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

Kenjiro Cho Kensuke Fukuda Hiroshi Esaki Akira Kato

Abstract

As peer-to-peer applications become popular, an unprecedented increase in user-to-user traffic has been observed worldwide, particularly in Japan due to its high penetration rate of broadband access. In this paper, we first report aggregated traffic measurements collected over 15 months from seven ISPs covering 41% of the Japanese backbone traffic. The backbone is dominated by symmetric residential traffic which increased 45% 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. The fiber users account for 86% of the inbound volume, and 62% of the total volume is user-to-user traffic. The dominant applications have poor locality and communicate with a wide range and number of peers. The distribution of heavy-hitters follows power law without a clear boundary between heavy-hitters and normal users, which suggests that ordinary 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 a new attractive application has drastically affected traffic usage and capacity engineering requirements.

1 Introduction

As peer-to-peer applications have become popular 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 broadband access. The traffic growth in Japanese backbones is illustrated by the aggregated peak traffic at major IXes — JPNAP[10], JPIX[9], and NSPIXP[17] — shown in Figure 1.

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 that has 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 their usage are emerging to take advantage of low-cost high-speed connectivity. In Japan, the number of FTTH subscribers is increasing exponentially while the increase in DSL subscribers is slowing down as shown in Figure 2 [28].

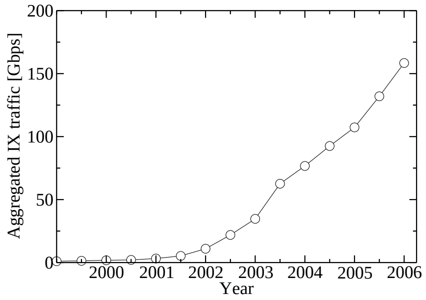

Figure 1: Traffic growth of the peak rate at the major Japanese IXes

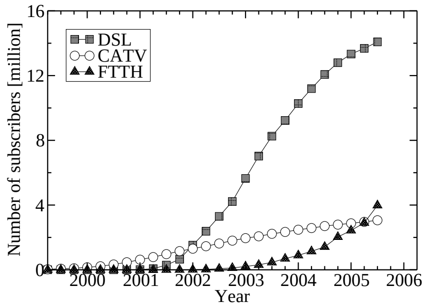

Figure 2: Increase of residential broadband subscribers in Japan: 21 million total broadband subscribers, 14 million for DSL, 3 million for CATV, and 4 million for FTTH as of September 2005.

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 to support this traditionally low-profit customer segment.

It is essential to sustain evolution of the Internet infrastructure 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 are collecting traffic information for their internal use but such data contain sensitive information and are seldom made available to others. In addition, measurement 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 commercial Japanese ISPs in order to identify the macro-level impact of residential broadband traffic on ISP backbones. Specifically, we are trying 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 show that the backbone traffic is dominated by symmetric residential traffic that is increasing at 45% per year.

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 a macroscopic view of residential traffic. The other set was collected by Sampled NetFlow [1] 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 [20] or RRDtool [19]. 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 able to be aggregated with those provided by other ISPs so that the share of each ISP is not revealed.

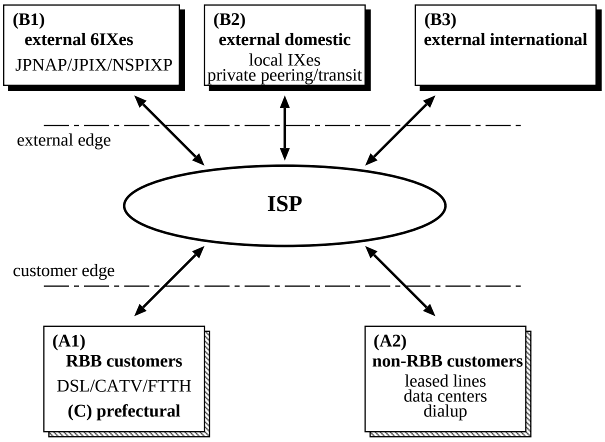

Figure 3: Five traffic groups at ISP customer and external boundaries for data collection

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 5 traffic groups shown in Figure 3 for data collection.

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

(A2) non-RBB customers are 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 are links for 6 major IXes, namely JPNAP, JPIX 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 links are domestic external links other than the 6IXes, including regional IXes, private peering and transit. We used the term “domestic” to mean both ends of a link are in Japan. This group also includes domestic peering with global ASes.

(B3) external international are international external links.

(C) prefectural links are RBB links divided 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 whose links are not based on prefectures 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 relatively small in this particular result since the ISPs in our data sets are peering, 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 is 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 the 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 five times over 15 months; September, October, November in 2004, May and November in 2005. To check consistency, we collected the data separately in each month. These results are consistent so that we are fairly confident about their accuracy. After the first three months, we confirmed the consistency in the results and decided to collect data only twice a year to reduce the workload of the participating ISPs.

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. The data were 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.

