Inter-AS Traffic Patterns and Their Implications

Wenjia Fang **Larry Peterson**
Princeton University
{wfang|llp}@cs.princeton.edu

Abstract: This paper reports on a study of traffic patterns among Autonomous Systems (ASes), based on traces taken at various points in the Internet. The traces display a highly non-uniform distribution of traffic on flows between pairs of hosts, networks, and ASes. Aggregation along coarser granularities, such as networks or ASes, accentuates this non-uniform distribution. In one typical trace, for example, the top 9% of flows between ASes accounts for 86.7% of the packets and 90.7% of the bytes transmitted. A highly non-uniform traffic pattern suggests that routers need to maintain only limited QoS flow state. The paper discusses the implications of this phenomenon on different proposed QoS mechanisms.

1.0 Introduction

Routing and data delivery are two integral parts of the Internet architecture. Routing is concerned with building collective information on where end hosts are, and data delivery is the process of forwarding packets to their respective destinations. A succinct observation is that routing advertisement information always travels in the opposite direction of data traffic [Rek95].

Most recent research has focused on routing. This includes the separation of inter-AS (Autonomous System) routing from intra-AS routing [RFC896, RFC1771], policy-based routing [ERH92, Rek95], and studies of routing behavior and instabilities [Paxson96,LMJ97]. However, few studies have been done on large-scale traffic patterns, due in large part to a lack of monitoring and measurement infrastructure. The exceptions are studies by [KleNay74] and [RekChi92], but these are becoming dated as the Internet has recently undergone several significant changes. Most notably, the NSFNET has been replaced by multiple commercial backbones, the adoption of CIDR now allows aggregation of routes and requires longest prefix matching in routers, BGP (Border Gateway Protocol) is now widely deployed as an inter-AS routing protocol, and the WWW has become a significant traffic source.

Despite all these changes, however, we suspect that the non-uniformity observed in earlier studies is still present on the Internet. One probable reason is that certain popular servers are much more likely to send/receive traffic than others. A second is that traffic patterns on any communication network must bear some relationship to the geographical and demographical distribution of its users. Since we expect to see more Internet users at crowded metropolitan areas than rural areas, there would naturally be a greater concentration of traffic flowing between ASes of metropolitan areas than those of rural areas, for example.

A non-uniform distribution pattern usually suggests that caching can be employed to improve performance. We obverse, however, that caching is not relevant for emerging real-time applications like video conferencing and Internet Telephony. For such applications, mechanisms like RVSP [RSVP], Diff-Serv [Diff-Serv] and MPLS [MPLS] must be used, but they require a trade-off between the state information needed in the middle of the network and the performance benefit of such effort. Fortunately, a non-uniform traffic distribution suggests that only limited state information needs to be kept for only certain flows to give a better performance to a large fraction of the traffic.

We used Internet traces taken at both research and commercial networks to study topological spatial localities—defined to characterize the topological closeness among many endhost flows—and non-uniform distribution of packets among flows. Both sets of traces confirm older studies that non-uniform distribution of packets among flows is a widespread phenomenon at various locations of the Internet. Traffic at those locations also displays various degree of topological spatial locality. Thus, aggregation along coarser granularities, such as networks or ASes—accentuates this non-uniform distribution. In one typical trace, for example, the top 9% of flows between ASes accounts for 86.7% of the packets and 90.7% of the bytes transmitted. In addition to presenting these results in more detail, this paper also discusses implications of this phenomenon on current and proposed QoS mechanisms.

The paper is organized as follows. Section 2 discusses related work. Section 3 describes background information, defines terminology, and presents our research methodology. Section 4 then reports some findings from two sets of Internet traces, one from campus and regional networks, and the other from a major commercial backbone. Finally, Section 5 offers arguments for aggregating traffic along networks and ASes and discusses the implications on different available mechanisms, and Section 6 makes some concluding remarks.

2.0 Related Work

Definitions and discussions of "flows" appeared in [CBP95] and [Clark88]. One definition of a flow is a TCP connection, which serves as a good abstraction from a host-centric perspective. [CBP95] introduced a network-centric perspective definition of "flows", where the observation point is a router in stead of a host. A flow is considered inactive or terminated if no packets belong to this flow appeared in the router during a time-out period. [CBP95] defined flows can be of different end points granularities: applications, endhosts, networks, and so on. They also observe that aggregated flows along different endpoint granularities do not proportionally reduce the number of flows, again suggesting a non-uniform distribution traffic pattern.

As early as in 1974, when the ARPANET consisted of approximately 40 switches and 50 hosts, [KleNay74] reported non-uniform distribution among site pairs.[1] The trace was a collection of information gathered over a seven-day period between 1 August, 1973 and 7 August, 1973. The authors observed that "90 percent of total traffic is between 192 (12.6 percent) of the site-pairs."

In [RekChi92], Internet traces were collected to evaluate different schemes of injecting inter-AS routes into intra-AS routing system. When this work was done, the Internet topology consisted of NSFNET backbone, and traces collected on the backbone showed that there was a highly non-uniform distribution of packets among different routes. As a consequence, a few selected routes, with high *route concentration* (a large number of packets forwarded using this route) could be injected into intra-AS routing to avoid the encapsulation/decapsulation cost for a majority of packets that are forwarded onto those inter-AS routes.

Many things have changed since the time when these two pieces of work were carried out. Most prominent changes includes 1) the replacement of the NSFNET backbone by many commercial Internet Service Providers (ISPs); 2) a wide adoption of BGP (Border Gateway Protocol), an inter-AS routing protocol, and a separation of inter-AS routing from the intra-AS routing; 3) the adoption of CIDR addressing scheme, which enables hierarchical route aggregations, and requires longest prefix matching algorithm in routers; 4) a phenomenal growth in the number of routers and hosts; and 5) the emergence of new applications, such as the WWW.

