Pop-Level and Access-Link-Level Traffic Dynamics in a Tier-1 POP

Supratik Bhattacharyya, Christophe Diot, Jorjeta Jetcheva, Nina Taft

Sprint Advanced Technologies Laboratories
Burlingame, CA USA

Abstract— ***In this paper, we study traffic demands in an IP bacbkone, identify the routes used by these demands, and evaluate traffic granularity levels that are attractive for improving the poor load balancing that our study reveals. The data used in this study was collected at a major POP in a commercial Tier-1 IP backbone. In the first part of this paper we ask two questions. What is the traffic demand between a pair of POPs in the backbone? How stable is this demand? We develop a methodology that combines packet-level traces from access links in the POP and BGP routing information to build components of POP-to-POP traffic matrices. Our analysis shows that the geographic spread of traffic across egress POPs is far from uniform. In addition, we find that the time of day behaviors for different POPs and different access links also exhibit a high degree of heterogeneity. In the second part of this work, we examine commercial routing practices to assess how these demands are routed through the backbone. We find that traffic between a pair of POPs is engineered to be restricted to a few paths and that this contributes to widely varying link utilization levels. The natural question that follows from these findings is whether or not there is a better way to spread the traffic across backbone paths. We identify traffic aggregates based on destination address prefixes and find that this set of criteria isolates a few aggregates that account for an overwhelmingly large portion of inter-POP traffic. We demonstrate that these aggregates exhibit stability throughout the day on per-hour time scales, and thus form a natural basis for splitting traffic over multiple paths to improve load balancing.***

Jorjeta Jetcheva is currently at Carnegie Mellon University. This work was done while she was at Sprint ATL.

I. Introduction

Internet backbones continue to grow at explosive rates, fueled by the bandwidth demands of new applications and by the advent of faster access technologies. To accomodate such growth while preserving the robustness of the network, most IP backbone operators have chosen a simple approach to traffic engineering: overprovisioning. Overprovisioning is the adopted approach because very little information exists today about the dynamics of the traffic in an IP backbone. This is primarily due to the lack of measurement infrastructure and techniques for collecting and processing data from backbones. To address this deficiency, we study traffic traces collected at a Point of Presence (POP) in a commercial Tier 1 IP backbone network. A passive monitoring system is used to collect packet-level traces on a number of access links within the POP [1]. The data is then analyzed offline in order to understand the dynamics of traffic entering the backbone at this POP. We describe a methodology for extracting information about routing and traffic flow within the backbone. This methodology forms the basis for building components of POP-to-POP level traffic matrices, which are key to studying a variety of traffic engineering and routing issues. We investigate how much can be said about the traffix matrix just from the data collected at a single POP.

It has been observed ([2], [3], [4]) that obtaining information about traffic patterns in both time and space is critical for most traffic engineering functions. Traffic engineering typically operates on long time-scales such as minutes, hours, weeks or longer. Examples of traffic engineering functions include dimensioning, provisioning, route optimization, where to best add new customer links, load balancing policies, designing POP architectures, and selecting failover strategies. The particular application deter-

mines the level of information needed about traffic patterns. Since IP networks do not typically generate feedback state information, traffic engineering has to rely on traffic measurements [2]. It has been observed that simulation data cannot provide substitutes [5]. Therefore collecting traffic measurements spanning multiple hours in order to build network-wide views of the traffic flows is central to being able to efficiently engineer an IP backbone.

A network-wide view is typically expressed in the form of a traffic matrix ([6], [3], [7]). A variety of information can be represented in it. For example, the traffic volume captured in the matrix can refer to any level of flow granularity. A traffic matrix also has an associated time granularity that specifies the measurement interval over which bandwidth measurements were averaged. The choice of exactly what is represented in the matrix depends upon the traffic engineering task to be performed with this matrix. In a POP-to-POP traffic matrix the rows represent ingress POPs and the columns represent egress POPs. Since our data was collected at a single POP in our network, we build one row of a POP-to-POP traffic matrix. Due to the cost of such equipment, the enormous difficulties involved in deploying the equipment in commercial backbones and the scarce availability of this backbone data, even this component of a traffix matrix constitutes a significant amount of useful information.

We decompose and study this data along a number of different dimensions. The work in this paper can be viewed as a search of answers to the following questions, each of which logically follows from the next. In the first part we ask, what is the traffic demand between a pair of POPs? How stable is this demand? The traffic matrix compiled in this part only describes the *demand* or how much traffic wants to go from one POP to another; it says nothing about how the traffic is routed. Thus in the second part we ask, how are these demands routed in our commercial backbone? Are link utilization levels similar throughout the backbone? Our observations from these two parts are that traffic is highly non-uniform in a geographic sense yet the ranking of POPs (in terms of volume) remains fairly stable in time; and that few routes are used and link utilization levels vary widely throughout the backbone. These findings motivate the third part which asks, is there a better way to spread the traffic across the paths? And at what level of granularity should this be done?

For the first part, we proceed to study the partition of traffic throughout the backbone as follows. We examine incoming traffic at a single POP at different levels of granularity. First, we analyze the spatial characteristics of POP-level traffic. We discover a large disparity in the spatial distribution of the ingress POP's traffic across the egress POPs. Second, we break up the ingress POP's traffic according to access link, and examine the spatial distribution of the traffic from specific types of access links across the egress POPs. We find that the same disparity appears at this level of granularity. We compare the access links and find that they behave differently from one another with respect to three metrics considered. For example, we find that one cannot isolate a single probability distribution to characterize the geographical fanout of the traffic from access links. We also examine time of day behavior of the traffic at both the POP-level and access link level. We find that egress POPs can be ranked roughly into three categories (large, medium and small) based on the amount of traffic they receive from the ingress POP, and that POPs generally remain in the same category through the entire day. A stronger statement can be made about many of the POPs - if they are ranked by the volume of traffic they receive, they maintain their ranking throughout the day. We also find that at night the overall traffic load is reduced by only 15-50% depending upon the access link.

For the second part of our work, we begin by checking whether or not overprovisioning has led to a disparate use of resources on a network-wide basis. By examining both SNMP data, we do indeed find that the amounts of excess link capacities are inequitably distributed throughout our backbone. We then study IS-IS routing behavior to understand how IS-IS is engineered to influence path selection, and how the routing impacts the link utilization levels. We find that the backbone is carefully engineered using IS-IS weights to restrict traffic between POP pairs to a few paths even though many alternate paths exist.

In the third part of our study, we return to our traffic data to assess at what granularity level it is desirable to do load balancing. We want to determine a traffic granularity that defines a unit of flow (or stream) that could be rerouted on an alternate path. Having examined our data at both the POP-level and the access-link-level, we now study the data at the granularity level of destination address prefixes. We find that a small number of these aggregate streams, called *elephants*, generate a large fraction of the total traffic, while a large number of these streams, called

mice, generate a small fraction of the total traffic. The elephants and mice phenomenon has been observed before in Internet traffic at the inter-AS level [4], at the level of multipoint demands from one router node to a set of router nodes [3] and in the Internet as it was many years ago [8]. Here we demonstrate this phenomenon at the granularity level of specific prefixes. We also demonstrate the stability of these aggregates throughout the day. The stability of these elephants makes them well-suited as a basis for routing traffic on alternate paths and thus improving the load balance in the backbone.

The rest of the paper is organized as follows. Our measurement infrastructure is briefly presented in Section II. Section III describes a methodology for building a POP-to-POP view of traffic flow across the backbone, based on observations at an ingress POP. Our technique makes extensive use of BGP and IS-IS routing information. The space and time characteristics of traffic at the POP-level and the access link level are analysed in Section IV. In Section V we study IS-IS routing in order to understand how routing practices influence the partition of traffic across the backbone. In Section VI we aggregate the traffic based on destination address prefixes, and demonstrate the existence of the elephants and mice phenomenon at this granularity level. We analyze properties of these aggregates and discuss their application to load balancing. Section VII discusses related work, and Section VIII discusses some of the implications of our results and identifies directions for future work.