Data from February 2005 was used to analyze per-customer behavior from 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 November 2005 were 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 SUTFPOINT 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. Here, *domestic* corresponds mainly to data-centers and leased lines in Japan, but it also includes residential address blocks not listed in the geo-IP databases.

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.

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 growth rates are shown in Figure 4. The growth rate of the RBB customer traffic (A1) is 38% for inbound, 51% for outbound, and 45% for the combined volume between November 2004 and November 2005. The difference between inbound and outbound slightly widened in the first 6 months. The increase of the inbound traffic in (A2) is attributed to a popular video-streaming service started in November 2005. 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) is 59% for inbound and 64% for outbound in November 2005.

Table 1: Average rates of aggregated customer traffic over 15 months

		(A1)customer-RBB (7 ISPs) inbound	 outbound	(A2)customer-non-RBB (4 ISPs) 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

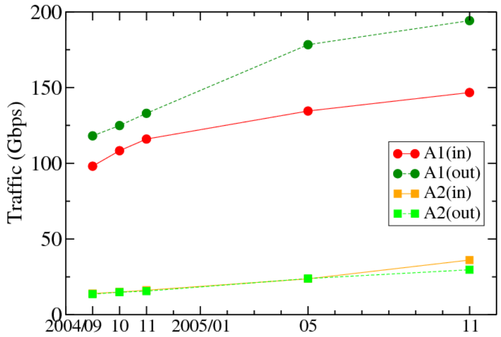

Figure 4: Growth of customer traffic: (A1) RBB customer and (A2) non-RBB customer

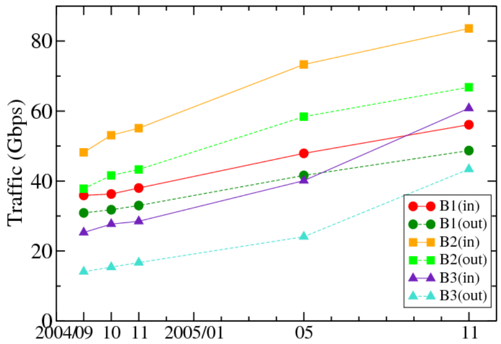

Figure 5: Growth of external traffic: (B1) 6 major IXes, (B2) other domestic and (B3) international

Table 2 summarizes the average rates of aggregated external traffic, and the growth rates are 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 volume of private peering is larger in our measurement than 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 26% for inbound and 30% for outbound in November 2005.

Table 2: Average rates of aggregated external traffic over 15 months

		(B1)ext-6ix (7 ISPs) in	 out	(B2)ext-dom (7 ISPs) in	 out	(B3)ext-intl (7 ISPs) 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

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). These relationships are used for consistency checking of our measurements. However, the 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 this is probably because the total customer traffic contains traffic whose source and destination belong to the same ISP.

Table 3: Average rates of total customer traffic and total external traffic over 15 months

		(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

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 [28]. In comparison with the published total incoming traffic of these IXes, our data consistently represent about 41% 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 is roughly estimated to be 353Gbps for inbound and 468Gbps for outbound in November 2005. We are not aware of available data from other countries against which to compare these numbers.

Table 4: IX traffic observed from ISPs and from IXes over 15 months

		(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

3.2 Customer Traffic

Figure 6 shows the weekly traffic of RBB customers, (A1), consisting of DSL/FTTH/CATV residential users. For weekly data analysis, we took the averages of the same weekdays in the month. We excluded holidays from the weekly analysis since their traffic pattern is closer to that of weekends. 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 close 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.

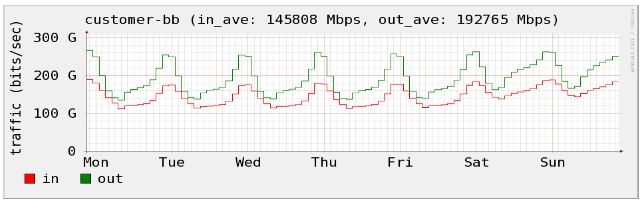

Figure 6: Aggregated RBB customer weekly traffic in November 2005. Darker vertical lines indicate the start of the day (0:00 am in local-time).

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.

Figure 8 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 commercial traffic appears to be smaller than residential customer traffic. The traffic patterns common to Figure 6 and 8 are different from well-known academic or business usage patterns in which the peak is found

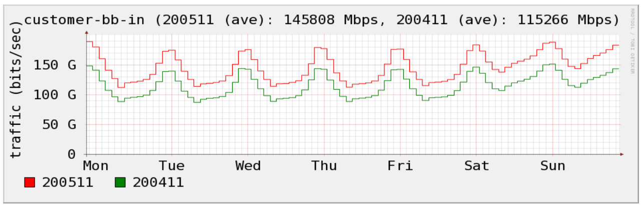

Figure 7: Growth of inbound traffic of RBB customers between November 2004 and November 2005

during office-hours.