These changes have consequences on our experiments. First, Internet traces should be taken at the backbone of the Internet, where traffic is highly aggregated and representable, as in [RekChi92]. However, despite growing collaboration efforts to provide traces, for example, [NLANR] and [CAIDA], traces are not easily available from many commercial providers. Many are concerned with the sensitive information an Internet trace can reveal. We were allowed to analyze traces

1. In 1974, a site, a host, a network, and an autonomous system could be considered equivalent abstractions.

taken from a backbone router under the condition that we not disclose the identities of the networks or ASes captured in the traces.

Second, it should be noted that given the distributed nature of the Internet, and the widely varying topology and technologies deployed, there will be no single truth about what Internet traffic really looks like. The answer is always "its depends on where you observe". We choose two sets of traces, one from mostly research and educational institutes and another from a commercial ISP's exchange point, in an effort to be more complete about the data points on which we base our conclusions. A research methodology of using traces is always limited by the representativeness of its traces. We do not claim that our traces are representative of Interne traffic as a whole, thus, confining our conclusions to where they are relevant.

Third, the introduction of AS, although used strictly as a routing concept up to this point, leads us to think of it as a more general architectural abstraction. Since an AS is assumed to be under one administrative domain, traffic originating from one AS or destined to another AS can be used as a possible granularity subjected to policies. We use a timeout methodology in [CBP95] to define the a *jumbo flow* as a sequence of packets flowing between two networks or ASes. They are "jumbo" because they are an aggregation of all individual flows between endhosts belonging to the two ASes, and therefore they are both more persistent and bigger than the individual flows.

3.0 Background

3.1 Internet Topology

The current Internet can be modeled as an arbitrary interconnection of Autonomous Systems (ASes). Within one AS, an interior gateway protocol (IGP) with a single routing metric is used. Between ASes, an exterior gateway protocol (EGP) [RFC895] is used. The most widely accepted and deployed exterior gateway protocol is the Border Gateway Protocol (BGP) [RFC1771]. AS was originally a purely routing concept, but eventually, it became synonymous with administrative domains, which refer to a single administrative authority. Each AS has a distinct 16-bit Autonomous System Number (ASN). ASNs are used in BGP to refer to ASes. As of January, 1999, more than 4400 ASes appeared in the global BGP routing tables.

Topologically, the Internet consists of a large number of interconnected regional and national backbones provided by different Internet Service Providers (ISPs). Several backbone providers exchange traffic and routing information at specific *exchange points*, so called *peering*. Therefore, topologically, we can imagine several parallel backbones make up the core of the Internet. Many regional, campus or corporate networks are ASes connecting to these backbone networks; others, are connected to small regional providers. The core of the Internet is also called the *default-free zone* (DFZ) because backbone service providers must maintain a complete map of all globally visible network prefixes reachable throughout the Internet: no default routes are allowed.

TABLE 1. Classification of ASes. Data set from a global routing table take on 12/18/98, of total 4427 ASes.

Class	Degree Range	Number	Approx. fraction	Descriptions
1	>79	15	0.34%	National and international backbone providers
2	22-79	58	1.31%	Large regional providers and European national networks.
3	7-22	226	5.1%	Smaller regional networks, metropolitan providers
4	<7	4128	93%	Smaller metropolitan, campus, and corporate networks.

A more detailed classification of ASes using the methodology in [GovRed96] categorizes ASes into four classes, depending on the *degree* of each AS. A degree of an AS is the number of other ASes it peers with. The degree of an AS can be inferred from the global BGP routing tables at the DFZ. We did a recent study on ASes, and approximated the distribution function of ASes with a piecewise linear function with four components. This gives us four classes of ASes by degree range, which are listed in Table 1.

Compared to the results reported on Internet of September, 1995 [GovRed96], we observe the following changes in the topology of Internet:

- A much larger number of ASes (909 in 09/95 and 4427 on 12/98). This reflects both the growth of the Internet and a wider deployment of BGP and the notion of Autonomous Systems.
- An increase in the degree of ASes in all classes. The average degree in 09/95 was 2.99 and 4.118 in 12/98. This reflects a growing trend in multi-homing among all ASes.
- A significant increase in degree for class 1 ASes, especially the upper bound of the degree of ASes can be as high as 1017. This means that backbone ISPs have a much greater number of stub ASes than before, and possibly much more states if reservations were made between pairs of those stub ASes.

3.2 Flows

There are at least two definitions of *flows*, one from a host-centric perspective and one from a network-centric perspective. From a host-centric perspective, a flow is simply a sequence of packets from one host to another; for example, a TCP connection. In a datagram network, a host knows the start and end of a flow originated from itself, as it has all the state information for this flow. From a network-centric perspective, a flow is observed within a router, and no connection state is explicitly maintained. This definition, given in [CBP95], relies on a timeout mechanism to determine whether the flow has stopped. More specifically, "a flow is active as long as observed packets that are meeting the flow specification are observed separated in time by less than a specified timeout value." For the rest of the discussion, we only use the network-centric definition of flows.

In 1988, [Clark88] proposed using flows as a building block for future network designs, and keep softstates for flows. Essentially, a flow is a new network abstraction in which packets belong to this flow can receive similar treatments with regard to routing, QoS mechanisms, accounting and billing policies, and security policies. A flow is a better building block than a datagram, as articulated in [Clark88], because "the goals of resource management and accoutability have proved difficult to achieve in the context of datagrams". A commercial network will ultimately be concerned with allocation of its resources and the recuperation of its cost. [Clark88] suggests keeping softstate for flows so as to permit the network to retain its primary connectionless nature for survivability and flexibility, while at the same time doing a better job dealing with the issue of resource management and accountability.

