

# Understanding Traffic Dynamics at a Backbone POP

Nina Taft<sup>a</sup> and Supratik Bhattacharyya<sup>a</sup> and Jorjeta Jetcheva<sup>b</sup> and Christophe Diot<sup>a</sup>

<sup>a</sup>Sprint Advanced Technologies Laboratories, Burlingame CA, USA.

<sup>b</sup>Carnegie Mellon University, Pittsburg PA, USA

## ABSTRACT

Spatial and temporal information about traffic dynamics is central to the design of effective traffic engineering practices for IP backbones. In this paper we study backbone traffic dynamics using data collected at a major POP on a tier-1 IP backbone. 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 results show that there is wide disparity in the volume of traffic headed towards different egress POPs. At the same time, we find that current routing practices in the backbone tend to constrain traffic between ingress-egress POP pairs to a small number of paths. As a result, there is a wide variation in the utilization level of links in the backbone. Frequent capacity upgrades of the heavily used links are expensive; the need for such upgrades can be reduced by designing load balancing policies that will route more traffic over less utilized links. 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 also demonstrate that these aggregates exhibit stability throughout the day on per-hour time scales, and thus they form a natural basis for splitting traffic over multiple paths in order to improve load balancing.

**Keywords:** traffic engineering, load balancing, network measurements, BGP, traffic matrix

## 1. INTRODUCTION

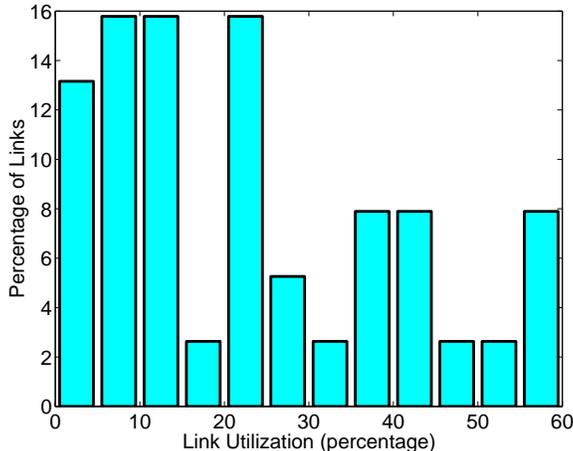
Internet backbones continue to grow at explosive rates, fueled by the bandwidth demands of new applications and by the advent of faster access technologies. To accommodate such growth while preserving the robustness of the network, most IP backbone operators have chosen a simple approach to traffic engineering: overprovisioning. Together with a highly meshed backbone and well understood routing protocols (such as OSPF and IS-IS), overprovisioning succeeds in providing lossless and low delay transmission. Overprovisioning means that there is a lot of excess capacity in the core that results in underutilized links, some of which are consistently underutilized. Whenever backbone links start to regularly bypass specified load levels, those links are upgraded. Because upgrades are done on a very coarse grained level (OC-12 to OC-48, OC-48 to OC-192, etc.), upgrades frequently perpetuate the existence of unused resources because the demand curve grows in a smoother fashion. To reduce the cost of this inefficient approach, it is advantageous to postpone the frequency of such upgrades, by reducing the likelihood that links reach the threshold load levels defined for upgrades.

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 to collect and process 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.<sup>1</sup> The data is then analyzed offline in order to understand the dynamics of traffic entering the backbone at this POP.

In order to get an understanding of how the excess capacity due to overprovisioning is distributed throughout our backbone, we examined the average link utilisation using SNMP data. Figure 1 provides a histogram of this data. 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%. Clearly there are some links that are consistently underutilized and the range of utilizations levels from smallest to largest is very wide. To understand why this happens one needs to know what the traffic demands are between pairs of POPs in the network and what paths the traffic is routed on. If it the traffic

---

Corresponding authors : Supratik Bhattacharyya, Nina Taft. The third author was at Sprint Advanced Technology Laboratories when this work was done.



