On the Feasibility of Identifying Elephants in Internet Backbone Traffic

K. Papagiannaki, N. Taft, S. Bhattacharya, P. Thiran, K. Salamatian, C. Diot

***Abstract*—The elephants and mice phenomenon is known as one of the few invariants of Internet traffic. No systematic method has been proposed so far for defining a boundary to separate flows into two distinct classes. We consider four single-feature classification methods and examine network-level prefix traces collected from a Tier-1 backbone. We illustrate that one cannot simultaneously achieve consistency in the fraction of total load due to elephants and the number of elephants. These methods also result in short average holding times in the elephant state. To avoid reclassification due to short term fluctuations we propose a two feature classification method that incorporates a metric, called latent heat, that is a function of the difference in elephant bandwidth and the class separation threshold over a number of time intervals. Our first contribution is to illustrate the volatility of elephants and hence the difficulty in isolating them for traffic engineering purposes, especially if one wants to achieve consistency with respect to load, number and duration of elephant flows. Our second contribution is to propose a classification method that achieves a reasonable tradeoff between fluctuations in elephant load and fluctuations in the number of elephants while simultaneously yielding longer average holding times.**

I. Introduction

Traffic engineering of IP networks has attracted a lot of attention in the recent past. Network operators are seeking ways in which they could make optimal use of their networks' capacity, without introducing complexity and maintaining the simplicity of the packet switching paradigm. With that in mind, researchers have proposed different algorithms for traffic engineering tasks that could lead to better utilized networks by exploiting certain properties of the underlying traffic [2], [6], [10], [12].

Studies of the Internet traffic at the level of network prefixes, fixed length prefixes, TCP flows, AS's, and WWW traffic, have all shown that a very small percentage of the flows carries the largest part of the information [2], [6], [10], [12]. This behavior is known as "the elephants and mice phenomenon", and is considered to be one of the few invariants of Internet traffic [7]. In statistics, this phenomenon is also called the "mass-count disparity". The "mass-count disparity" states that for heavy-tailed distributions the majority of the mass in a set of observations is concentrated in a very small subset of the observations [3].

In order to be able to exploit such a property for traffic engineering purposes, one has to be able to identify which flows carry the majority of the bytes. Examples of traffic engineering applications include re-routing [10] and load balancing of elephant flows. The attraction of this approach to traffic engineering is that by treating a relatively small number of flows differently one can affect a large portion of the overall traffic. For this to be practical, elephant flows would need to remain elephant flows for reasonably long periods of time, so that a traffic engineering application need not change its policy or state frequently.

To the best of our knowledge no systematic way has been proposed so far, for choosing the criterion that isolates the high-volume flows. The criterion should allow one to consider any particular flow and to decide if it should be categorized as an elephant or a mouse. Other studies have selected a particular criterion and then examined other problems given this fixed criterion. For example, in [5] an elephant is any flow whose rate is larger than 1% of the link utilization; and in [9] an elephant is any flow whose peak rate exceeds the mean plus three standard deviations of the aggregate link flow. These studies do not address the question of how to choose the criterion for defining an elephant. Instead they focus on, given a fixed criterion, how should one sample high-volume flows [5], and on how to mathematically model the two types of flows [9].

Our aim is to separate out a small number of high-volume flows that account for most of the traffic, from a very large number of flows that contribute negligibly to the overall load. In the rest of the paper we refer to the former as the *elephant* class, and the latter as the *mouse class.* The term *elephants and mice* has been used differently throughout the literature. In this work, elephants and mice are defined according to their average bandwidth (as opposed to file sizes or flow durations). In [9] they use the terms α and β flows when this phenomenon is applied to bandwidth rates of 5-tuple flows.

We investigate a number of schemes for defining a threshold value that allows one to separate out the elephants from the mice. We collect packet measurements from Sprint's IP backbone, and study flow bandwidths at the network prefix level. Initially we consider two single-feature classification methods. The first method exploits the property that the network-level prefix flow bandwidth distribution is heavy tailed. It identifies the threshold separating elephants from mice, as a cut-off point in the Complementary Cumulative Distribution Function (CCDF). This technique makes no assumptions about how high or low the threshold should be. Instead, it defines the threshold as the first point in the flow bandwidth distribution after which power-law behavior can be witnessed. The second method classifies as elephants the high-volume flows whose cumulative bandwidth is equal to a certain fraction of the overall traffic. This is motivated by practicality as it is simple for carriers to implement.