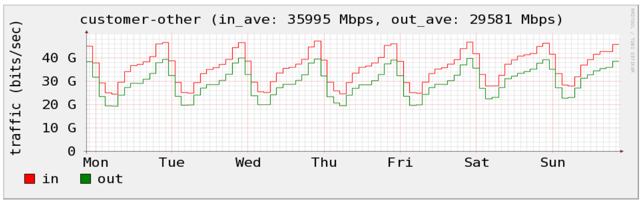

Figure 8: Aggregated non-RBB customer weekly traffic in November 2005 (from 4 ISPs)

3.3 External Traffic

The external traffic groups are used to understand the total traffic volume in backbone networks. Figure 9 shows traffic to and from the six major IXes (B1). It is apparent that the traffic behavior is strongly affected by residential traffic.

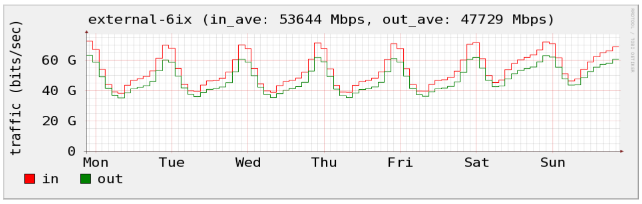

Figure 9: External weekly traffic to/from the 6 major IXes in November 2005

Figure 10 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 Figure 9.

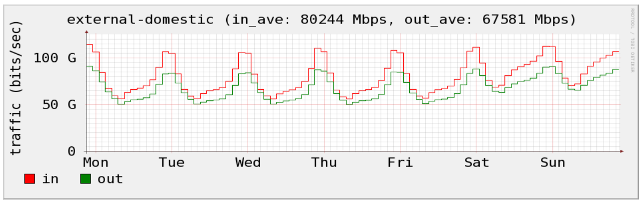

Figure 10: External domestic weekly traffic in November 2005

Figure 11 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 fluctuates less than inbound traffic, suggesting that the traditional behavior of Japanese users downloading content from overseas is still non-negligible part of international traffic. At the same time, the constant part is about 70% of the average inbound rate so that machine-generated traffic could be a large part of it.

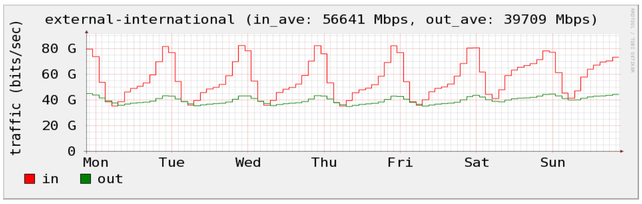

Figure 11: External international weekly traffic in November 2005

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 12 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.

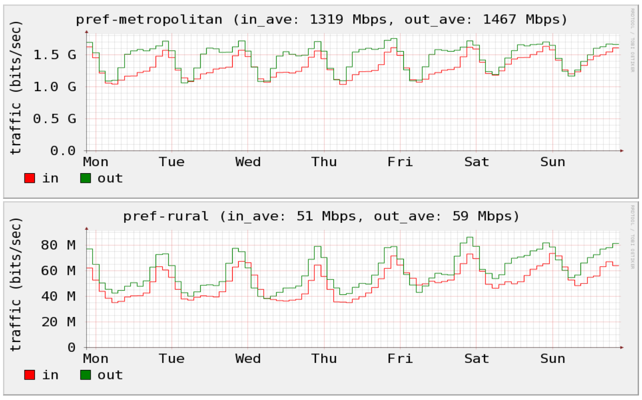

Figure 12: Example prefectural traffic: a metropolitan prefecture (top) and a rural prefecture (bottom)

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.

In order to analyze the scaling property of traffic volume, we show the (complementary) cumulative distribution of prefectural traffic on a log-log scale in Figure

13. The plot conforms to a power law distribution with a cutoff point at 700Mbps, meaning that there is no typical size of prefectural traffic volume. It is also observed that the plots for the top 5 largest prefectures deviate from the power law. To investigate this power law decay, we show the cumulative distribution of prefectural population in the sub-panel. The plots reveal that the power law decay appearing in the traffic volume is derived from the power law decay of prefectural population, as can be inferred from the linear relationship between traffic and population. Thus, we can conclude that 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 fairly universal access services in Japan; 100Mbps fiber access is available in most city areas.

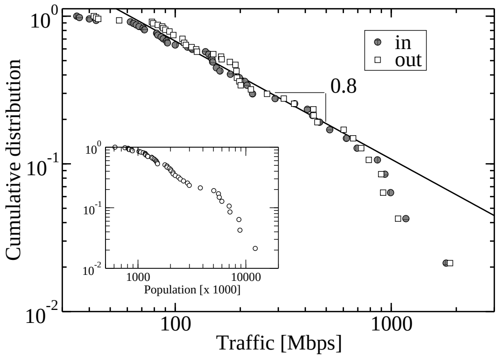

Figure 13: Cumulative distribution of prefectural traffic. Sub-panel shows the cumulative distribution of population for comparison.