**Figure 1.** Histogram of Link Utilizations in the Backbone

demands are amenable to rerouting, then the network would benefit by moving some of the traffic onto the lightly loaded links.

Before discussing the traffic demands, we first make some comments about the paths that traffic is routed on, based upon our observations of IS-IS routing practices. IS-IS is the Interior Gateway Protocol used in our backbone. IS-IS is carefully engineered to influence path selection by a judicious choice of links weights that effectively constrain traffic between an ingress-egress POP to only a few (often overlapping) paths. It is difficult to achieve satisfactory load balancing, since IS-IS routes all traffic on the minimum cost path, and does not take into account traffic demands. The choice of paths can be heavily influenced by the POP design itself because POPs contain many nodes and links that impact the cost (e.g., when measured in terms of hop count). Because our topology is highly meshed, there are many alternate paths that do not get selected.

Given our observations on highly disparate link utilizations and the restrictions on routing, we were motivated to explore the possibilities of improving load balancing practices. To understand whether or not the traffic is amenable to alternate load balancing practices, we study both geographic and temporal properties of the traffic demands. We discovered a large disparity in the spatial distribution of the monitored POP’s traffic across the egress POPs. The time of day behavior of the traffic at the POP-level revealed that POPs can be ranked roughly into three categories (large, medium and small), and that POPs remain stable throughout the day in the sense that they don’t change categories. A stronger statement can be made about many of the POPs: that if they are ranked by the volume of traffic they receive, they maintain their ranking throughout the day. To address the issue of whether load balancing policies can remain unchanged throughout an entire 24 hour period, we compared day and night behavior of our traffic. We found that at night the overall traffic load is reduced by only 15-50%.

We conclude that the wide disparity in the spatial distribution of traffic and current routing practices together explain the large differences observed in backbone link utilizations. Better load balancing policies would help to avoid having to upgrade the more heavily used links while plenty of underutilized links coexist by reducing the frequency at which such expensive upgrades occur.

In order to assess at which level of traffic granularity load balancing should be applied, we examine traffic aggregates based on destination address prefixes. We find that a small number of these aggregate streams, which we call *elephants*, generate a large fraction of the total traffic, while a large number of these streams, which we call *mice*, generate a small fraction of the total traffic. Traffic with this property is also often referred to as non-uniform. The elephants and mice phenomenon has been observed before in Internet traffic at the inter-AS level,<sup>2</sup> at the level of multipoint demands from one router node to a set of router nodes<sup>3</sup> and in the Internet as it was many years ago.<sup>4</sup> Here we demonstrate this phenomenon at the granularity level of specific prefixes. The elephants and mice phenomenon is useful for load balancing because load balancing only need be applied to a few flows yet it can affect a sizeable amount of traffic. This is a scalable approach to load balancing. To ensure that such an approach would remain valid over long time periods, we examine the stability of these aggregates throughout the day. We find that