Therefore, two necessary requirements for flows as a network building block are: (1) packets belong to a flow can be treated similarly with regard to many types of policies (e.g., routing policies, QoS policies, security policies, and accounting policies), and (2) the overhead of establishing and maintaining states for flows can be amortized over many packets belong to the same flow so that cost for keeping softstate is justifiable.

3.3 Aggregation and Jumbo Flows

Early studies show that most flows between two endhosts are short-lived, transactional in nature, and contain a small number of packets [CBP95]. This calls into question the cost of maintaining states for such short-lived flows. As the Internet grows to include millions of endhosts, the magnitude of states for potential flows between endhosts can be formidable for any router to maintain. It is clear that the traditional definition of flow is not sufficient. Thus, we need an abstraction with coarser granularity.

Aggregating individual endhost flows into networks—–using the network-centric definition of a flow but with end-point granularity of networks instead of end-host IP address—–is attractive. Internet forwarding is essentially based on networks, with all packets heading towards the same network forwarded along the same route together. (Note that the above statement is true for both inter-AS and intra-AS routing. In other words, networks are the smallest granularity in routing.) This makes a nice network-level abstraction subjected to QoS policies and accounting policies. Furthermore, this aggregated flow may be potentially persistent and contain many packets. We call this as a jumbo flow between networks.

Aggregating endhost flows into ASes—–jumbo flows between ASes—–is another possibility. Since ASes usually mirror some external administrative entities, jumbo flows defined between two ASes is a good abstraction for specifying and applying policies to respective entities: QoS, accounting and billing, and security.

In current Internet routing, an AS can advertise many networks to other ASes. BGP allows routing policies to be applied to the granularity of networks (called a *route* in BGP). Therefore, different networks belong to the same AS can be routed differently. At first, this might seem to pose difficulty to jumbo flows between ASes, because packets belong to the same jumbo flows might be forwarded over different routes. However, we observe that ASes are the smallest granularity for routing at inter-AS level, whereas networks are the smallest granularity for routing at intra-AS level. In other words, packets destined to networks from the same AS will always be treated together at the inter-AS level, even though at the intra-AS routing level, packets destined to networks from the same AS might be routed differently.

The subtlety of this observation is depicted in Figure 1, in which, routing advertisements are shown as dotted lines with arrows and packet forwarding is shown in solid lines. There are four ASes, and AS2 is advertising three networks A, B & C to its neighbor, AS1. AS1 applies its inter-AS policy and decides that networks A and B should be further advertised to a neighbor AS3, and network C should be further advertised to another neighbor AS4. Clearly, when packets destined to networks A,B and packets destined to network C are forwarded within AS1, they will have different inter-AS routes as well as intra-AS routes. However, observing from the inter-AS level, only two jumbo flows exist: one between AS4 and AS2, and another between AS3 and AS2. At the inter-AS level, packets belong to either of these two jumbo flows can be treated together. For example, on the link between AS1 and AS4, packets destined to networks A and B can be aggregated. However, within the transit AS1, packets destined to networks A and B (both belong to the same jumbo flow) can be routed via different routes if the IGP of AS1 has different cost for networks A and B [Rek99].

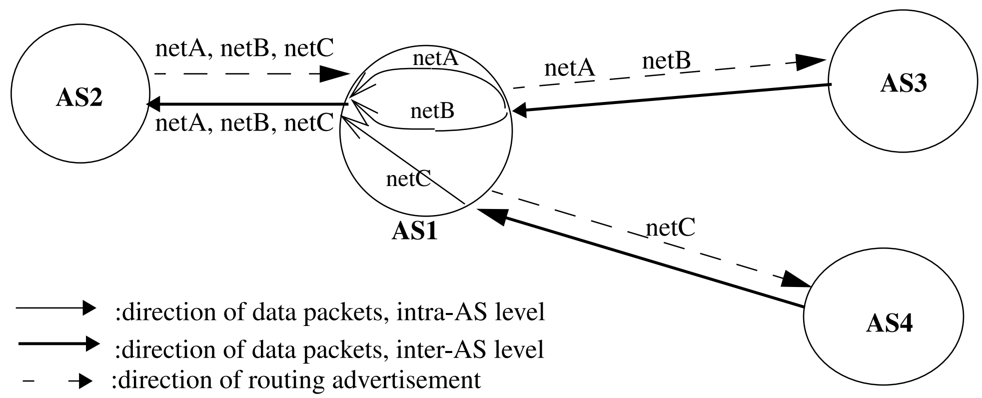


Figure 1. networks are the smallest granularity for routing at intra-AS level; ASes are the smallest granularity for routing at inter-AS level

3.4 Topological spatial locality and non-uniformity

We now distinguish between two attributes of the traffic patterns: topological spatial locality and non-uniformity of traffic.

We define the *topological spatial locality* of the traffic seen by a router, denoted E, as the likelihood that packets originating from different endhosts on the same network (or the same AS) go to endhosts on another network (or AS). This variable can be quantified by dividing the number of flows between endhosts by the number of jumbo flows (networks or ASes) observed for a given period of time. The value of E is in the range of $[1, \infty)$. Effectively, E is the reduction factor in states if packets are aggregated into jumbo flows. A value approaching infinity means that traffic between this network/AS and other networks/ASes shows a very high topological spatial locality, and aggregating endhost flows into jumbo flows is effective in reducing the number of states involved. We can visualize that individual flows originating from this network/AS show a topological closeness to each other, that is, they are in fact destined to another network/AS. A value close to 1 means that the flows between endhosts on this network/AS and other endhosts show a "fan-out" pattern; that there is no topological closeness among the destination endhosts.