II. Measurement Infrastructure

The data used for this study was gathered from an operational IP backbone using the passive monitoring infrastructure described in [1]. The backbone topology consists of a set of nodes known as Points-of-Presence (POPs) connected together by high bandwidth backbone links. Each POP also locally connects customers through access links, ranging from large corporate networks to regional ISPs and webservers. Peering at a POP is provided either through dedicated links to another backbone (private peering) or through public Network Access Points (NAPs). Each POP has a two-level hierarchical structure (Figure 1). At the lower level, customer links are connected to access routers. These access routers are in turn connected to the backbone routers. The backbone routers provide connectivity to other POPs and to the peers. The backbone links

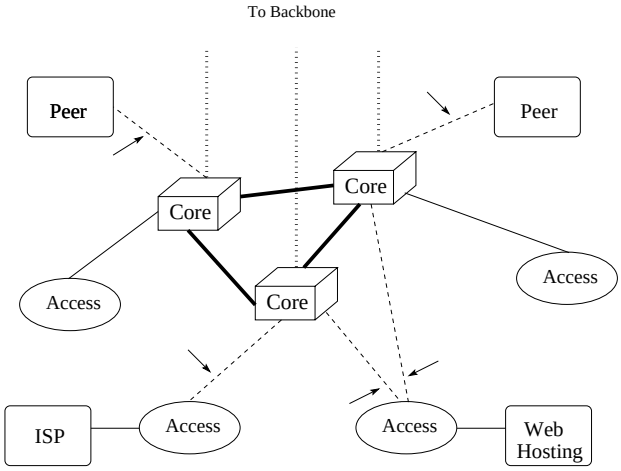

Fig. 1. Monitored POP Links/Architecture of a POP

Link	Trace Length (hours)	Trace Size	# Packets (10^6)
Peer 1	16	51 GB	853
Peer 2	24	47 GB	794
Web Host 1	19	51 GB	853
Web Host 2	13	51 GB	853
Tier 2 ISP	8	17 GB	284

Fig. 2. Summary of Data

connecting the POPs are optical fibers with bandwidths of 2.5 Gbps (OC-48). They carry IP traffic using the Packet-over-SONET (POS) protocol. The exterior and interior gateway protocols for the backbone are Border Gateway Protocol (BGP) and IS-IS respectively.

The infrastructure developed to monitor this network consists of passive monitoring systems that collect packet traces and routing information on the links located between the access routers and the backbone routers, or on the peering links. The monitoring systems tap onto the selected link using optical splitters, and collect the first 44 bytes of every packet on these links. Every packet record is timestamped using a GPS clock signal which provides accurate and fine-grained timing information. The format of the packet record is as follows.

- GPS timestamp : 8 bytes
- Size of record : 4 bytes
- Size of POS frame : 4 bytes
- HDLC header : 4 bytes
- IP packet header : 44 bytes

BGP tables were downloaded from one router[1] in the POP once per hour during the time the packet traces were

[1] all routers in the POP have the same view of BGP routes

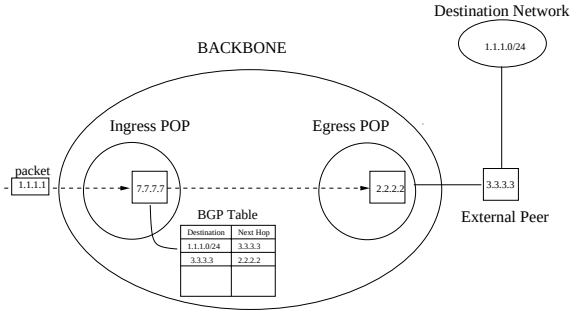

Fig. 3. Example of destination to egress POP mapping

collected. In this study, we used data from five access links, collected on August 9, 2000, starting at 10:00am US Pacific Standard Time (UTC-7). The arrows in Figure 1 indicate the monitored links used in this study. The table in Figure 2 provides a summary of our traces. The traces are of different lengths because packets were collected until the hard disk in each monitoring system filled up. Therefore heavily loaded links filled up the disk faster than lightly loaded links. We have collected many other traces during August and September 2000. The results in this paper have been verified against one other day. We present the data from a single day to avoid overloading the paper with excessive graphs.

III. Methodology

In this section we explain how we construct the row of our POP-to-POP traffic matrix representing our backbone traffic. This row in the traffic matrix corresponds to data that originates at the monitored POP (i.e., the ingress POP) and leaves the network through each of the other egress POPs (including itself). To do this, we need to map each packet received at the monitored POP, to the egress POP through which it leaves the network. All backbone routers participate in the BGP protocol and exchange information with each other about routes learned from routers external to the network, called external peers. This information is kept in the BGP table of each router and can be used to determine the last egress router for each packet destination. However, information on mapping an egress router to an egress POP is not readily available and has to be derived from the values of standard BGP attributes.

We start by illustrating with an example the methodology we use to determine the egress POP for a packet entering the backbone (Figure 3). Consider a packet with a destination address of $1.1.1.1$. Suppose that the BGP table at the ingress router (7.7.7.7 in the figure) for this packet, identifies the destination subnet for this packet as $1.1.1.0/24$. The BGP table entry at the ingress router, which corresponds to this subnet contains a *Next-Hop* entry which is typically the address of the external BGP peer from which a border router in our backbone first learned the route to the destination and injected it into the I-BGP[2]mesh. This border router is in the egress router for destination $1.1.1.1$, since it is the router that packets for subnet $1.1.1.0/24$ need to go through in order to reach the external peer on their way to the destination network. Suppose the address of the border router is $2.2.2.2$ and that the address of the external peer recorded in the *Next-Hop* entry in the BGP table is $3.3.3.3$[3]. The BGP table at $7.7.7.7$ also contains an entry for $3.3.3.3$ (or the subnet containing it), whose *Next-Hop* field identifies the address of the border router, i.e. the egress router for this packet. To find the egress POP for the packet destination, we use the BGP *Community Id* attribute which is an identifier shared by all routers at a POP[4] . This attribute is recorded in each BGP table entry and identifies the egress POP for all destinations that map to that entry. In our example, the *Community Id* allows us to identify the POP to which the egress router $2.2.2.2$ belongs.

However, that are many cases when the *Community Id* attribute for a route is not set to an identifier that specifies the egress POP (due to internal policies) for the BGP entry it belongs to. In such cases, we extract the *Originator* attribute for the route announcement to a given *Next-Hop*. The *Originator* attribute value corresponds to the address of a router in our backbone. In the above example, the router $2.2.2.2$ would be the *Originator* for the route to $3.3.3.3$. Querying the BGP table returns the *Community Id* attribute associated with the *Originator*, and hence the POP at which the *Originator* is located. This POP is the egress POP associated with the *Next-Hop* that we are interested in ($3.3.3.3$ in our example).

Note that there are a few cases, when BGP attributes

[2]backbone routers use BGP to exchange information about routes to external networks and policies internal to the backbone

[3]Typically a BGP table will contain a number of alternate paths for a destination subnet. However we consider here only the route chosen as the ”best” or ”preferred” route based on BGP policies and attribute values.

[4]This information is not available in response to “show ip bgp”. However, knowing all possible values for the community id attribute, it is possible to use the *“show ip bgp community x”* command to determine the attribute value for each route.

fail to reveal the POP name. In these cases, we perform a *Traceroute* to the *Next-Hop* router address associated with the BGP entry for the packet destination. We can extract the name of the last hop router within the backbone from the output of *Traceroute*, and derive the identity of the POP from the name. The name of each router at a POP contains a sub-string that is derived from the name of the city in which the POP containing the router is located (for example, a router's name in a POP in Miami would contain the string *mia*).