4 Analysis of per-customer traffic

This section analyzes NetFlow data from one of the ISPs. Although the data sets are from only one ISP, the traffic characteristics appear to be consistent with the aggregated results in the previous section so that the results are likely to represent Japanese residential traffic.

The number of unique active users observed in the February data set is shown in Table 5. As we explain later, users are classified into two groups by average daily inbound traffic usage, 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.

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	46.4	3.66	42.79
DSL	53.6	0.80	52.75

4.1 Distribution of Heavy-hitters

Figure 14 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.

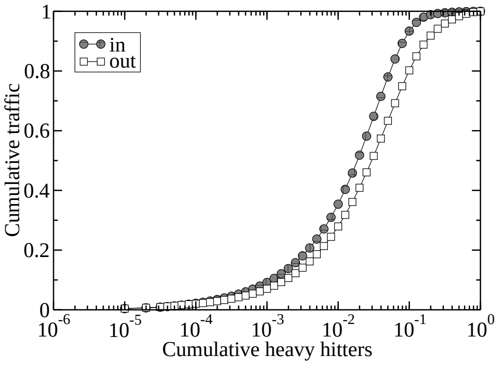

Figure 14: Cumulative distribution of traffic volume of heavy-hitters in decreasing order of volume

Figure 15 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 of the month, and the distribution is computed independently for inbound and outbound traffic.

The distributions follow power law 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

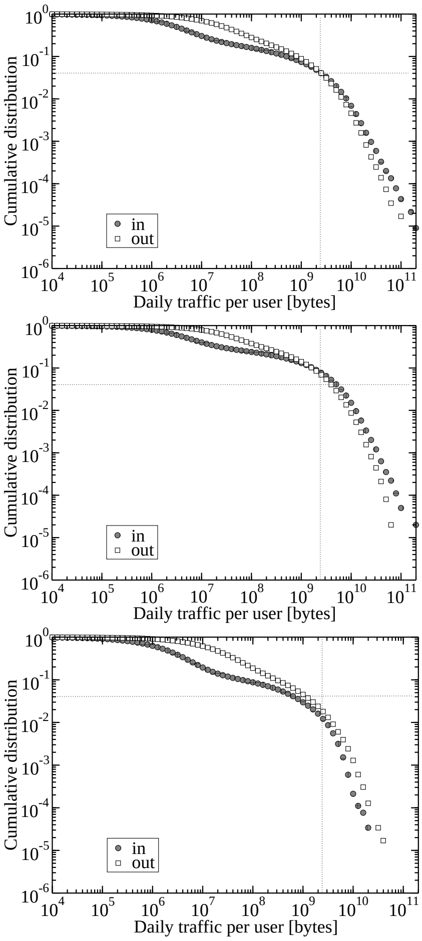

Figure 15: Cumulative distribution of daily traffic per user: total users (top), fiber users (middle) and DSL users (bottom). The lines are drawn at 2.5GB/day and the top 4% heavy-hitters, the knee of the total users' slope.

the users. In the rest of the paper, the *heavy-hitter* group is used to denote users uploading more than 2.5GB/day on average, and the *normal user* group is used for users uploading less than 2.5GB/day on average. 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. There is no typical daily traffic volume per user that can be identified by a concave in the slope.

As for prefectural differences, the distributions look similar across different prefectures as shown in Figure 16 which compares one metropolitan prefecture (top) with one rural prefecture (bottom). One difference is the tail length affected by 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 which in turn affected by the ratio of fiber users including larger heavy-hitter population.

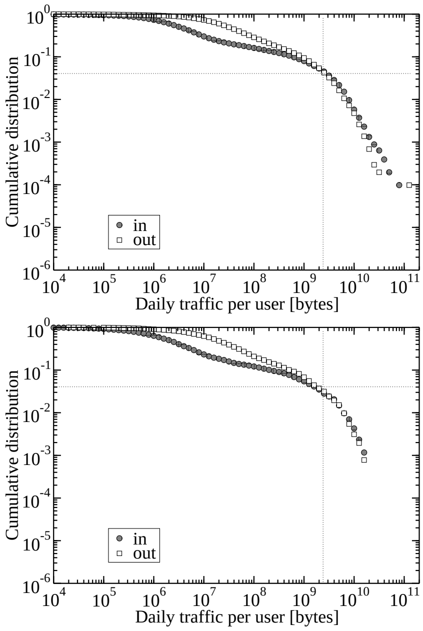

Figure 16: Cumulative distribution of daily traffic per user: a metropolitan prefecture (top) and a rural prefecture (bottom)

4.2 Correlation of Inbound and Outbound Volumes

The correlation between inbound and outbound traffic volumes for each user is shown as log-log scatter plots in Figure 17. These plots are taken from a metropolitan prefecture 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 the volume of outbound (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 15. 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