We define *flow concentration*, denoted *P*, as the ratio of packets belong to this flow to the total number of packets observed for a fixed time interval. If we consider the set of all flows observed within a time interval, then the distribution of flow concentration can be characterized by either a frequency distribution function or a cumulative distribution function. For a given flow concentration *P*, the value of *F(P)* is defined as the ratio of the number of flows with flow concentration equals to *P*, to the total number of flows observed. The flow concentration cumulative distribution *F(P)* is defined as the ratio of the number of flows that have flow concentration less than or equal to *P*, to the total number of routes. That is,

$$F(P) = \sum_{m \le P} F(m) \qquad \text{Eq. 1}$$

It should be noted that the topological spatial locality and the flow concentration distribution are independent. Topological spatial locality depends on the location of the observation point. It describes the nature of the destination patterns for flows originated from a given network/AS. Flow concentration, on the other hand, describes the nature of the distribution of packets in different flows observed at a particular observation point.

4.0 Inter-AS traffic patterns

4.1 Experimental Setup

We use two sets of traces. The first set, which we call "vBNS traces" are taken from nine research and educational institutes connected to the vBNS [vBNS]: a nationwide network that supports high-performance, high-bandwidth research applications as a result of a collaboration between NSF and MCI. Some of those institutes deploy OCXMON monitors, traffic measurement and monitoring devices provided by NLANR [NLANR], to make desensitized Internet traces available for analysis and studies. Using the classification presented in section 3.1, the nine sites where we had taken traces from can be classified as class 4 ASes, with the exception of FXW and MRT, which are class 2 and 3 ASes, respectively. We have an hour-long trace each from the nine sites on Friday, January 15th, 1999 at 15:00 PST. Early studies showed that Internet traffic fluctuates on daily basis, and there is a significant decrease in traffic on weekends. The traces are taken on a weekday to be more representable.

The second set of traces are taken from MCI's commercial network at MAE-West. It comes from one direction of an OC12c link between two of the three MAE West ATM switches, carrying several inter-ISP peering sessions. The trace was taken between approximately 22:22 and 23:22 UTC (14:22-15:22 PST) on Wednesday, January 27th, 1999. The link was carrying, on average, about 60Mbps of traffic during that period.

TABLE 2. vBNS traces information and the MCI trace information

Sites name	Description	duration (sec)	Number of Pkts	Number of Bytes
ANL	Argonne National Labs	1648	1,371,854	751,628,957
FLA	U. of Florida, Miami, Florida State U.	3600	3,294,418	1,852,493,352
FXW	Original FIX-West FDDI ring	3599	43,075,924	16,934,334,898
MRT	Merit Computer Networks in Michigan to AADS NAP in Chicago	3600	58,980,772	23,835,022,940
NCL	MCNC, North Carolina State U., U. of North Carolina, Duke U.	3602	6,919,914	3,485,306,669
NRN	NASA Research and Education Network, AIX connection, NASA Ames, California	3600	33,600,336	17,581,244,607
ODU	Old Dominion University, Virginia	3600	4,808,121	1,380,815,799
OSU	Ohio State University	3598	32,765,590	13,554,573,762
SDC	San Diego supercomputer center	3600	36,044,403	10,910,468,426
MCI	MCI MAE-West OC12 link	3600	45,382,571	21,071,510,683

4.2 Analytical Tools

We wrote a program, *fs2jumboflows*, which maps packets into endhost flows, jumbo flows between networks, and jumbo flows between ASes. To map packets into endhost flows, we use a timeout definition of flows, with a timeout of 32 seconds. To map an IP address into its respective network number, we obtained a global routing table from MERIT [MERIT], and used a longest-trie-based prefix matching algorithm [WVTP97].

To determine the appropriate AS, we use the global BGP routing table from the Default Free Zone. This table is collected daily from route servers and published daily. We consolidated different views of routing tables from five route servers: there were 43,154 unicast routes advertised on January 15th, 1999, and 43,703 unicast routes advertised on January 27th, 1999. These BGP routing tables include the ASpath for every network number. For any BGP routing advertisement, an AS prepends its own AS number in the beginning of ASpath, so the last AS in ASpath is the origin AS of the network. When we built the trie structure for all 43,154 network prefix entries, we also include their corresponding origin AS as a separate field, so for a given IP address, we can also find out the origin AS it belongs to. This allows us to map an IP address to the corresponding originating AS number.

4.3 Results

4.3.1 Topological Spatial Locality

Using the definition in Section 3.2.3, we calculate the topological spatial locality factor for each site in the vBNS traces. The results, including the average packets per flow, are listed in Table 3.

TABLE 3. Topological spatial locality factor, *e* for 9 sites in vBNS traces and the MCI trace. timeout = 32sec.