The complete algorithm for determining the egress POPs for destination networks in the BGP table is omitted due to space constraints.

Recall that the BGP tables we used were collected once an hour, a time-scale on which they have been observed to be relatively stable [9]. The number of unique *Next-Hop* entries in each table was on the order of a few thousands. Of these, about 98% were resolved to egress POPs using BGP attributes, and the rest were resolved using *Traceroute*. Overall, more than 99% of the destination networks in the BGP tables were resolved to egress POPs using our technique.

Once we obtain a mapping of destination networks to egress POPs, we can apply it to the packet traces to determine the traffic flowing from the monitored POP to each other POP. This task is analogous to the problem of performing lookups on packets arriving at a router to determine the egress port. For this purpose, we used the IP lookup technique described in [10]. This technique uses an LC trie-based longest prefix match and a software implementation is publicly available[5]. We modified this software to perform a longest prefix match on packets in a trace using our destination-to-POP map. The output consists of per-egress-POP packet traces. These can be further analyzed to understand the dynamics of traffic between any two POPs in the network. We have developed tools to subdivide the traffic between two POPs based on various criteria such protocol number, destination prefix, etc. Tools have also been designed and implemented, to study the variation of traffic on different timescales. These analysis tools were used to compute all of the results presented in this paper.

[5] http://www.nada.kth.se/ snilsson/public/soft.html.

IV. Traffic in Space and Time

A. Geographic Spread

We look at the geographical spread of traffic demands across egress POPs, or *fanout*, first at the POP-level and then on an access-link level. Since most of our traces span from 13 to 24 hours each (depending upon the link), we are also able to study the time of day behavior for these demands. Our goal in this section is to classify the basic behaviors we observe into a few categories, and to understand the range of behaviors that can occur. We are also interested in comparing different types of access links and different egress POPs to see if and where commonalities lie.

First we consider the traffic demands on all five access links together as one input stream. Note that this constitutes a significant portion of the input traffic at our monitored POP. Given the variety of access links chosen, this is also highly representative of the total input traffic entering the POP. The monitored POP is henceforth refered to as the *ingress* POP.

We use the methodology described in the previous section to classify all the packets in a trace by their egress POPs. We then determine the total number of bytes headed towards each egress POP using the packet length information in the IP header of each packet record. This gives us the fanout of traffic demands by volume (Figure 4). The values presented in this figure are bandwidth values that were averaged over the duration of the entire trace for every link. This fanout constitutes the row on our POP-to-POP traffic matrix.

For the purposes of display we have organized the POPs into 3 groups: the west, midwest and east regions of the United States. The monitored POP is located in the west coast of the US. For proprietary reasons the POPs are only identified with numbers. Within each of the 3 regions the ordering is arbitrary and does not have any geographic signifigance.

We observe that there are two POPs that are clearly dominant, and receive a large amount of traffic (over 35 Mbps). Among the remaining POPs about half receive quite a small amount of traffic (under 5 Mbps) and the other half receive a moderate amount of traffic (10-20 Mbps). Our data suggests that ingres POPs could be roughly categorized as *large*, *medium* and *small*, where (i) roughly the same number of POPs fall into the small and

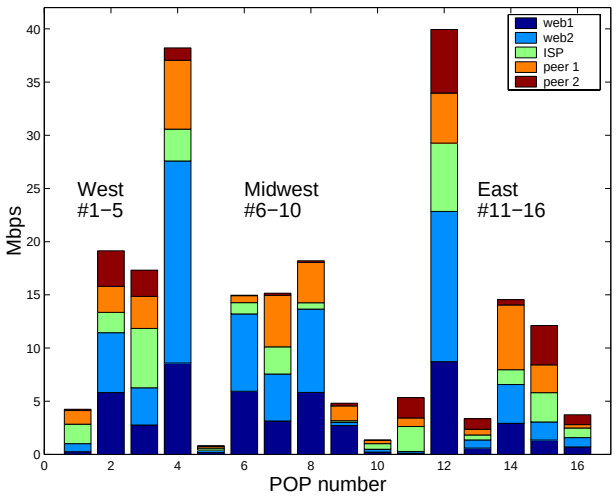

Fig. 4. Fanout of 5 Access Links Combined

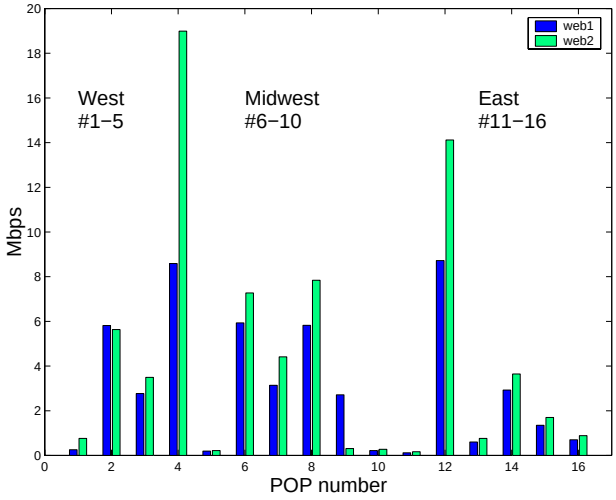

Fig. 5. Fanout of Web Host Links

medium categories and only a few fall into the large category; and (ii) each category carries approximately twice the volume of the category below it. This simple characterization will prove useful in our interpretation of other data below. (We will discuss the stability of these characteristics in the next subsection.)

Often in simulation environments, researchers assume a traffic matrix. In the past, in absence of data, the most common model is, given a source, pick a destination at random according to a uniform distribution. This histogram reveals that such an approach does not at all match Internet behavior. Moreover, thinking about how the Internet is designed, it is easy to understand why we see this non-uniform behavior. First, one would expect that some POPs would sink higher traffic demands than others because of their geographic location. For example, dominant POPs are expected to be located on the two coasts of United States because this is typically where international trunks terminate, and because the coasts are more heavily populated than the center of the country. Secondly, one would expect this distribution to exhibit a significant degree of variation. The volume of traffic an egress POP receives (from other ingress POPs) depends upon a large number of factors, such as the number and type, of customers and servers attached to it. Similarly, the amount of traffic an ingress POP generates can also vary enormously depending upon the number and type, of customers and servers, on its access links. Thus we expect the inter-POP flows to vary dramatically from one to another, and to depend on the (ingress POP, egress POP) pair.

Note that for the purposes of bandwidth prediction, the (ingress POP, egress POP) pair might represent a level of granularity that is too coarse for accurate traffic estimation. It is natural to hypothesize that the access links at the ingress POP may differ from one another, and may affect the traffic flowing to each egress POP differently. We thus next consider the fanout of traffic at the ingress POP on a per-access-link-type basis. To compare these links we considered three metrics: (i) the total volume of traffic per link (summing across all egress POPs); (ii) the max/min ratio of the average bandwidth headed towards an egress POP[6] and (iii) the distribution among the egress POPs. The total volume and max/min ratios are given in Table I. We see that the access links differ from one another with respect to these simple measures that span a range of values.

Figure 5 demonstrates that the rough categorization we applied to egress POPs at the POP-level (i.e., the large, medium and small categories) continues to hold at the level of input access link type. (The same is true for the peering links, however we exclude the fanout plot due to space restrictions.) In other words, a very small number (between 1-3) of POPs receive a large amount of traffic and the rest of the POPs are evenly split between the medium and small categories. To compare the fan-out of the different access

[6] In computing the max/min ratio we ignored the three smallest POPs for a given access link because there were typically a few POPs that receive a negligible amount of traffic and this creates ratios that are not representative.