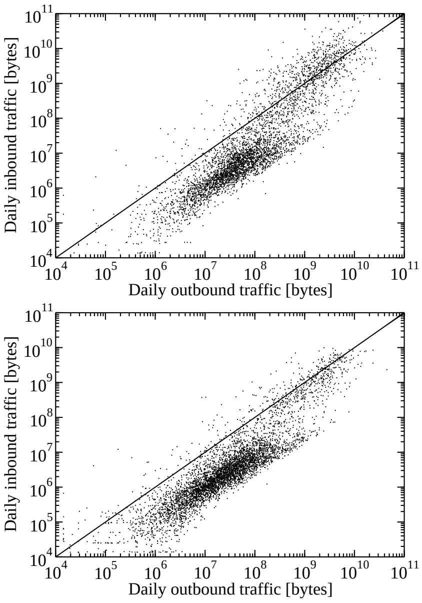

Figure 17: Correlation of inbound and outbound traffic volumes per user in one metropolitan prefecture: fiber (top) and DSL (bottom)

but, as expected, the high-volume cluster is larger in the fiber 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 allows them to compensate for the shortage of upstream bandwidth of DSL lines. 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 18 and Figure 19 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 traffic volume of the ISP.

The plots show that the inbound and outbound volumes are almost equal for fiber traffic but the inbound is about 60% larger for heavy-hitters and the outbound is about 166% larger for the normal users. The total is counterbalanced by the two groups. In the DSL users, the outbound volume is about 83% larger for the total users, only about 11% larger for the heavy-hitters and about 180% 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.

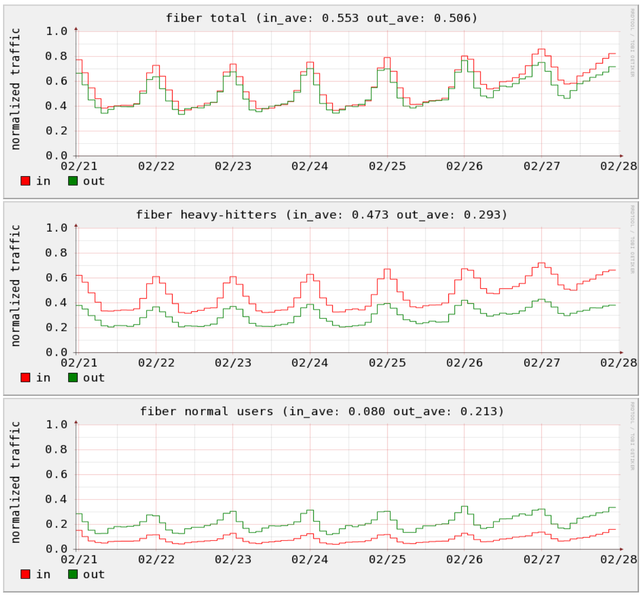

Figure 18: Fiber weekly traffic: total fiber users (top), heavy-hitters (middle) and normal users (bottom)

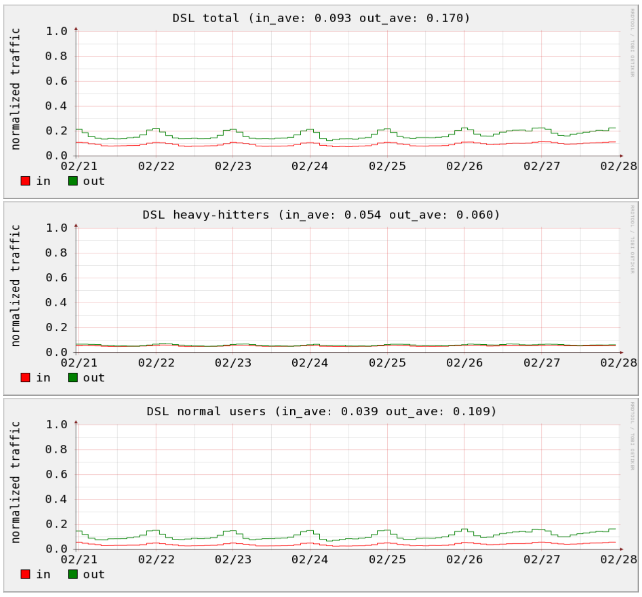

Figure 19: DSL weekly traffic: total DSL users (top), heavy-hitters (middle) and normal users (bottom)

Figure 20 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. When the active user number is compared to the traffic volume, the increase is larger in the morning and smaller in the evening. This behavior suggests that bandwidth use is more intense, i.e., higher bandwidth demand per user, in the evening.

