetiker. MRTG: The multi router traffic grapher. In *USENIX LISA*, pages 141–147, Boston, MA, Dec. 1998.

[23] V. Paxson. Growth trends in wide-area TCP connections. *IEEE Network*, 8(4):8–17, July 1994.

[24] L. Plissonneau, J.-L. Costeux, and P. Brown. Analysis of peer-to-peer traffic on ADSL. In *PAM2005 (LNCS3431)*, pages 69–82, Boston, MA, Mar. 2005.

[25] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang. Experience in measuring Internet backbone traffic variability: Models, metrics, measurements and meaning. In *International Teletraffic Congress 18*, 2003.

[26] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A measurement study of peer-to-peer file sharing systems. In *MMCN'02*, pages 156–170, San Jose, CA, Jan. 2002.

[27] S. Sen and J. Wang. Analyzing peer-to-peer traffic across large networks. In *IMW*, pp. 137–150, Marseille, France, Nov. 2002.

[28] D. Stutzbach, R. Rejaie, and S. Sen. Characterizing unstructured overlay topologies in modern p2p file-share systems. In *IMC2005*, pages 49–62, Berkeley, CA, Oct. 2005.

[29] Statistical Outlook of the Internet in Japan (in Japanese). The Ministry of Internal Affairs and Communications, 2005.