Ingress Link	peer #1	peer #2	ISP	webhost #1	webhost #2
volume (Mbps)	40	22	32	50	70
max/min	13	50	13	35	63

TABLE I
COMPARISON OF ACCESS LINKS

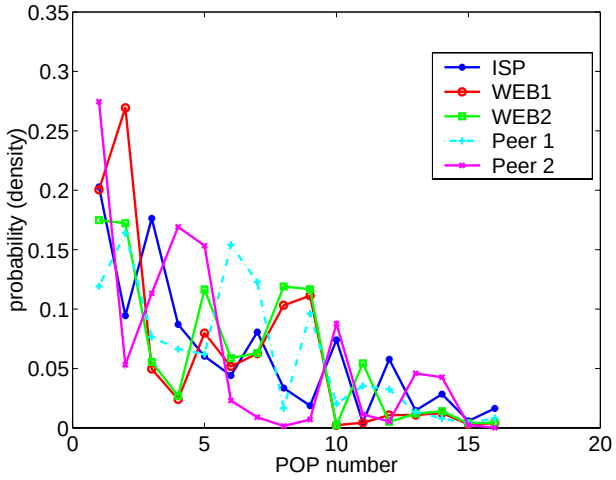

Fig. 6. Probability density of POP fanout per access link type

links numerically, we normalize and convert the fan-out information per link into a probability distribution. Let b_{ij} denote the average bandwidth that access link i sends to POP j during one day. For a given link i, the probability that a byte of data chosen at random gets sent to POP j is given by $P_i(j) = b_{ij}/\sum_j b_{ij}$. The density curves for each of the five links is given in Figure 6. The ordering of the POPs here is different than in the previous graphs, and thus it no longer represents an east/midwest/west organization. This ordering was selected to try to isolate a pattern.

To facilitate the discussion, we use the term *popularity* of a POP to refer to the likelihood that a byte of data from an ingress access link will be sent to that egress POP. On the one hand, we see a few similarities among the five links. The first two POPs are the most popular among all the access links. POPs #11-16 are fairly unpopular for all links. For all other POPs, the popularity ordering jumps around quite a bit for each link. For example, the likelihood that a packet on a given link will choose POP #4 can vary from 0.02 to 0.17. This graph indicates that POP #3 is most likely to be chosen by our ISP link, POPs #4 and #5 are most likely to be chosen by the peer 2 link, POPs #6 and #7 are most likely to be chosen by the peer 1 link, and #8 and #9 by the second web host link. In general, for POP's #3-#10, the likelihood of being chosen can vary about 10%. We believe that these differences are substantial and that the fanouts from the different links are sufficiently different so that one cannot conclude that there is a single underlying distribution that represents all the access links. Note that the categorization of egress POPs according to large/medium/small *is the same* for different access links. The access links differ in their geographic spread primarily in how they distribute traffic among the medium sized POPs.

We thus infer that when studying traffic demands for load balancing, and more generally, when designing bandwidth predictors for traffic engineering, the pair (ingress POP access link type, egress POP) should be explicitly accounted for rather than simply using the (ingress POP, egress POP) pair.

From this section, we conclude that in terms of geographic distribution there is a large disparity among the traffic sent to the egress POPs from a single ingress POP, and that the access links differ from one another significantly according to three different metrics. The exception is for the two web host access links; however, these two links carry traffic from the same client, which reinforces our notion that links generate different traffic demands based on their types.

B. Time of Day Behavior

In the previous section, the fanouts we examined were computed based on day-long averages. In order to examine the consistency of the fanout throughout the day, we look at inter-POP flows on an hourly basis. In Figure 7 we consider just four of our input links (because the 5th has too few hours) and examine the behavior throughout the day of three representative POPs, one in the large category, one in the medium and one in the small. First, we observe that the large POP is the most volatile, that the medium POP experiences a long slow small decline, and that the small POP remains fairly stable. We examined a number of other POPs and found this behavior to be consistent of POPs within their respective categories. Second, we observe that during the day the distinction between large,

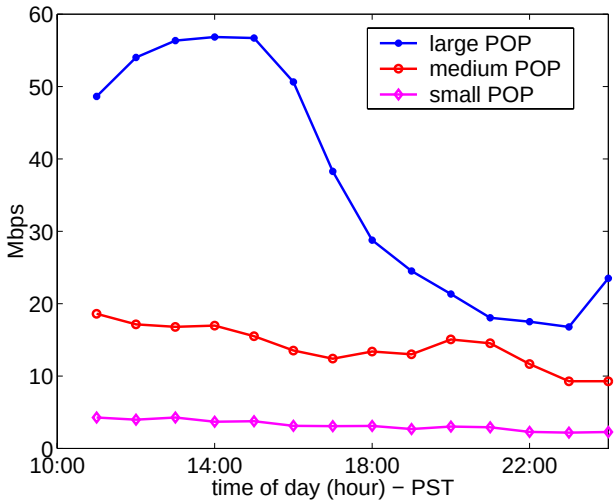

Fig. 7. POP-level Time of Day

trace	peer #1	peer #2	web host #1	web host #2
% *reduction*	30%	30%	50%	16%

TABLE II
NIGHT VS. DAY TRAFFIC

medium and small remains, whereas at night the large and medium POPs become less distinguishable.

Figure 7 indicates that some POPs do not experience much decrease in traffic volume at night, while others (particularly those in the large category) do. When we considered traffic volume between 10 AM to 6 PM (daytime) and 6 PM to 6 AM (nighttime) separately, we found that the nighttime peak traffic is about 30 Mbps, about half of the daytime peak. The average percent reduction on a per-link basis is shown in Table II. The table indicates that the average volume of night time traffic is anywhere from 15-50% less than the average volume of day time traffic. This is surprising since it is counter to the widely held belief that traffic in the backbone reduces by a few factors at night. This may arise from an increase in international traffic and night-time backup activities.

We now examine time of day behavior (Figures 8-11) at the access link level to examine the variations and if our previous observations hold at the access link level as well. In order to compare different access links, we provide a separate figure for each of four access links (identified in the figure caption). For each access link, we plot the hourly bandwidth averages for six different POPs, hence each curve corresponds to a single egress POP. Some of the curves on these graphs exhibit a sharp drop around 2:00am. This is due to maintenance activities at the POP. From these four plots we observe the following:

- A number of POPs have traffic that remains fairly constant throughout the day.
- A number of POPs experience a long slow decline of loading throughout the day.
- The most popular POPs are usually the most volatile.
- If we were to rank the POPs by volume received, then most of the POPs (excepting the few large ones) maintain their rank throughout the day.
- POPs can experience an increase at night (see Figure 10 and 11).

These observations are interesting in that they reveal counter-intuitive things about busy periods. Our experience from telephone networks leads us to expect peak period behaviors in time-of-day plots. These figures reveal that some POPs do not experience any busy periods, some POPs experience one busy period, and others can experience two.

We see that the category that an egress POP falls into can depend upon the access link. For example, consider POP #13. On the two peering links, this POP is a small one. On the two web hosting links, it would be considered a medium one. This indicates that the fraction of traffic than an egress POP draws from an ingress POP depends upon the number and type of input access links. An alternative way to see this is given in Figures 12 and 13. In these plots we compare the traffic destined for a single egress POP originating from each of the access links. This illustrates that an egress POP's behavior can differ dramatically depending upon which access link on an ingress POP it receives traffic from. For some access links, an egress POP receives a roughly constant amount of traffic while for others its traffic experiences peaks and dips over the course of the day. Thus the incoming traffic on an egress POP is directly dependent upon the type of access link at the ingress POP.