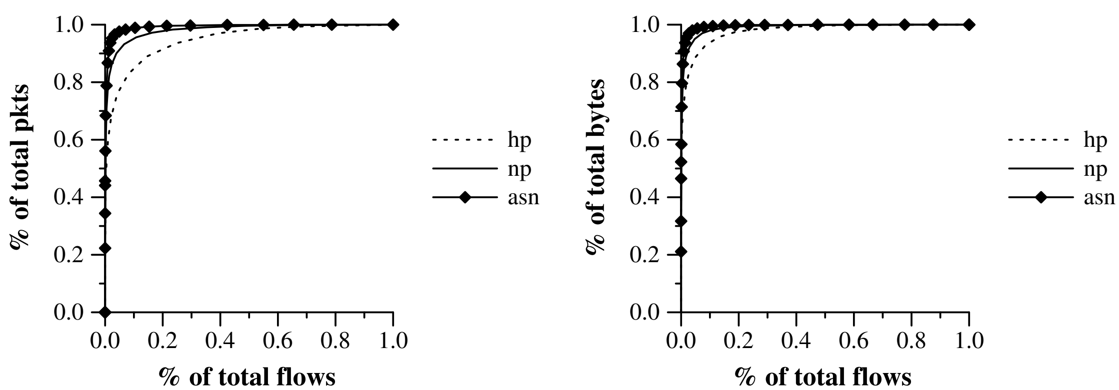
Site	# of endhost flows	pkts per flow	# of networks jumbo flows	pkt per jumbo flow	E for net-work	# of AS jumbo flows	pkts per jumbo flow	E for ASes
ANL	16778	81.77	7312	187.62	*2.294*	2662	515.35	*6.302*
FLA	14956	220.27	5663	581.74	*2.641*	3624	909.06	*4.127*
FXW	1782502	24.17	474871	90.72	*3.753*	118726	362.86	*15.01*
MRT	1231978	47.88	191998	307.19	*6.416*	27291	2161.18	*45.14*
NCL	76913	89.97	17523	394.90	*4.389*	3060	2261.4	*25.14*
NRN	19496	1723.4	6375	5270.64	*3.058*	1941	17310.8	*10.04*
ODU	56705	84.79	20525	234.26	*2.762*	11126	432.15	*5.096*
OSU	458610	71.45	117455	278.96	*3.904*	28797	1137.8	*15.924*
SDC	536060	67.24	146426	246.16	*3.660*	50364	715.68	*10.643*
MCI	795007	57.08	183718	247	*4.327*	40639	1116.72	*19.562*

As expected, the number of flows and number of packets per flow observed at different sites is dependent on the idiosyncrasies of the sites. For example, the FXW (FIX-West) trace consists of a large amount of DNS traffic because it is close to many root DNS servers. This explains why the average packets per flow is alarmingly small: there are many short DNS queries in the FXW trace. NRN is another outlier with a very high packets per flow. This is because there are a significant amount of active multicast flows during the trace, each consists of millions of packets. This brings up the average number packets per flow despite the fact that majority of the flows are much smaller.

As observed in those traces, the majority of the endhost flows are usually small, consisting mostly of tens of packets. Aggregating along the networks increases the number of packets per flow, and consequently, reduces the number of states needed to keep the flow information. There could be a number of factors influencing the rate of reduction *E*: the time when the trace was taken, the type of active flows during the trace, the specific location of the site (for example, whether the site hosts some major web servers). Regardless of those factors, the reduction in states when aggregating endhosts flows into network jumbo flows is between 3 and 6. Aggregating along the ASes will increase the average number of packets per flow to hundreds (with the exception of FXW, which was explained above). As expected, the reduction factors for ASes are larger than those for networks, because there could exists some topological spatial locality between network jumbo flows, so that they can be aggregated into one AS jumbo flow. Typically, aggregating endhosts flows into AS jumbo flows results in a reduction factor between 5 and 25.

4.3.2 Non-Uniformity

We did a non-uniformity study using the MRT trace. We plot the cumulative flow concentration function, as defined in Section 3.2.3, for host pair flows (hp), network pair jumbo flows (np) and AS pair jumbo flows (asn). Figure 2 shows the graphs in packets and bytes.



pkts cumulative distribution   bytes cumulative distribution

Figure 2. Non-uniformity distribution in flow concentration (MRT)

First, all three cases, hp, np and asn, display a highly non-uniform distribution of flow concentration. That is, a very small number of flows have a very high concentration of packets and bytes, whereas the majority of flows carry very few packets. Second, aggregating individual host flows (hp) along coarser granularity as networks (np) and ASes (asn) tends to accentuate such non-uniformity. For example, 10% of host pair flows account for 59.8% of packets, 9% of network flows accounts for 77.1% of packets, and 9% of AS flows account for 86.7% of the total packets.

This non-uniformity is observed across the nine vBNS traces and the MCI trace. MRT is one of the more typical cases in this regard, traffic in a few other traces is completely dominated by a handful multicast flows, and has a more extreme non-uniformity. More interestingly, the graphs above look remarkably similar to that reported in 1974 in [KleNay74], despite tremendous development and changes on the Internet during the intervening 25 years.

4.3.3 Correlation Between Flow Concentration and Sending Rates

There are only two possible ways for a flow to have a high flow concentration: either it lasts a long time, or it sends at a fast rate. We suspect it is the combination of both factors which determines flow concentration. We calculate the average flow concentration and the average duration of flows for the top N percent of the flows (N ranging from 0 to 100). We plot the distribution of average flow concentration, in packets and bytes, and average duration of flows against the cumulative distribution of flows, and show the graphs in Figure 3. The Y axis is in logarithmic scale. We observe that flows with high concentration of packets and bytes tend to last long as well.