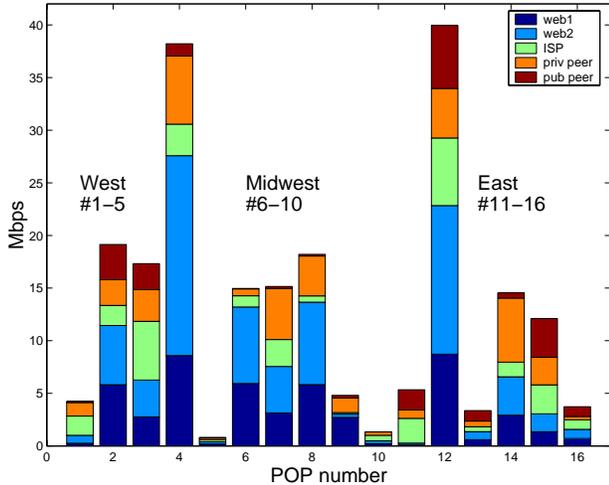


Figure 2. Fanout of Ingress POP traffic

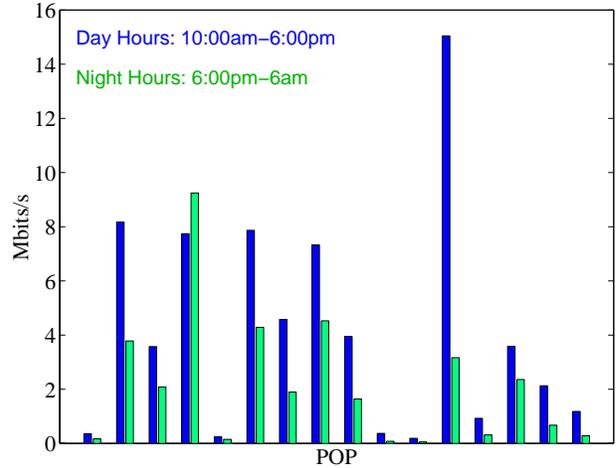


Figure 3. Day and Night Traffic Histogram

the elephants remain elephants throughout the day which makes them well-suited as a basis for routing traffic on alternate paths.

The data used for this study was gathered from an operational IP backbone using our passive monitoring infrastructure.<sup>1</sup> 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 web servers. Peering at a POP is provided either through dedicated links to another backbone (private peering) or through public Network Access Points (NAPs).

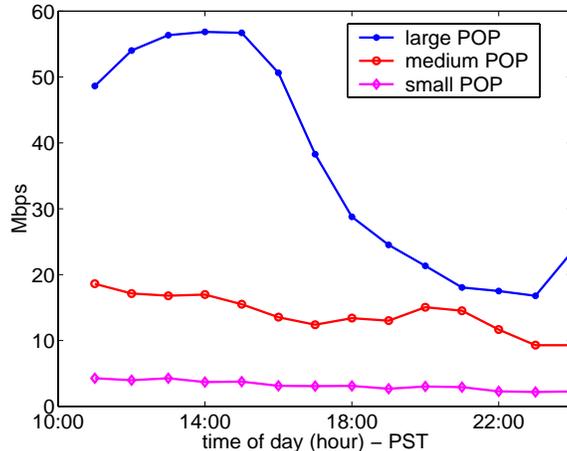
The infrastructure developed to monitor this network consists of passive monitoring systems that collect packet traces and routing information on the links located between the access routers and the backbone routers, or on the peering links. The monitoring systems tap onto the selected link using optical 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. BGP tables were downloaded from a router in the POP once per hour during the time the packet traces were collected. The methodology developed to combine the packet traces with BGP information in order to determine traffic volume between ingress-egress POP pairs is described in a detailed technical report.<sup>5</sup>

## 2. RESULTS

In this section, we present some of our findings on the geographical distribution and temporal properties of traffic between ingress-egress POP pairs. Our results are based on day-long packet traces collected on five representative access links at one of our backbone POPs. The links include two peering links, two web hosting access links and a Tier-2 access link. These five links constitute a significant portion of the input traffic at our monitored POP. Given the variety of access links chosen, this data is also highly representative of the total input traffic entering the POP. The monitored POP is henceforth referred to as the *ingress* POP. For the complete set of results, the reader is referred to the detailed technical report.<sup>5</sup>

We look at the geographical spread of traffic demands across egress POPs, or *fanout*, is depicted in Figure 2. Each column represents the average amount of bandwidth that our ingress POP sent to a particular egress POP. The bandwidth values that were averaged over the duration of the entire trace for each access link. 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 significance.