V. OBSERVATIONS ABOUT IS-IS ROUTING IN THE BACKBONE

In the previous section we examined properties of traffic *demands*, i.e. how much traffic wants to go from one end of our network to another end. This says nothing about how that demand is routed through our network. The inte-

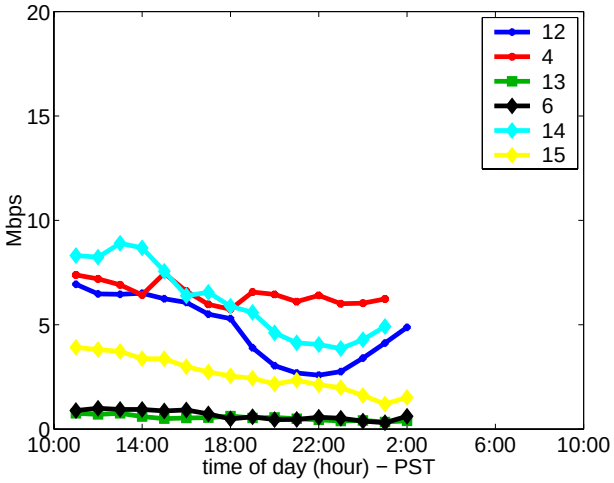

Fig. 8. Peer 1

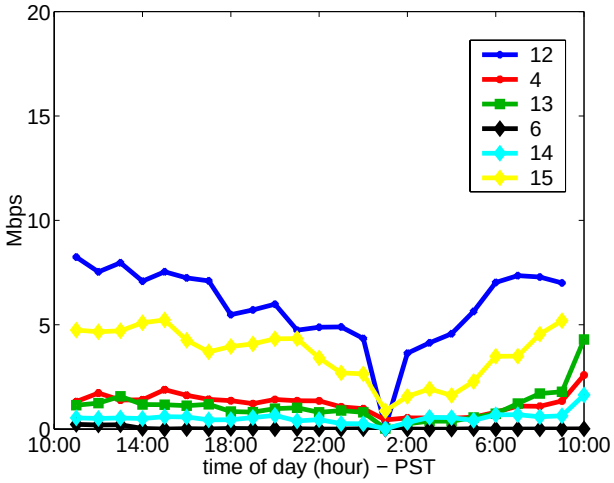

Fig. 9. Peer 2

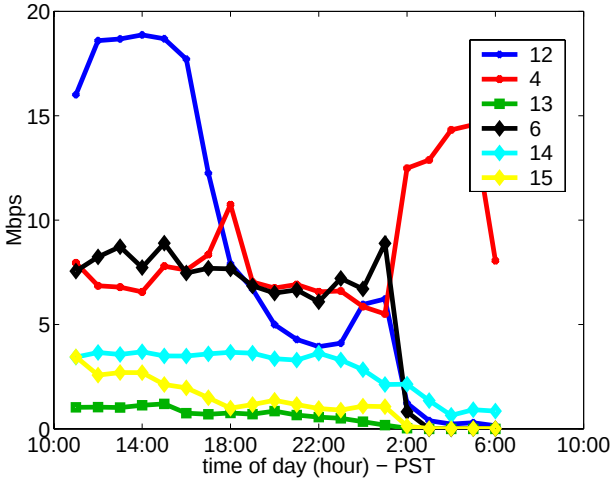

Fig. 10. Web Host Link #1

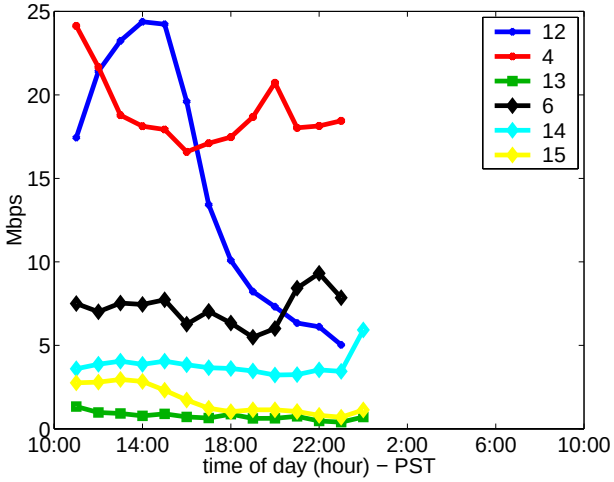

Fig. 11. Web Host Link #2

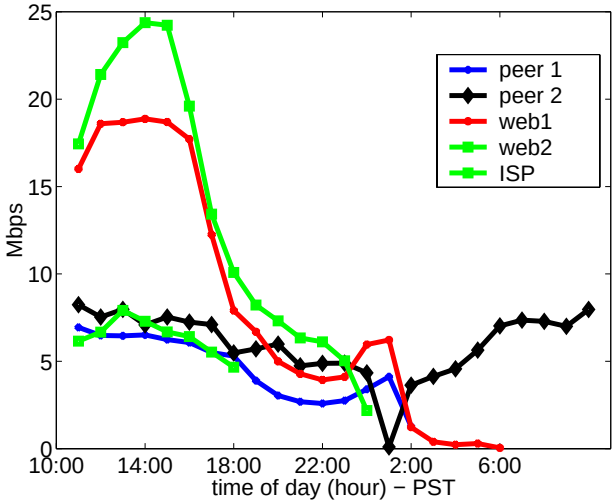

Fig. 12. East Coast POP

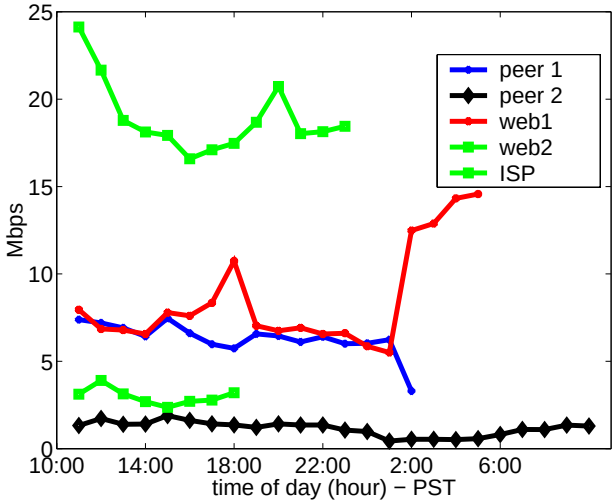

Fig. 13. West Coast POP

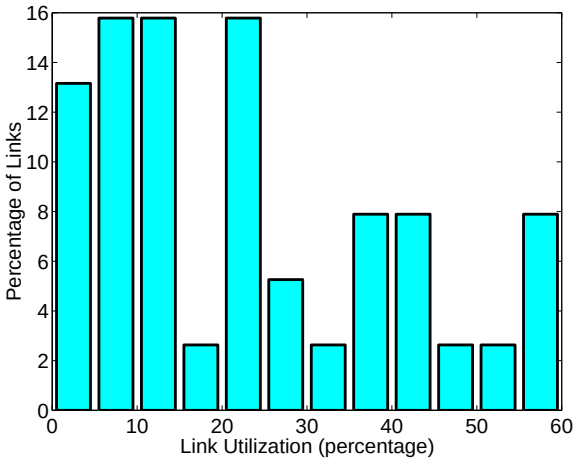

Fig. 14. Histogram of Link Utilization Across in the Backbone

rior gateway protocol used for routing traffic in a backbone has a direct effect on link load levels. We have studied the configuration of IS-IS, the internal gateway protocol used in our backbone, and have reached a few conclusions about current routing practices. Currently, the IS-IS link weights are handcrafted by network operations experts such that (i) the weights are chosen such that traffic between any ingress-egress POP pair is restricted to only a few paths through the backbone; and (ii) the intra-POP link weights can heavily influence the path chosen to traverse the backbone. While this approach has certain advantages such as ease of management, it may drive up link utilization levels on paths between POPs when the inter-POP traffic demands are heavy. In particular, we have found that for some heavy demand POP-pairs, a number of alternate paths exist, but many of them are underutilized, while a few have high utilization levels.