We use four performance metrics to evaluate these schemes. Our findings illustrate the volatility of elephants in a number of ways. The number of elephants at any time is highly variable, elephant flows need frequent reclassification, and elephants do not remain elephants sufficiently long for traffic engineering applications. We quickly consider two other single feature classification schemes (without examining all four metrics in detail), to illustrate that these other methods also yield highly volatile elephants. A comparison of all these schemes reveals an interesting tradeoff, namely that it is difficult to keep both the elephant load and the number of elephants consistent (i.e., small fluctuations) at the same time.

Because persistence in time is an important property, and simple single feature classification schemes cannot achieve this, we next propose a two feature classification method. We suggest a metric, called *latent heat,* that incorporates temporal behavior of a flow. By adding this into the classification we avoid unnecessary reclassification of flows due to short-term fluctuations that do not change a flow's essential nature. We show that this scheme achieves a better tradeoff in terms of fluctuations in the elephant load and the number of elephants, while simultaneously yielding elephant flows with much larger average holding times in the elephant state.

In Section II, we describe the data analyzed throughout the paper. In Section III we present our methodology, and in Section IV, we present initial results. We improve on the proposed approaches in Section V, and characterize the identified elephant network prefixes in Section VI. We conclude in Section VII.

II. Measurement environment

The data used in this paper comes from packet traces collected in the core of Sprint's Tier-1 IP backbone network [8]. Optical splitters are used in conjunction with passive monitoring equipment to collect 44-byte headers from every IP packet traversing the monitored link. Monitoring equipment has been installed in three major POPs in the USA. The analysis described throughout the paper has been performed on 6 traces collected on OC-12 and OC-48 links, yielding similar results. In this paper we present results only from two different OC-12 links. The specifics of those two traces are presented in Table I. The links used are two hops away from the periphery of the network, and traffic is captured on its way towards the core of the network. Therefore, traffic towards specific destinations should exhibit a sufficient level of aggregation. The utilization levels of these two links are given in Figure 1 (The times displayed in the figures are always in EDT.). We selected those two particular links because they exhibit different behavior, which has implications on the classification process as will be seen in later sections. The first of these links provides a data set that varies in a smooth fashion throughout the day. The second link exhibits a more pronounced bursty or busy behavior.

Trace label	Start time	End time	Link Speed
west coast	Jul 24 08:00:35 2001 EDT	Jul 29 02:42:55 2001 EDT	OC-12
east coast	Jul 24 08:00:34 2001 EDT	Jul 25 13:21:55 2001 EDT	OC-12

TABLE I
DESCRIPTION OF COLLECTED TRACES.

The first issue addressed is the granularity level of the flows to be classified. For the purpose of intra-domain traffic engineering the natural granularity level is that of network prefixes appearing in routing tables. Other candidates

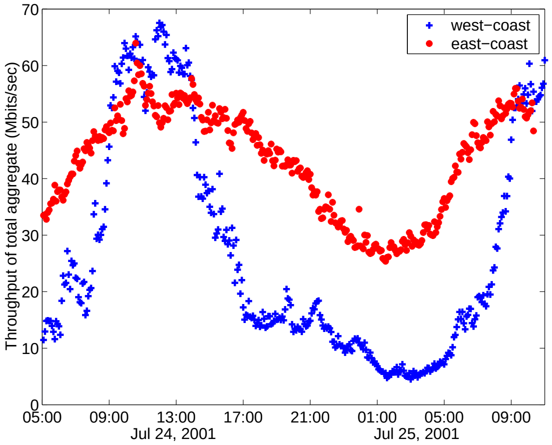

Fig. 1. Link utilization for the west and east coast trace.

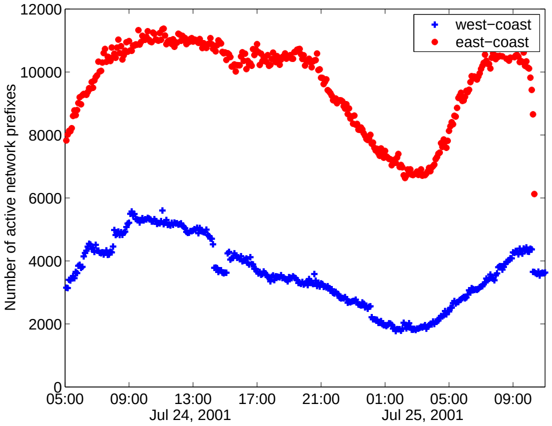

Fig. 2. Number of active network prefixes for the west and east coast trace.

could include flows defined by their destination address, a fixed prefix length, the usual five-tuple (source address, destination address, source port, destination port, protocol id), or the AS level.