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



**Figure 4.** Time of Day Behavior of POPs

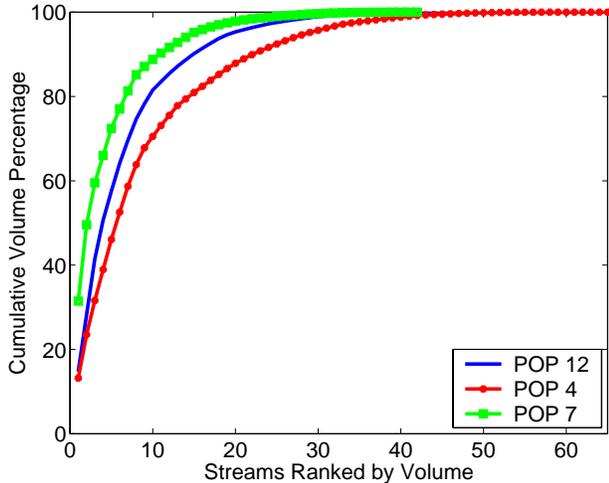
half receive a moderate amount of traffic (10-20 Mbps). Our data suggests that ingress POPs could be roughly categorized as *large*, *medium* and *small*, where (i) roughly the same number of POPs fall into the small and medium categories and only a few fall into the large category; and (ii) each category carries approximately twice the volume of the category below it.

In many ways this histogram matches our intuition. First, one would expect that some POPs would generate 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.

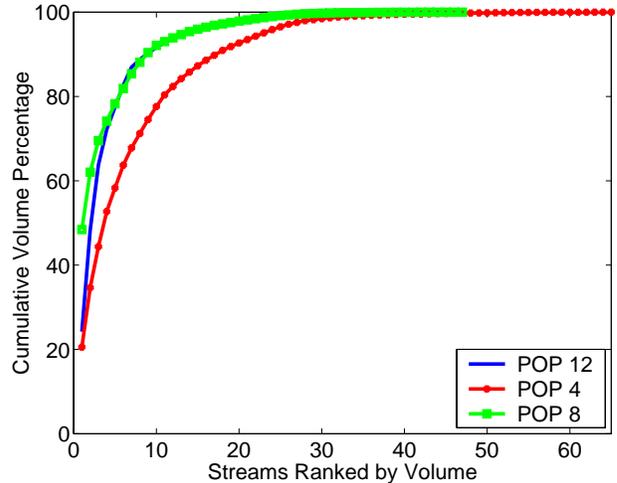
In order to study the stability of this fanout over time, we next look at daytime and nighttime behaviour of POP-to-POP traffic. For this purpose, we separate out our previous fanout plot into two fanouts, one representing the day hours and the other representing night hours (Figure 3). When we compare the total amount of day time traffic against the total amount of night traffic across all egress POPs, we find that the traffic at night is about 50% less than during the day, i.e., a reduction factor of about two. This figure is from a single access link. The range of traffic reduction across all five monitored links was between 15-50%. The amount of nighttime reduction appears to be link dependent.

To examine the temporal stability of POP traffic at a finer granularity level, we plot the traffic from three representative POPs (one from each category) on per-hour time scales (Figure 4). We find here that POPs remain in their assigned category. This also holds true for the remaining POPs not included in this figure. More specifically, a similar plot including all the POPs (not included here due to the difficulty of including graphs with 16 curves in a small space) demonstrates that POPs, ranked by volume, retain their rank throughout the day.

In summary, we derive the following insights from our results. We find that the traffic fan-out is highly heterogeneous in space yet stable in time. We thus conclude that the high concentration of traffic between certain POP pairs, coupled with current routing practices that concentrate the traffic on a few paths, gives rise to heterogeneity in link utilization levels. This high concentration of traffic between certain POP pairs, indicates that it might be sufficient to reroute traffic from just a few POP pairs and still make significant improvements in rendering network-wide link utilization levels more similar. Note that the goal should not be to *equalize* link utilizations levels which is clearly impossible. However reducing the disparity would improve the efficiency with which network resources are currently being used.