To get a better sense of the joint impact of the traffic demand and the routing on link utilization levels, we collected SNMP data on link load levels from all the backbone links in our network. Figure 14 provides a histogram of this data, averaged over an entire day. We find that the majority of the links have an average utilization under 25%, and that the link utilization levels can vary from 3% to 60%. This histogram reveals (i) the extent of link underutilization and (ii) the extent of the disparity in the utilization levels of the backbone links.

It is clear from these findings, combined with our findings from Section IV, that better load balancing schemes are needed in the network. There are different approaches to load balancing. Enforcing a change in load balancing via the IS-IS routing protocol has difficulties. IS-IS does not have the capability to balance traffic across all of these paths unless they all have exactly the same cost. Currently, the IS-IS weights are handcrafted by network operations experts. Moreover altering IS-IS weights has potential repercussions on the entire backbone. We therefore search for a more policy-based approach towards load balancing.

Clearly, in order to use some of the underutilized links and paths, a load balancing scheme would have to deviate from using shortest hop paths. It is important to ensure that significant delays are not introduced to traffic that is rerouted on longer paths. We believe that this will not happen for two reasons. First, the backbone is highly meshed, and thus most alternate paths between an ingress-egress POP pair are likely to be only one or two hops longer than the min-hop path. Second, [1] shows that the average delay across routers in the backbone is on the order of a few milliseconds. Therefore, the additional delay that a packet will incur by traversing a few more routers is likely to be within acceptable limits.

VI. Traffic Aggregates for Load Balancing

In order to realize effective load balancing in the backbone, it is necessary to understand how traffic should be split over multiple alternate paths. In this section, we address this issue by examining techniques for creating aggregate traffic streams between (ingress link, egress POP) pairs. The aggregation of packets into streams can be based on a variety of criteria and can lead to streams with different levels of granularity. At the coarsest level, we can aggregate all the packets into a single stream. On the other hand, using the classic five-tuple of (source address, destination address, source port, destination port, protocol) leads to very fine-grained streams. The criteria used for creating traffic aggregates depends largely on the purpose of such aggregation. For example, when the goal is to provide different levels of service to different types of traffic, packets may be aggregated based on the (TOS) field or the protocol field in the packet IP header. Since we are interested in the routing of these aggregate streams across the backbone, it is natural to consider the destination address of packets as the basis for aggregation. Moreover routes are determined according the destination subnets (as advertised through BGP), each of which is an aggregate over

a range of IP addrreses. Subnets in turn can be grouped on the basis of IP address prefixes. Therefore we consider destination address prefixes of different lengths as the basis for aggregating POP-to-POP traffic. For example, streams may be created based on an 8-bit destination address prefix, in which case all packets sharing the same first octet value for their IP address belong to one stream. We shall henceforth refer to such a stream as a $p8$ stream. In general, when an N-bit prefix is used for aggregation, we refer to the aggregate stream as a pN stream.

Aggregate traffic streams thus created would be assigned to different paths in order to balance the network load. Before adopting this approach to load balancing, we need to examine properties of these aggregates such as their traffic volume and their stability over the time interval for which such load balancing would be carried out.

We first consider $p8$ streams and rank them in decreasing order of traffic volume (so that stream #1 is the largest). Figure 15 shows the cumulative percentage of traffic of $p8$ streams from the private peer access link and the webhost access link 1, respectively. For this access link, the traffic demand to three of the busiest egress POPs is presented. We see that for every egress POP pair, a few of the top-ranked flows account for an overwhelmingly large share of traffic. We have observed that this phenomenon is widespread across most other (ingress POP access link, egress POP) pairs. This brings us to an important result - the existence of a few very high-volume traffic streams, and many low-volume traffic streams in the backbone. We refer to the former as *elephants* and to the latter as *mice*. As mentioned in Section I, the phenomenon of "elephants and mice" has been reported at other granularity levels in other traffic studies [4], [3], [8]. Here we demonstrate the existence of elephants and mice at specific IP destination address prefix levels in a commercial IP backbone.

The existence of elephants has important implications for traffic engineering in general, namely that in order to realize most of the benefits, we can focus primarily on engineering the network for the elephants. Many of the difficulties in providing quality of service in the Internet today stem from scalability issues. One cannot exert fine grained control because of scalability problems that arise with keeping too much state information. The elephants and mice phenomem means that one can try to exert more careful control on the elephants and that coarse control is sufficient for the mice. Although this has been observed before, we are not aware of any concrete suggestions or examples of using this traffic behavior to influence control. Elephants streams provide a basis for load balancing since once the elephants are identified, they can be rerouted over underutilized portions of the network. The criterion for identifying the elephants – destination address prefix – is simple enough for use in practice without new and complex protocols.

For simplicity of implementation, it is attractive to have a load balancing policy that is applicable over long time scales, such as a few hours, or even potentially throughout the day-time. Of course, our approach of load balancing via rerouting elephants, cannot be applied unless the ranking of elephants and mice remains fairly stable on these timescales. Figure 16 show the time-of-day variation of bandwidth for some of the elephants and mice to a busy POP from webhost 1 access link. In the graph, the one-hour average of the bandwidths of these streams is plotted against time for 18 hours. We find that throughout this period, the elephants retain a large share of the bandwidth, and that they maintain their relative ordering. In other words, the elephants remain elephants and the mice remain mice. We have verified this behaviour for a large number of ingress-egress POP pairs. This result encourages us to focus our attention on just a few streams in the backbone for the purposes of load balancing.

Interestingly, we discover that the phenomenon of elephants and mice is recursive. In other words, if we consider a $p8$ elephant stream, and then further subdivide it into sub-streams based on say a 16 bit prefix, then we find elephants and mice again among these substreams. In Figure17 we consider the three largest elephants to each of the POPs 4 and 12 for the peer 1 access link, subdivide each into $p16$ streams, rank them, and plot the cumulative volume for the ordered streams. Thus each curve in Figure 17 corresponds to the $p16$ substreams from a single $p8$ stream for a given POP. We find that 10 of the largest flows account for 80% or more of the bandwidth in every case. As with the $p8$ streams, these $p16$ elephants and mice exhibit stable behavior over many hours (figures omitted due to space considerations), even though the bandwidth of some of the elephants decreases substantially at night.

We further examine this recursive behaviour by taking some of the $p16$ streams from the previous step and dividing them into substreams based on a 24-bit prefix. We find that although the elephants and mice phenomenon

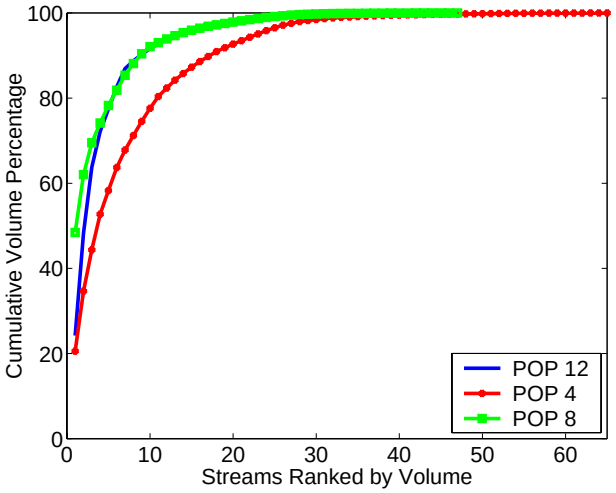

Fig. 15. Distribution of traffic across p8 streams for Webhost access link 1

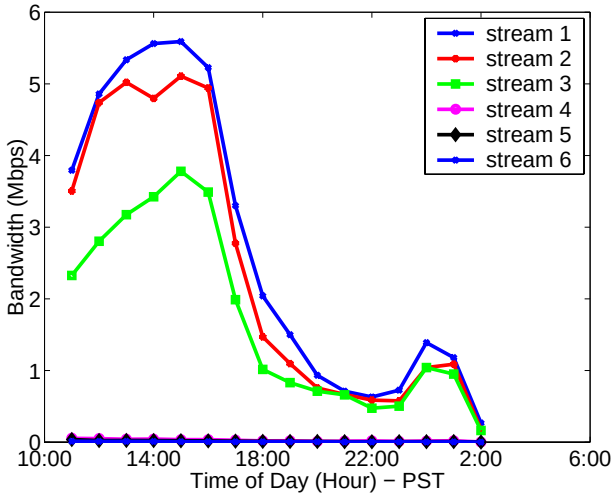

Fig. 16. Time of day variations for p8 elephants and mice for Webhost access link 1