Network prefixes is a natural candidate for a few reasons. First, such flows can easily be manipulated by simply changing the next hop address in the routing table for that particular flow. Second, routing policies tend to be applied to a network prefix as a whole, since network prefixes are the smallest routable entities in the Internet. Therefore any change in policy would affect the entire flow as we define it. This would not the case, for example, for flows defined by destination address or prefixes of a fixed length. Third, it has been observed that elephants and mice do exist at this level of granularity [6].

In parallel to the packet trace collection, we collect the BGP routing tables at the corresponding POPs. Those BGP tables are default-free and contain approximately 120K entries. We calculate the volume of traffic headed towards each BGP destination and compute the average bandwidth of each flow over 5 minute time intervals. We select 5 minutes as the measurement interval, since this is the default time interval at which SNMP information is usually collected. We find that in any given measurement interval, approximately 90% of the network prefixes have no traffic traveling towards them. We define a flow to be *active* if it receives at least one packet during the measurement interval. Figure 2 presents the number of active prefixes at each measurement interval over a two day period. The observation that there are only a few flows active at any given time slot is further confirmation that this aggregation level is appropriate for the identification of elephants in that it produces a limited number of flows on which computations need to be carried out.

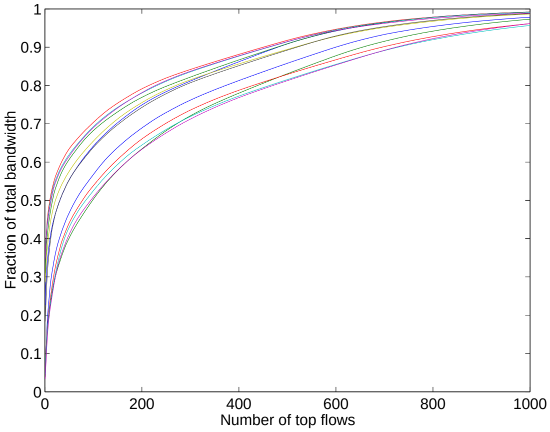

Fig. 3. Cumulative Distribution Function of flow bandwidths for the first 12 5-minute intervals of the east coast trace.

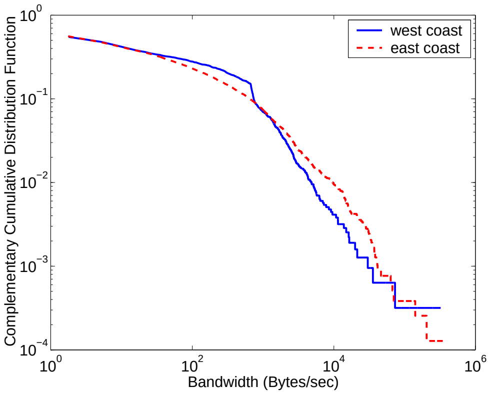

Fig. 4. Complementary Cumulative Distribution Function of flow bandwidths for the east and west coast trace.

The cumulative distribution function (cdf) for the bandwidth of the derived flows is presented in Figure 3. Each curve corresponds to a different 5 minute interval within the first hour of the trace. We see that at any time during this hour, the top 100 flows account for 50% to 70% of the total traffic, while the top 400 flows account for 75%-85%

of the total traffic. This confirms that elephants and mice do exist in our data set at the chosen granularity level.

We further examine our dataset for heavy-tail properties. In Figure 4 we present the complementary cumulative distribution function plot of the bandwidth values of our flows. The straight line at the end of the distribution indicates that it may be heavy-tailed. We use the *aest* test designed by Crovella and Taqqu (described in [4]) to estimate the scaling parameter of the distribution of the flows' bandwidths. We find a scaling parameter whose values vary between 1.03 and 1.2, verifying that the bandwidth distribution is indeed heavy-tailed. In Section III, we describe the properties of the heavy-tail distributions, and we devise a method in which we makes use of these properties in the classification process.