**Figure 5.** Distribution of traffic across  $p8$  streams for a peer access link



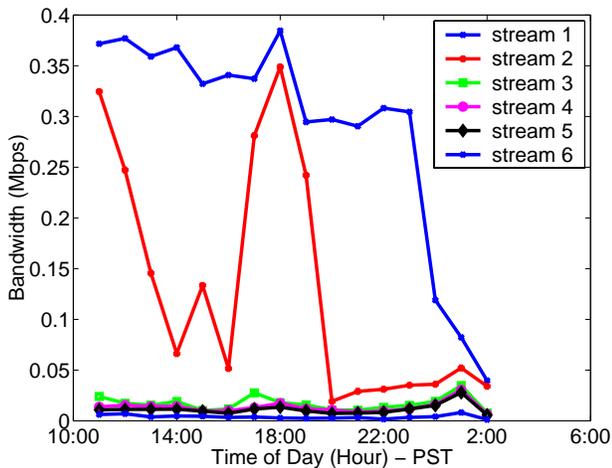
**Figure 6.** Distribution of traffic across  $p8$  streams for a Webhost access link

### 3. OF ELEPHANTS AND MICE

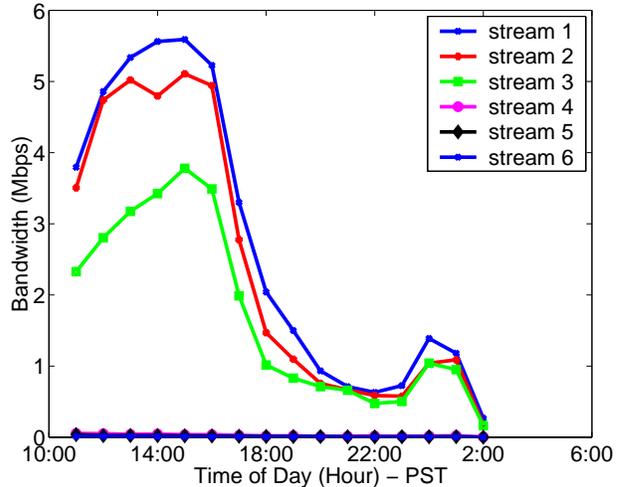
In order to realize effective load balancing in the backbone, it is necessary to understand how traffic should be split over multiple alternate paths. In this section, we address this issue by examining techniques for creating aggregate traffic streams between (ingress link, egress POP) pairs. The aggregation of packets into streams can be based on a variety of criteria and can lead to streams with different levels of granularity. The criteria used for creating traffic aggregates depends largely on the purpose of such aggregation. 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 addresses. 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.

We first consider  $p8$  streams and rank them in decreasing order of traffic volume (so that stream #1 is the largest). Figures 5 and 6 show the cumulative percentage of traffic of  $p8$  streams from a peer access link and a web-host access link, respectively. For each access link traffic, the data from three of the busiest egress POPs is presented. We see that for every (ingress POP access link, 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 1, the phenomenon of “elephants and mice” has been reported at other granularity levels in other traffic studies.<sup>2-4</sup> Here we demonstrate the existence of elephants and mice at specific IP destination address prefix levels.

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. This implies that one need not keep state information for all flows, but only a few. 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,



**Figure 7.** Time of day variations for p8 elephants and mice for a peer access link



**Figure 8.** Time of day variations for p8 elephants and mice for a Webhost access link

they can be rerouted over underutilized portions of the network.

We observed in section 2 that there is a 15 – 50% reduction in traffic on various links from daytime to nighttime. Therefore there is more to be gained by load balancing during the day than at night. Given the wide differences in traffic volume reduction on different links, it is difficult to predict whether any benefit (or how much) can be accrued from load balancing at night. The same load balancing policy may be applicable both during the day and night, or a separate policy may be beneficial for each. In any case, our work points to the usefulness of load balancing policies that are applicable on the timescale of multiple hours. This is an important observations since frequent policy changes would be infeasible in an operational network.

Of course, such load balancing policies spanning multiple hours cannot be applied unless the ranking of elephants and mice remains fairly stable on these timescales. Figure 7 and 8 show the time-of-day variation of bandwidth for some of the elephants and mice to a busy POP from the one of the access links at the monitored POP. In the graphs, 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.