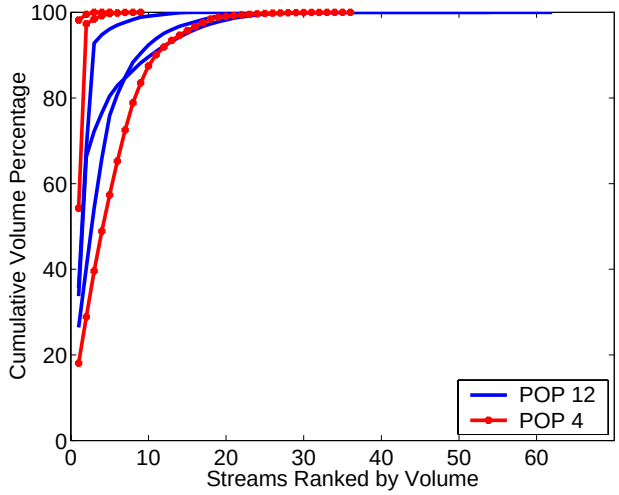

Fig. 17. Distribution of traffic across p16 streams for peer 1 access link

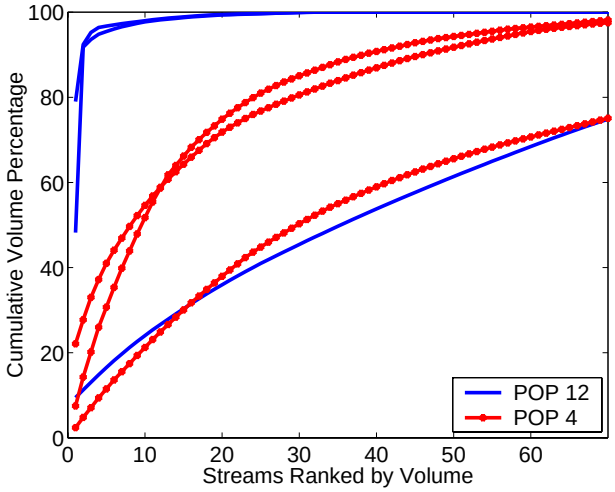

Fig. 18. Distribution of traffic across p24 streams for peer 1 access link

still exists, it becomes less pronounced as traffic becomes somewhat more uniformly distributed across streams (Figure 18). Although we investigate 1, 2 and 3 byte masks, there is no particular association with Class A, B and C addresses that have become less meaningful with the advent of CIDR. In fact, we expect that this phenomenon will manifest itself at other prefix levels as well, certainly those between 8-24, but probably less so at prefixes longer than 24.

A different way of studying the stability of elephants and mice is to look at the frequency and size of rank changes at a given prefix level. Suppose that we divide time into equal-sized slots and compute the average bandwidth for all the streams in each time slot. We can then rank the streams according to bandwidths and examine the change in rank of streams from one time slot to another. More precisely, let $R_i(n)$ be the rank of flow i in time slot n, where $n = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, M$. Let us define $\delta(i, n, k) = |R_i(n) - R_i(n + k)|$, where $1 \leq k \leq (N - n)$. For a given value of k, we examine the probability distribution for $\delta(\cdot, \cdot, k)$.

Figure 19 applies this technique for $p8$ traffic streams

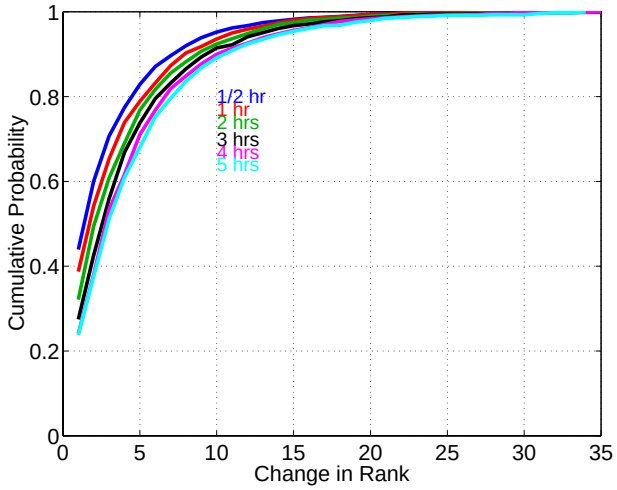

Fig. 19. Cumulative distribution of rank changes for p8 traffic streams between peer 1 access link and a busy egress POP

between the peer 1 access link and POP 12 for an averaging interval of 30 minutes, and $k = 1, 2, 4, 6, 8$ and 10. The results show that most of the rank changes are small – in fact, rank changes of size 5 or less account for about 70% of the changes. Moreover, this is true for rank changes over different time intervals ranging from 30 minutes to five hours. Besides the obvious implication for load balancing, this result has a powerful implication for building traffic matrices. It provides a direction for predicting the rank of a stream at a future time based on the current rank. Development of such prediction techniques requires extensive analysis and sophisticated statistical models, and is beyond the scope of the current paper.

This distribution contains all the elephants and mice mixed together. To isolate the rank change behavior of the elephants alone, we checked the ranking of particular elephants over entire traces. We found that 70% of the top 15 elephants remain among the top 15 streams throughout the day. We verified this behavior for a number of p8 streams and a number of egress POPs.

In summary, we demonstrate the existence of the elephants and mice phenomenon at the granularity of $p8$, $p16$ and $p24$ prefixes. We show that this phenomenon is recursive in that elephants themselves are composed of elephants and mice at a lower granularity level. We verify the stability of the ranking of streams a few ways: by examing the time-of-day behavior, by calculating the density on the change of rank process over multiple time intervals, and by examining the top elephants and their ranking throughout the day. Our results indicate there exist natural and simple aggregates, based on p8 and p16 aggregates, that constitute a large fraction of traffic and that remain stable throughout the day. Many of today's routers provide the capability to do per-prefix load balancing over equal/unequal cost links (e.g., Cisco's Express Forwarding). Load balancing at the traffic granularity that we identify here can be implemented by extending this capability.

VII. Related Work

Starting with pioneering work by Paxson ([11], [12]), network measurement and monitoring has received widespread attention from Internet researchers and practitioners ([13], [3], [14], [4]). However, much of this work relies on data collected at public peering points, edge routers and from academic networks. Our work is unique in that the analysis is based on traces and routing tables collected from a operational IP backbone. In this respect, our paper comes closest to the work in ([2], [3]). In [2] Grossglauser et. al. propose a method for following the trajectories of a subset of the packets flowing through a backbone network. The method is based on using a single hash function, based on packet content, at multiple nodes in the network. It has direct applicability to the problem of determining traffic demands on different paths through

a network.