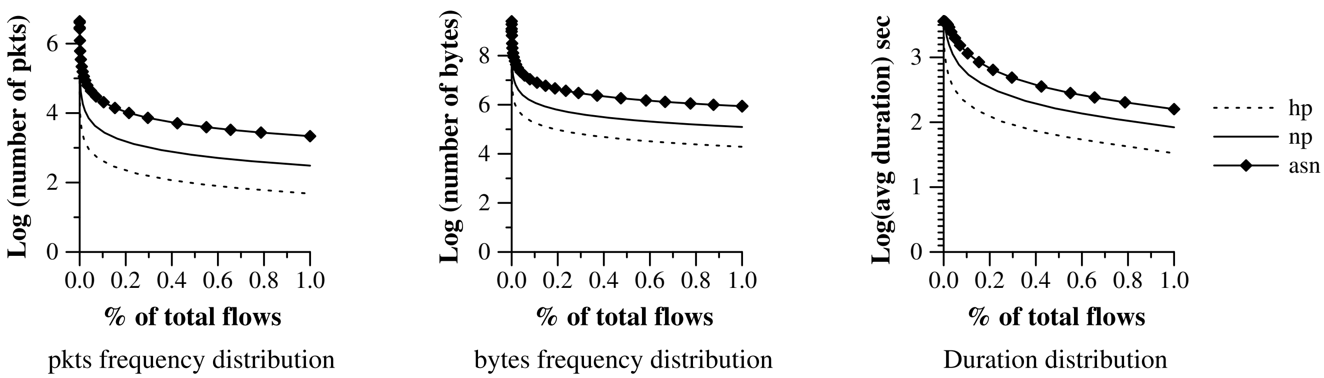


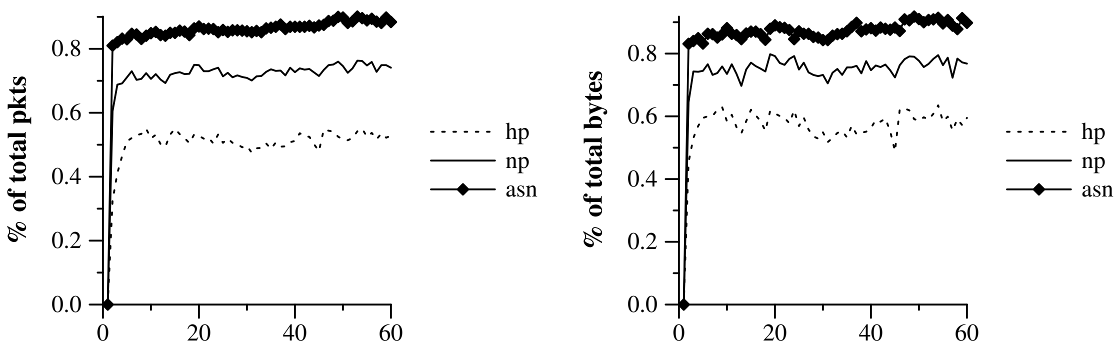
Figure 3. Cumulative distribution of avg. packets, bytes and duration per flow

Suppose a router wants to apply some QoS mechanisms to a limited set of flows, and the ideal choices of such flows are those that have the largest flow concentration by the time the flows terminate. However, at any instant in time, a router has no knowledge of how long an existing flow is going to last. The only heuristic a router can use is a flow's past sending rate. If, for example, a high past sending rate is likely to predict that a flow is going to be a highly concentrated flow by the time the flow terminates, then it is a safe bet for routers to start specially treating those particular flows. Therefore, we study the correlation between the concentration of flows and their average sending rate. The average sending rate of any particular flow is calculated using a low pass filter with a weight w of 0.5, i.e.,

$$avgRate = w \bullet currentRate + (1 - w) \bullet avgRate \qquad \text{Eq. 2}$$

We use a 60-second-timer mechanism to examine flow states. Whenever a timer goes off, all non-expired flows are checked, and their average rates are updated. The flows are expired by a different mechanism: if no packets have arrived in the last 32 seconds on this flow, the flow is automatically timed out. We then keep track the top 10% of flows with the highest sending rates, and calculate the packets and bytes they contributed as a percentage of entire traffic on a minute by minute basis. This gives us a time series of traffic contributed by the top 10% of flows at hp, np and asn granularities, which is shown in Figure 4.

We observe that the top 10% hp flows will amount to close to 59.8% of total pkts and 73% of total bytes, throughout time. When aggregated, the top 9% of jumbo flows accounts for 77.1% of packets (np) and 86.7% of packets (asn). This is consistent with our finding from previous section that the packet distribution among flows is highly non-uniform, and aggregation makes this phenomenon more extreme. Another observation is that the top 10% of the jumbo flows have a smaller variation in their aggregated traffic than the top 10% of endhost flows.



Time series of top 10% flows

Time series of top 10% flows

Figure 4. Time Series of MRT trace

A more subtle point that complements the above graph is made in the following two tables. We show the average number of active flow states a router would have to keep, and the change rate of flow states (i.e, establishing and tearing down flow sates per second) across all flows, and for only the top 10% flows. In other words, we know that the composition of top 10% of flows will change throughout time, but we expect that the top flows––those sending at higher rate––also last longer, so the change rate for top 10% of flows is much smaller than the average for all flows. Another way to visualize this is that there is a non-uniform distribution across the lifespan of flows, and those sending at high rates are likely to last long as well. There values are directly correlated with the memory and CPU cycles needed to maintain those flow states.

TABLE 4. Flow statistics for all flows, and top 10% flows (MRT)

	Across all flows		Across top 10% of flows	
	active flows on entry	**flow change rate (#/sec)**	**active flows on entry**	**flow change rate (#/sec)**
hp	21195	678.77	2119.5	18.725
np	5834.6	105.08	583.46	1.26
asn	1381.3	14.77	138.13	0.109

Table 4 shows the results for the MRT trace, in which the aggregate sending rate is 6.6Mbytes/sec. The left half of the table shows the number of active flows states that would need to be kept, and the average change rates of flow states for hp, np and asn. The right half of the table shows the counterpart values for only the top 10% of the flows. If the timespan distribution of flows were uniform, we would have expected to see the flow change rates in the right half of the table to be exactly 10% of the corresponding values in the left half of the table. But in fact, the values on the right half of the table are much smaller (by almost a factor of 10), suggesting that flows with high concentrations are those both last long (persistent in time) and send at high rate (large in volume).

TABLE 5. Flow statistics for all flows and top 10% flows (MCI trace)

	Across all flows		Across top 10% of flows	
	active flows on entry	**flow change rate (#/sec)**	**active flows on entry**	**flow change rate (#/sec)**
hp	13965.7	431.198	1396.57	15.77
np	5031.5	99.05	503.15	1.901
asn	1857.7	21.708	185.77	0.192

Table 5 shows the results from the MCI trace, in which the aggregate sending rate is 5.8Mbytes/sec. We can make essentially the same observations as those for the MRT trace.