III. Single Feature Classification

Using our packet-level measurements, we aggregate traffic into "flows", based on its network prefix as announced through BGP [11]. This gives us a set of flows, and an associated bandwidth for each flow. We now describe techniques that could be used to classify these flows into "elephants", and "mice".

Before proceeding we introduce some notation. Let i denote the index of a network prefix flow, i.e., a BGP routing table entry. Let τ denote the length of the time interval over which measurements are taken. Time is discretized into these intervals, and n is the index of time intervals. We define $X_i(n)$ to be the average bandwidth of the traffic destined to a particular network prefix i during the n^{th} time slot of length τ. We use C_1 to represent the set of all flows considered elephants, and C_2 to represent the set of flows considered mice.

Our methodology consists of two phases: 1) threshold detection phase, and 2) threshold update phase. In the first phase, we seek for a bandwidth value $v(n)$, such that if a flow's average bandwidth in the chosen time interval exceeds the threshold, it is classified as an elephant. Otherwise, it is considered a mouse. In the second phase, we update the threshold, i.e., at time $n+1$, based on the current threshold detected plus past threshold values.

A. Threshold Detection Phase

Previous attempts at classifying flows into elephants and mice use a single classification feature, which is either 1) bandwidth [6], [12], or 2) duration [10]. Given that we are interested in high-volume flows, for the first step of our methodology we focus on techniques that identify elephants based on their bandwidth alone. Issues regarding the duration of the elephant flows are investigated in later stages.

In the detection phase of our classification process, we analyze the properties of the dataset collected and calculate a threshold value $v(n)$ accordingly. This value $v(n)$ is likely to change with time always isolating those flows that contribute the highest amount of information in each time interval.

We propose two different techniques to identify the initial threshold value.

- The first method takes into account the heavy tail property of the collected data set. The threshold value is set in such a way so that a flow is characterized as an elephant, only if it is located in the tail of the flow bandwidth distribution.
- The second technique is more intuitive, and could be preferred in operational environments. It requires the setting of an input parameter corresponding to the fraction of total traffic we would like to place in the elephant class. The threshold is set in such a way that all the flows exceeding it account for the requested fraction of the total traffic.

1) AEST approach: A random variable X follows a heavy-tailed distribution (with tail index α) if

$$P[X > x] \sim cx^{-\alpha},\ as\ x \to \infty,\ 0 < \alpha < 2 \qquad (1)$$

where c is a positive constant, and where $\sim$ means that the ratio of the two sides tends to 1 as $x \to \infty$. AEST is a method for the estimation of the scaling parameter α proposed by Crovella and Taqqu [4].

The particular value of α is important in many practical situations, and a number of methods are commonly used to estimate its value. For instance, values of α lower than 1 indicate that the random variable X has an infinite mean. Values of α between 1 and 2 indicate infinite variance.

The AEST method identifies the portion of a dataset's tail that exhibits power-law behavior. It's based on the fact that the shape of the tail determines the scaling properties of the dataset when it is aggregated. By observing the distributional properties of the aggregates one can make inferences about where in the tail power-law behavior begins. We use this method to identify the first point in the distribution after which power law properties can be witnessed. We set this particular bandwidth value as the threshold value separating elephants and mice in that given measurement interval.

In other words, if we define $\bar{F}(x)$ as the complementary cumulative distribution function of the flow bandwidths, i.e. $\bar{F}(x)$ = 1 - F(x) = P[X>x], then $v(n)$ = $\hat{x}$, so that

$$\hat{x} = \min_{x}(\frac{dlog\bar{F}(x)}{dlogx} \sim -\alpha)$$

The AEST methodology has the advantages of being non-parametric and easy to apply. Moreover, it has been shown to be relatively accurate on synthetic datasets. More importantly, there has been evidence that the scaling estimator, as calculated by AEST, appears to increase in accuracy as the size of the dataset grows. Given that our datasets feature at least two thousand measurements, such a technique will provide us with reasonably accurate values for the cutoff points in the heavy tail flow bandwidth distribution measured in each time interval.