Internet measurement data can be broadly divided into routing data and traffic data. The former consists of data about routing protocols and their behaviour. Enormous amounts of such data has been collected and analysed to understand the behaviour of large-scale routing in the Internet ([9], [13], [11], [14]). Traffic data, consisting of packet traces, is not as widely available, especially from operational backbones. However, both traffic and routing data are need to construct traffic matrices. The use of traffic matrices as a systematic way of representing and analyzing Internet traffic has been gaining attention recently ([3], [15], [7]), and [3] is an important recent work in this area. There are strong similarities in the overall goal of the work in [3] and our work – collecting and processing data from an operational backbone in order to understand traffic demands and improve traffic engineering. However our work differs from [3] in a number of ways. First, [3] uses data from peering links at different points in the backbone to construct point-to-multipoint traffic demands across these peering links. These traffic demands comprise only of the transit traffic through their backbone. On the other hand, we collect data from a diverse set of access links (peering, web-hosting, ISPs, etc.) in our backbone, and study the geographic spread of this traffic over the entire backbone. As we showed in this paper, the spatial and temporal characteristics of traffic depends on the type of originating access link; this makes it important to study traffic from different types of access links. Secondly, [3] seeks to build a traffic matrix representing multipoint demands from one router to a set of egress router nodes; this captures all the alternate egress points to a destination network beyond the backbone. In our backbone, I-BGP policies are used to pick one of many egress points to a destination network at any given time. We are interested in studying internal routing and traffic behaviour, given that this egress point has already been determined by I-BGP. Hence we focus only on point-to-point POP-level flows. These differences notwithstanding, both our work and [3] represent important first efforts in understanding backbone traffic demands.

VIII. Conclusions

In this paper, we used packet-level traces collected at a large POP of a tier-1 IP backbone to understand the dynamics of ingress traffic at a backbone POP. In order to study geographical and temporal properties of POP-to-POP traffic, we introduced a methodology for combining our data with BGP information to classify the incoming traffic according to egress POPs. We found that there is a wide disparity in the volume of traffic headed towards different egress POPs. We analyzed the traffic at three granularity levels, the POP level, the access link level, and the destination prefix level. A contribution was to demonstrate different types of temporal stability for each of these on long time scales.

We also examined our network to see how the traffic demands are routed through the backbone. We found that the POP topology and IS-IS link weights are carefully chosen to constrain traffic between most ingress-egress POP pairs to a few paths across the backbone. The combined effect of such routing practices and overprovisioning means that there is a lot of excess capacity in the core that results in (i) underutilized links, (ii) a wide range of link levels within the underutilized range, and (iii) some links being consistenly underutilized. Our findings on the disparate geographic spread of traffic demands combined with current routing practices indicate that there is a lot of room for improvement in load balancing in the backbone. Current routing practices today do not take into consideration the traffic demand matrix because such matrices are typically not available. We believe this is one of the key reasons why we see large variation in link load levels.

Our main findings can be summarized as follows.

- The geographic spread of traffic from one ingress POP across all egress POPs is far from uniform. A few egress POPs sink large amounts of traffic, while the majority sink either small or medium amounts of traffic. Our initial assessment of POPs indicates that a simple categorization, in which each category draws about twice the volume of traffic as the one below it (i.e. large/medium and medium/small ratios are approximately two), is possible. Further work needs to be done to model POPs in finer detail. This data is important in that it confirms empirically our intuition (based on internet practices) about how traffic is distributed across backbones. However, it also contradicts the widely used simulation model that assumes uniform distribution of traffic among destination nodes.
- Access links do not distribute traffic similarly across egress POPs; some access links are more likely to send to one set of POPs, while others are more likely to send to a different set of POPs. This differentiation occurs mostly

in medium sized egress POPs, and not in large or small POPs.

- We found that the large egress POPs can exhibit a large variability during the day, whereas the medium and small POPs exhibit little variability over a full day. More importantly, we found that when POPs are ranked according to volume, then they maintain their rank throughout the day. With respect to their rank, POPs appear quite stable.
- The elephants and mice phenomenon that we found among streams aggregated on destination prefixes is a natural basis for splitting traffic over multiple paths in the backbone, using routing policies. Identifying reroutable flows at this level of traffic granularity is attractive because such flows exhibit stable behavior throughout the day. Load balancing this way would require early identification of the elephants in the access links of the ingress POPs.

The value of our methodology, observations and analysis extends beyond load balancing to other aspects of backbone engineering. For example, we found a close connection between traffic patterns amongst POPs and the architecture of the POPs themselves. This can help in architecting POPs, or in adding new customers and provisioning backbone capacity when the backbone is upgraded. Also, our analysis of POP behavior, its spatial and temporal characteristics, and its underlying dependence upon access links can be incorporated into capacity planning models.

REFERENCES

[1] C. Fraleigh, C. Diot, B. Lyles, S. Moon, P. Owezarski, D. Papagianaki, and F. Tobagi, "Design and Deployment of a Passive Monitoring Infrastructure," *Passive and Active Measurement Workshop, Amsterdam, Netherlands*, April 2001.

[2] N. Duffield and M. Grossglauser, "Trajectory Sampling for Direct Traffic Observation," *ACM SIGCOMM*, 2000.

[3] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving Traffic Demands for Operational IP Networks: Methodology and Experience," *ACM SIGCOMM*, August 2000.

[4] W. Fang and L. Peterson, "Inter-AS Traffic Patterns and Their Implications," *Proceedings of Global Internet*, December 1999.

[5] V. Paxson and S. Floyd, "Why We Don't Know How to Simulate the Internet," *Proceedings of the 1997 Winter Simulation Conference*, December 1997.

[6] D. O. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "A Framework for Internet Traffic Engineering," *Internet Draft draft-ietf-tewg-framework-02.txt*, May 2000.

[7] R. Sabatino, "Traffic Accounting using Netflow and Cflowd," *Fourth International Symposium on Interworking, Ottawa, Canada*, July 1998.

[8] L. Kleinrock and W. Naylor, "On Measured Behavior of the Arpa Network," *AFIPS Conference Proceedings, National Computer Conference*, vol. 43, December 1999.

[9] G. Huston, "Tracking the Internet's BGP Table," *ISMA Winter Workshop, San Diego, USA*, December 2000.

[10] S. Nilsson and G. Karlsson, "IP-address lookup using LC-tries," *IEEE Journal on Selected Areas in Communication, 17(6):1083-1092*, June 1999.

[11] V. Paxson, "End-to-End Routing Behavior in the Internet," *IEEE/ACM Transactions on Networking, vol. 5, pp. 601-615*, October 1997.

[12] V. Paxson, "End-to-End Internet Packet Dynamics," *ACM SIGCOMM, Cannes, France*, September 1997.

[13] K. Claffy, "Internet Measurement and Data Analysis : Topology, Workload, Performance and Routing Statistics," *NAE Workshop*, 1999.

[14] G. R. Malan C. Labovitz and F. Jahanian, "Internet Routing Instability," *ACM SIGCOMM, Canne, France*, September 1997.

[15] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: Traffic Engineering for IP Networks," *IEEE Network Magazine*, March 2000.

[16] O. Goldschmidt, "ISP Backbone Traffic Inference Methods to Support Traffic Engineering," *2nd ISMA Winter Workshop, San Diego, CA.*, December 2000.

[17] B. Fortz and M. Thorup, "Internet Traffic Engineering by Optimizing OSPF Weights," *IEEE Infocom*, March 2000.

[18] S. V. Wiel, J. Cao, D. Davis, and B. Yu, " Time-varying Network Tomography: Router Link Data," *Symposium on the Interface: Computing Science and Statistics*, June 1999.

[19] B. Chinoy, "Dynamics of Internet Routing Information," *ACM SIGCOMM*, 1993.

[20] A. Shaikh, J. Rexford, and K. Shin, "Load-Sensitive Routing of Long-Lived IP Flows," *ACM SIGCOMM*, pp. 215–226, September 1999.

[21] A. Sridharan, S. Bhattacharyya, C. Diot, R. Guerin, J. Jetcheva, and N. Taft, "On the Impact of Aggregation on the Performance of Traffic Aware Routing"," *Proceedings of the 17th International Teletraffic Congress (ITC), Salvador do Bahia, Brazil*, September 2001.