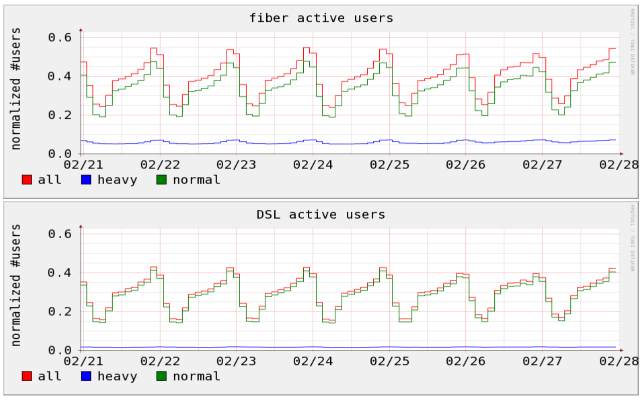

Figure 20: Normarized number of active users in fiber (top) and DSL (bottom): total fiber active users, heavy-hitters and normal users

4.4 Protocol and Port Usage

Table 6 shows the ranking of protocols and ports. To rank port numbers in TCP and UCP, 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 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 [11], 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 Traffic Matrices

To investigate geographic communication patterns among residential users, we classify traffic using the geo-IP databases. Table 7 shows the traffic matrix among residential users (RBB), domestic data-centers and leased-lines (DOM), and international addresses (INTL). Residential user-to-user traffic accounts for

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

protocol	ratio(%)	port #	name	ratio(%)
TCP	97.43			
(*port* < 1024	13.99)	80	http	9.32
		20	ftp-data	0.93
		554	rtsp	0.38
		443	https	0.30
		110	pop3	0.17
		81	-	0.15
		25	smtp	0.14
		119	nntp	0.13
		21	ftp	0.11
		22	ssh	0.09
		-	others	2.27
(*port* >= 1024	83.44)	6699	winmx	1.40
		6346	gnutella	0.92
		7743	winny	0.48
		6881	bittorrent	0.25
		6348	gnutella	0.21
		1935	macromedia-fsc	0.20
		1755	ms-streaming	0.20
		2265	-	0.13
		1234	-	0.12
		4662	edonkey	0.12
		8080	http-proxy	0.11
		-	others	79.30
UDP	1.38	6346	gnutella	0.39
		6257	winmx-	0.06
		-	others	0.93
ESP	1.09			
GRE	0.07			
ICMP	0.01			
OTHERS	0.02			

62% of the total residential traffic. This is a conservative estimate since the international group also includes residential users.

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 Japanese. However, there are many Japanese worldwide who want Japanese content, 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 are 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 with a similar error rate of 1.5%.

To show the geographic distribution of domestic user-to-user traffic, the prefectural traffic matrix is shown

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	62.2	9.8	3.9
DOM	18.0	16.7	1.1	0.2
INTL	5.0	4.8	0.2	0.0

in Figure 21 in which the prefectures are ordered by geographic locations. In order to observe differences among prefectures, the traffic volumes are normalized to the source prefecture so that the sum of the rows becomes 100%.

Most traffic flows to several prefectures with large populations and all the rows have similar distributions, which again confirms that the traffic volume is roughly proportional to the 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 communications.

Figure 22 shows the temporal behavior of the traffic matrix. For heavy-hitters, a large part of the traffic is user-to-user. There are some daily fluctuations even in user-to-user traffic in the heavy-hitter group. It might be that a certain part of heavy-hitter traffic is still triggered manually or it could be triggered by non-heavy hitters manually accessing heavy-hitters.

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. 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, independently for inbound and outbound. To observe differences in peer types, peers are classified into 4 groups by the geo-IP databases: residential *users*, *domestic*, *international* and *invalid*, and then, each user is marked by the largest group. The data were taken from one day of traffic on July 5th 2005, and those users who used more than 1GB were extracted for analysis.

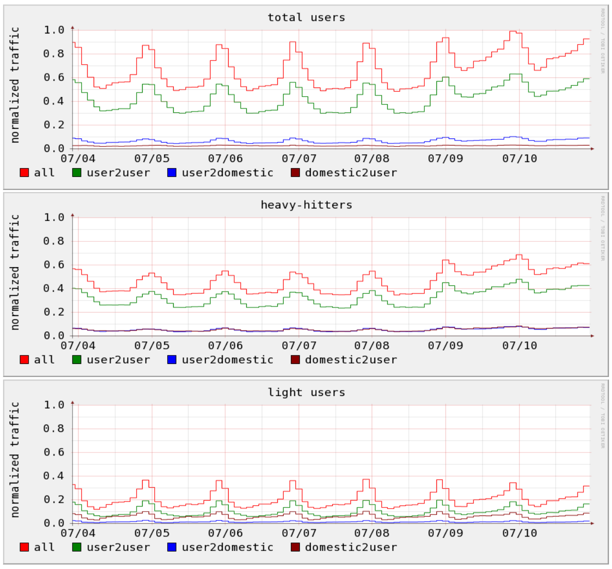

Figure 22: User-to-user weekly traffic for the total users (top), for heavy-hitters (middle) and for normal users (bottom)