4.3.4 Conclusions

We summarize our observations as follows:

- a) Endhost flows have a non-uniform traffic distribution among them, on the order of 10% of the flows accounting for 90% of the traffic.
- b) Two factors contribute to the non-uniform distribution of traffic: high sending rate and long timespan. We observe that flows which are sending at high rates tend to last long as well, resulting in a non-uniform distribution in terms of sending rates, timespan and volume among flows. This suggests that in order to identify flows with high flow concentration, we could use a high sending rate as a heuristic.
- c) Aggregating endhost flows into coarser granularities like networks and ASes accentuates such non-uniformity. If states were only maintained for jumbo flows, then we expect to see 1) a reduction in states of active flows; 2) smaller variations in sending rates; 3) smaller rates in establishing and tearing down states; 4) bigger flow concentration.

5.0 Discussion

5.1 Aggregation Among Networks/ASes

It should be realized that, due to the law of large numbers, when traffic is aggregated, we expect to see smaller variation in sending rates, a reduction in active flows, and a corresponding smaller rate of change. For example, the extreme case of aggregation is in fact, a FIFO queue for all packets, which has an active flow of 1. But this kind of aggregation is not useful engineering wise. In this section, we want to offer arguments for aggregating along the networks and AS granularities.

The argument for aggregating along networks is purely technical. Current Internet routing is done at the granularity of networks, so that packets aggregated between two end networks will still be forwarded together, just like packets between two endhosts. In the current Internet, QoS mechanisms applied to traffic are dependent on the fact that packets are forwarded together because packets taken different routes or put into different queues will likely to result in mis-ordering of packets arriving at the endhost, and current TCP interprets them as congestion signals, rather than merely mis-ordered packets, resulting in performance degradation.

In contrast, the argument for aggregating along Autonomous Systems is, at this point, non-technical, because current routing is done at a finer granularity than ASes as discussed in Section 3.2. However, it has the potential to be a technical problem, since it is easy to imagine that a transit AS has a routing policy of injecting routes into its internal IGP such that all packets between two ASes are forwarded together inside this AS (i.e, the case depicted in Figure 1 does not happen, meaning that packets destined to networks A and B are treated together inside the AS). In essence, this makes the granularity of ASes visible to IGP and requires IGP to treat packets from the same AS and destined to another AS together. If this can be done, then the transit AS can, at its boundary, aggregate packets into jumbo flows between ASes, knowing that they would be forwarded together. We will talk about possible mechanisms can be applied to those jumbo flows in Section 5.2.

On top of the technical arguments presented above, there are also business concerns. Resource allocation has to be done at a certain granularity, and for a commercial network, the ultimate goal of resource allocation is charging and billing. The granularity chosen for resource allocation in provisioning of a network will ultimately be used as the same granularity for charging and billing. ASes, chosen to mirror external administrative entities, are natural subjects for charging and billing. Similarly, network and IP host level granularity are useful if they in fact mirror, or can be used to a first approximation, to represent external entities like a department, a group or an individual.

5.2 Implications on QoS Mechanisms

One of the most pressing concerns about providing QoS mechanisms to network traffic is scalability of such mechanisms, especially at more aggregated points of the network. Efforts like Diff-Serv [Diff-Serv] attempt to separate policies from mechanisms by using simple queueing mechanisms—–preferential treatment of packets with certain TOS patterns—–in the middle of the network, and applying policies—–tagging packets with certain TOS patterns—–at the edge of the network. Diff-Serv assumes that networks are provisioned well as a whole to handle the traffic injected from different ingress points.

Results from our study suggest that by taking advantage of the topological spatial locality at aggregated points of the network, and by taking advantage of the non-uniformity in traffic distribution, we could significantly reduce the amount of state needed to provide better performance guarantees for the vast majority of traffic. This, combined with separating policies from mechanisms, can simplify the mechanisms needed in internal routers of a transit network, also scale well. To see this, consider the transit AS depicted in Figure 5, which contains two different technologies. The first technology is a set of internal routers that deploy Diff-Serv's preferential dropping algorithms (these routers are denoted R), and the second technology is a switch-based network like ATM or

Frame Relay. There are border routers, denoted BR, at the edges of the AS. Both technologies used inside the AS can benefit from the reduced state afforded by aggregation and non-uniformity.

In the case of the Diff-Serv routers, the border routers of the AS can keep flow states for––and give preferential treatment to––only a select number of flows within the AS. For example, if traffic is aggregated into AS jumbo flows, and tagging mechanisms are only applied for the top 10% jumbo flows, then the states needed to provide for 90% of traffic using Diff-Serv at the ingress points (border routers) of a transit domain are only 1857 states with a change rate of 0.192/sec (MCI trace).

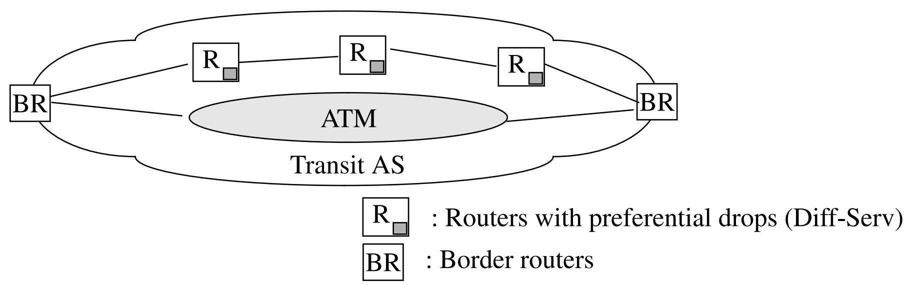


Figure 5: Taking advantage of non-uniformity in traffic

In the case of the ATM network, we have chosen aggregation granularities to be consistent with that of forwarding, so that packets mapped into the same jumbo flows will be forwarded together. Thus, such jumbo flows can be directly mapped onto *forwarding equivalent classes* in MPLS [MPLS], which uses fix-length labels to identify forwarding equivalent classes to speed up the forwarding process on those classes. On top of forwarding, queueing mechanisms like CBR, ABR can be applied to such jumbo flows within an ATM substrate to provide delay assurance.