#### 4. CONCLUSION

In this paper, we used packet-level traces collected at a large POP of a tier-1 IP backbone to understand the dynamics of traffic entering the backbone at that POP. By studying traffic demands between POPs, traffic dynamics at a POP, and correlating this with observations about routing practices, we are able to explain the wide disparity in backbone link utilization. Furthermore we used our data and its properties to gain insights into desirable approaches to load balancing. We summarize our main observations below :

- A few egress POPs account for a large portion of the traffic entering the backbone at all times. Load balancing policies can be applied at the ingress POP to spread the traffic headed towards these dominant POPs over different backbone paths.
- The traffic headed towards the dominant egress POPs can be highly variable – there is a upto a 50% difference in peak traffic between day and night. Hence different load balancing policies may be used during day and night, e.g., load balancing could be turned on during the day, and turned off at night.

- 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. This requires 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, e.g., improving the architecture of POPs, finding suitable locations for adding new customers to the backbone, provisioning capacity, etc. Also, our analysis of the spatial and temporal characteristics of backbone traffic, and its underlying dependence upon the physical characteristics of the underlying network (e.g., link capacities) can be incorporated as heuristics in bandwidth prediction models. Our work thus constitutes an important first step towards building models for populating traffic matrices at the POP-to-POP level. Development of more sophisticated models requires extensive analysis of data spanning days, or even months, and is a part of ongoing work.

## REFERENCES

1. C. Fraleigh, S. Moon, C. Diot, B. Lyles, and F. Tobagi, "Architecture of a Passive Monitoring System for Backbone IP Networks," *Sprint Technical Report TR00-ATL-101801*, October 2000.
2. W. Fang and L. Peterson, "Inter-AS Traffic Patterns and Their Implications," *Proceedings of Global Internet*, December 1999.
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. L. Kleinrock and W. Naylor, "On Measured Behavior of the Arpa Network," *AFIPS Conference Proceedings, National Computer Conference* **43**, December 1999.
5. N. Taft, S. Bhattacharyya, J. Jetcheva, and C. Diot, "Understanding Traffic Dynamics at a Backbone POP," *Sprint Technical Report TR01-ATL-020201*, February 2001.
6. O. Goldschmidt, "ISP Backbone Traffic Inference Methods to Support Traffic Engineering," *2nd ISMA Winter Workshop, San Diego, CA.*, December 2000.
7. B. Fortz and M. Thorup, "Internet Traffic Engineering by Optimizing OSPF Weights," *IEEE Infocom*, March 2000.
8. A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: Traffic Engineering for IP Networks," *IEEE Network Magazine*, March 2000.
9. N. Duffield and M. Grossglauser, "Trajectory Sampling for Direct Traffic Observation," *ACM SIGCOMM*, 2000.
10. 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.
11. 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.
12. V. Paxson and S. Floyd, "Why We Don't Know How to Simulate the Internet," *Proceedings of the 1997 Winter Simulation Conference*, December 1997.
13. R. Sabatino, "Traffic Accounting using Netflow and Cflowd," *Fourth International Symposium on Interworking, Ottawa, Canada*, July 1998.
14. V. Paxson, "End-to-End Routing Behavior in the Internet," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 601-615, October 1997.
15. V. Paxson, "End-to-End Internet Packet Dynamics," *ACM SIGCOMM, Cannes, France*, September 1997.
16. K. Claffy, "Internet Measurement and Data Analysis : Topology, Workload, Performance and Routing Statistics," *NAE Workshop*, 1999.
17. S. Nilsson and G. Karlsson, "IP-address lookup using LC-tries," *IEEE Journal on Selected Areas in Communication*, 17(6):1083-1092, June 1999.
18. C. Labovitz, G. R. Malan, and F. Jahanian, "Internet Routing Instability," *ACM SIGCOMM, Canne, France*, September 1997.
19. B. Chinoy, "Dynamics of Internet Routing Information," *ACM SIGCOMM*, 1993.
20. G. Huston, "Tracking the Internet's BGP Table," *ISMA Winter Workshop, San Diego, USA*, December 2000.
21. A. Shaikh, J. Rexford, and K. Shin, "Load-Sensitive ROuting of Long-Lived IP Flows," *ACM SIGCOMM*, pp. 215-226, September 1999.