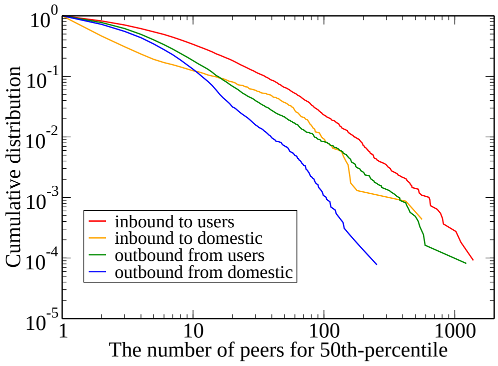

Figure 23: CDF of number of peers for 50th-percentile traffic for four groups

Figure 23 shows the (complementary) cumulative distribution of number of peers for four groups. The distributions have heavy tails, and even the domestic groups include many file-sharing users who are in organizations connected by leased-lines. The two application types we expected are discernible in the distributions only by the slight difference in the slope for more than 10 peers. The inbound-to-users group is, however, different from the others, especially in the upper-left region; the majority of the users have only a few peers. The results show that there is no clear boundary between streaming/downloading and file-sharing, and that there is no typical number of peers.

The corresponding log-log scatter in Figure 24 show the correlation between the peer numbers and the traffic volumes. There are a surprisingly wide range of peer

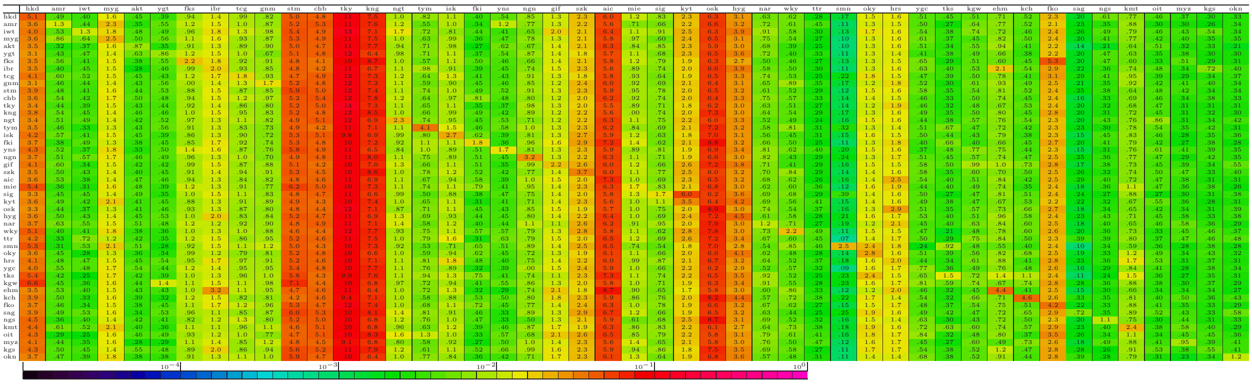

Figure 21: Traffic matrix of 47 prefectures normalized to the source prefecture. The columns have similar values so that the distributions of destinations are similar among different prefectures.

numbers regardless of the traffic volume; some users communicate with more than 1000 peers. 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. The positive correlation means that the number of peers is proportional to the traffic volume. If file-sharing applications have a typical transfer size for each peer, we should be able to observe positive correlation. Although a positive correlation can be observed, it spans a wide range of traffic volumes. In addition, the extremely heavy-hitters with 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 is that the majority of users use both file-sharing and downloading with different ratios.

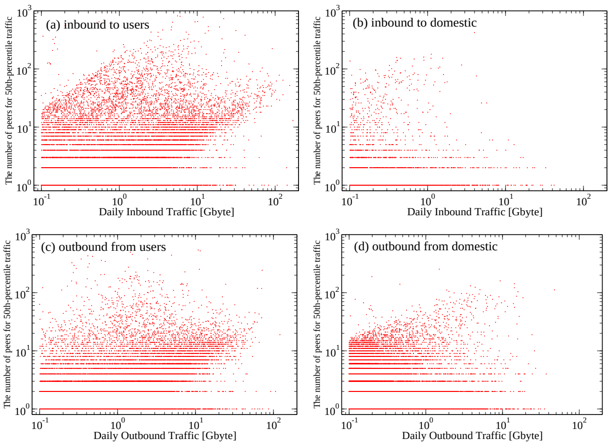

Figure 24: Number of peers for 50th-percentile traffic: inbound to users (top-left), inbound to domestic (top-right), outbound from users (bottom-left) and outbound from domestic (bottom-right)

5 Related Work

The impact of residential broadband traffic can be seen not only in volume but also in usage patterns. The peak hours have shifted from office to evening hours, and emerging file-sharing or other peer-to-peer communications with audio/video content exhibits behavior considerably different from traditional world wide web that was the dominant application in earlier traffic measurement studies [24, 4, 15].

There is little solid work in literature that tries to estimate the growth rate of Internet traffic. Odlyzko analyzes various aspects of the traffic growth, and reports the growth rate of 100% per year in the U.S. in 2003 [18].

Our results are consistent with earlier measurements of peer-to-peer traffic [6, 7, 22, 3]: peer-to-peer traffic is dominant in commercial backbones [23, 21], and highly variable and skewed among participating nodes [25, 26]. However, measurement techniques relying on known port numbers to identify peer-to-peer applications can no longer be applied since peer-to-peer traffic is shifting from known to arbitrary ports [12].