Our results also suggest that flows with high concentrations tend to last long as well. This means that we could use an *implicit* reservation mechanism which estimates the sending rate of this jumbo flow at the ingress points and uses this rate to make queueing reservation within a transit domain for the future. This reservation does not have to be very accurate since it is only a prediction on the future sending rate, but it gives the internal transit domain routers an estimated rate for which it can make reservations for. In contrast to explicit reservations mechanisms like RSVP [RSVP], this does not require prior knowledge of sending rates from endhosts, nor compatible reservation protocol between endhosts and routers. Instead, the routers can manage to establish reservations for certain flows, based on its policies, as they deem appropriate.

6.0 Conclusions

This paper presents a recent study of inter-AS traffic patterns at a number of locations across the Internet. It confirms older studies that Internet inter-AS traffic displays a high non-uniform distribution despite the fact that Internet went through many topological, architectural changes in the intervening years. It also studied the topological spatial locality among endhost flows and suggest that aggregating traffic along networks and ASes results in a significant reduction in number of states.

In a typical case, 9% of all jumbo flows––referring to packets originates from a common AS, and destined to a common AS––comprises 86.7% of the traffic. We observe a non-uniformity distribution of flows in terms of sending rates, timespan and volume of packets. Such highly non-uniform distribution suggests only limited reservation states are needed to improve the performance of a significant amount of traffic. Many proposed QoS mechanisms can take advantage of this phenomenon to make the implementations scalable. A discussion on the implication of this phenomenon on different QoS mechanisms is offered.

Acknowledgment:

The authors would like to acknowledge Hans-Werner Braun of NLANR for his tremendous efforts in making traces available, in numerous discussions and in providing a conducive environment to make experiments possible. Thanks also go to Kimberley Claffy of CAIDA and Steve Feldman of MCI for their assistance in providing a commercial trace. We also would like to thank Prof. Leonard Kleinrock of UCLA for a speedy provision of a copy of his 1974 paper.

References:

[CAIDA] http://www.caida.org/

[Chinoy91] Chinoy, B., "Dynamics of Internet Routing Information", Proceedings of SIGCOMM'93, pp. 45-52, September, 1993.

[CBP95] Claffy, K., Braun, H., Polyzos, G., “A parameterizable methodology for Internet traffic flow profiling”, IEEE JSAC, 13(8), pp. 1481-1494, October 1995.

[Clark88] Clark, D., “The design philosophy of the DARPA Internet protocols”, in Proceedings of ACM SIGCOMM ‘88, August, 1988, pp. 16-19.

[Diff-Serv] Nichols, K., Blake. S., "Differentiated Services Operational Model and Definitions", IETF Internet draft, draft-nichols-dsopdef-0.0.txt, February, 1998.

[ERH92] Estrin, D., Rekhter, Y., Hotz, S., "Scalable Inter-Domain Routing Architecture", Proceedings of SIGCOMM ’92, pp. 40-52, August, 1992.

[GovRed96] Govindan, R., Reddy, A., "An Analysis of Internet Inter-Domain Topology and Route Stability", University of Southern California Computer Science Dept. Technical Report, http://www.usc.edu/dept/cs/tech.html

[KleNay74]Kleinrock, L. and W. Naylor, "On Measured Behavior of the ARPANetwork", AFIPS Conference Proceedings, Vol. 43, National Computer Conference, Chicago, Illinois, May 1974, AFIPS Press, Montvale, New Jersey, pp. 767-780, 1974.

[LMJ97] Labovitz, C., Malan, G., Jahanian, F., "Internet Routing Instability", in Proceedings of SIGCOMM’97, September, 1997.

[MERIT] MERIT Network Inc. Internet Performance and Measurement and Analysis Project, http://www.merit.edu/ipma/

[MPLS] Callon, R., Doolan, P., Feldman, N., Fredette, A., Swallow, G., Viswanathan, A., "A Framework for Multiprotocol Label Switching", Internet-draft, http://www.ietf.org/internet-drafts/draft-ietf-mpls-framework-02.txt

[NLANR] NLANR Measurement and Operations Analysis Team, http://moat.nlanr.net/

[Paxson96] Paxson, V., "End-to-End Routing Behavior in the Internet", Proceedings of SIGCOMM ’96, Stanford, CA.

[Rek95] Rekhter, Y., "Inter-Domain Routing: EGP, BGP and IDRP", in *Routing in Communications Networks*, M. Steenstrup, editor, Prentice-Hall, 1995.

[Rek99] Rekhter, Y., private email exchange, 1999.

[RekChi92] Rekhter, Y., Chinoy, B., “Injecting Inter-Autonomous system routes into Intra-Autonomous system routing. A Performance Analysis”, SIGCOMM92, vol 22, No. 1

[RFC896] Rosen, E., "Exterior Gateway Protocol (EGP)", RFC896, October, 1982.

[RFC1771] Rekhter, Y, Li, T, “A Border Gateway Protocol 4 (BGP-4)”, RFC1771, March 1995

[RSVP] Braden, R., Zhang, L., Berson, S., Herzog, S., Jamin S., "Resource Reservation Protocol (RSVP) Version 1 Functional Specifications", IETF RFC2205. September, 1997.

[vBNS] http://www.vbns.net/

[WVTP97] Waldvogel, M., Varghese, G., Turner, J., Plattner, B., "Scalable High Speed IP Routing Lookups", Proceedings of SIGCOMM ’97, September, 1997, Cannes, France.