Among various peer-to-peer measurements, a study of France Telecom's ADSL networks [21] 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 are considerably different from ours, probably due to the fiber user ratio and differences in popular applications in our measurement.

Many studies report the asymmetric nature of peer-to-peer traffic [26, 25, 21]. From our comparison between fiber and DSL users, it is clear that the bandwidth demands of applications and users are not asymmetric, and the deployment of symmetric access lines will change traffic patterns in other countries.

It is known that, in general, peer-to-peer traffic has very poor geographic locality [14, 13]. The intersection between query sets across different regions is also very small [14] partly due to language and cultural barriers [8, 21]. However, modern peer-to-peer networks are characterized by small-world (i.e., small diameter and highly clustered network) [27], which could lead to heterogeneous behavior in different geographic regions [14]. 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

It has been reported worldwide that peer-to-peer traffic is taking up a significant portion of backbone networks. Our aggregated measurements indeed show that the backbone traffic is heavily impacted by residential customer traffic which accounts for about 60% of the total customer traffic. Residential customer traffic increased rapidly — 45% between November 2004 and November 2005.

The properties of residential broadband traffic differ considerably from those of academic or office traffic often seen in literature. The constantly flowing portion of daily traffic fluctuations is about 70%, much larger than those found in earlier reports [2, 5]. Research results obtained from campus or other academic networks may no longer apply to commercial traffic.

The inbound and outbound rates are roughly equal throughout our data sets. Many access technologies employ asymmetric line speed for inbound and outbound based on the assumption that content-downloading is dominant for normal users. However, this assumption does not hold in our measurements.

Our measurements also suggest that a large amount of traffic is exchanged by private peering which implies that data from IXes may not be an appropriate index of nation-wide traffic volume.

The prefectural results show that traffic volume is roughly proportional to regional population, and the distribution of customer traffic usage reflects the ratio of fiber and DSL populations. If this is the case, it would affect the design of capacity planning for backbone networks.

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 follows a power law 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 customers but simply by the larger percentage of heavy-hitters among fiber users. The large skew 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. At the same time, the entire 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 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. Or, sometimes users subscribe to fiber first, and then, look for applications to use the abundant bandwidth. The implication is that, if a new attractive application emerges, a drastic change could occur in traffic usage. For example, current peer-to-peer applications do not take locality into consideration, but future applications could as suggested in [13].

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 [16]. Japan is a model of widespread symmetric residential broadband access. Another 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, 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 62% of the residential traffic volume is user-to-user traffic that exhibits impressively diverse behaviors. The distribution of heavy-hitters follows power law without a clear boundary between heavy-hitters and normal 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.

References

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

[2] 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, 2000.

[3] F. L. Fessant, S. Handurukande, A.-M. Kermarrec, and L. Massoulii. Clustering in peer-to-peer file sharing workloads. In *IPTPS 04*, San Diego, CA, Feb. 2004.

[4] M. Fomenkov, K. Keys, D. Moore, and k claffy. Longitudinal study of Internet traffic in 1998-2003. In *WISICT*, Cancun, Mexico, Jan. 2004.

[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*, pages 6–16, 2003.

[6] 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.

[7] 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.

[8] 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.

[9] Japan Internet Exchange Co., Ltd. (JPIX). http://www.jpix.co.jp.

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

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

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

[13] 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.

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

[15] S. McCreary and k. claffy. Trends in wide area IP traffic patterns. In *ITC Specialist Seminar*, Monterey, CA, Sept. 2000.

[16] White Paper Internet KOREA 2003. , National Computerization Agency, Korea, Sept. 2003.

[17] NSPIXP: Network Service Provider Internet eXchange Project. http://nspixp.wide.ad.jp.

[18] A. M. Odlyzko. Internet traffic growth: Sources and implications. In *SPIE 2003*, 2003.

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

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

[21] 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.

[22] J. A. Pouwelse, P. Garbacki, D. H. J. Epema, and H. J. Sips. The bittorrent p2p file-sharing system: measurements and analysis. In *4th Int. Workshop on Peer-to-Peer Systems (IPTPS'05)*, Ithaca, NY, Feb. 2005.

[23] Sandvine inc. EDonkey — Still king of P2P in France and Germany. URL http://www.sandvine.com/news/pr_detail.asp?ID=88, Sept. 2005.

[24] S. Saroiu, K. P. Gummadi, R. Dunn, S. D. Gribble, and H. M. Levy. An analysis of Internet content delivery systems. In *OSDI 2002*, pages 315–328, Boston, MA, Dec. 2002.

[25] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A measurement study of peer-to-peer file sharing systems. In *Multimedia Computing and Networking (MMCN'02)*, San Jose, CA, Jan. 2002.

[26] S. Sen and J. Wang. Analyzing peer-to-peer traffic across large networks. In *Second ACM SIGCOMM Internet Measurement Workshop*, pages 137–150, Marseille, France, Nov. 2002.

[27] 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.

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