2) Constant load approach: This second approach is a more pragmatic approach towards identifying the threshold bandwidth values separating elephants from mice. For each time interval n, we define local threshold $v(n)$, such that the flows exceeding this threshold account for α% of the total traffic on the link. We sort the flows and starting from the largest, we proceed down the list adding flows into the elephant class. We stop when the volume of the elephant class reaches α% of the link load. If we rank bandwidths in ascending order, let j denote the index of a flow, and M denote the highest bandwidth flow in timeslot n, then $v(n) = X_{m-1}(n)$ where m is chosen such that: $\sum_{j=M}^{m} X_j(n) < \alpha * \sum_{j=M}^{1} X_j(n)$, and $\sum_{j=M}^{m-1} X_j(n) >= \alpha * \sum_{j=M}^{1} X_j(n)$.

B. Threshold Update Phase

To make use of the threshold derived for engineering purposes, we classify the data in time slot $(n+1)$ according to $v(n)$. This may not be adequate because we may not have a sufficiently large number of measurements in each time interval to yield representative estimates. In addition, we expect that correlations exist across time slots (shown later in Section IV-B) and are thus motivated to allow past data to influence our choice of threshold. We thus calculate a more general classification threshold $\hat{v}(n)$ using a simple first order auto-regressive filter on $v(n)$ as follows.

$$\hat{v}(n+1) = (1-\beta) * \hat{v}(n) + \beta * v(n) \qquad (2)$$

We call this the *update rule*. We found that a value of $\beta = 0.2$ leads to a sufficiently smooth $\hat{v}(n)$. We now use the threshold $\hat{v}(n)$ to classify the traffic during the $(n+1)^{st}$ time interval. In other words, the final *decision rule* is the following.

$$\text{if } X_i(n) > \hat{v}(n) \text{ then } i \in C_1 \text{ else } i \in C_2 \qquad (3)$$

C. Threshold Evaluation

In most applications of classification each data object inherently belongs to some particular class. This class is unknown to the observer whose job is to guess correctly the true state of nature. Our situation is different in that the flows do not naturally belong to a particular class. Instead we are imposing a categorization onto our flows because it is useful for traffic engineering purposes. Because of this, the traditional performance measures based on misclassification errors cannot be used. The evaluation of the type of classification techniques proposed here should be based on the needs of the particular traffic engineering application using the classification. The first two metrics we use for the identification of the elephants are the following.

1) The fraction of total traffic apportioned to the elephant class should be sufficiently large.
2) The number of elephant flows should be reasonably small, say less than a number F, where F is constrained by the router capabilities, and denotes a reasonable number of flows for which the router can keep state.

D. Persistence of Classification in Time

We explained earlier that for traffic engineering purposes we would like flows to retain their class as long as possible. Ideally the majority of the elephants should remain elephants for long periods of time, i.e. in the order of few hours. Intuitively this means that we would like to classify as elephants those flows whose volume is above the identified threshold consistently.

The length of time that an elephant can remain an elephant is both a function of the flow itself and of the classification. It is a function of the classification in the sense that a particular high-volume flow will remain in the elephant class as long as the continually adjusting threshold stays lower that its average bandwidth. Because traffic on the Internet is highly variable, we don't expect flows to retain their classification indefinitely. We know that there will be many flows that will burst for a small amount of time, and become quiet afterwards.

To define a performance metric to capture the persistence of the classification in time, we proceed as follows. Note that the classification scheme we have proposed, induces the following underlying two-state process on each flow. Let $I_i(n) = 1$ if $i \in C_1$, and $I_i(n) = 0$ if $i \in C_2$. At each classification time interval, the process either transitions to the other state or stays in the same state. We define the following two performance metrics in terms of this process.

3) For each flow i, we calculate the autocovariance for a lag of 1, namely $Cov_i(1) = E[I_i(n) * I_i(n+1)]$. This metric reflects how predictable the class of flow i is at time $n+1$ based on its class at time n. The classification scheme with the larger autocovariance is the one in which the flows are more likely to retain their classification.
4) For each flow i, we calculate the average holding time the flow spends in the elephant state, and its average holding time in the mouse state. The classification scheme that yields larger holding times is more attractive for our purposes.

IV. Results

In this section we present the results derived applying the two methods described in the previous section. We compare the two methods with respect to the four performance metrics identified. We discuss temporal issues related to classification such as the duration of the validity of the classification, and the timescale of the measurement and classification intervals.

A. *Classification Results*

Figures 5, 6, and 7 present the threshold values computed using the two proposed technques, the number of elephants, and the fraction of total traffic apportioned to elephants for both traces.

For the west coast trace, the threshold obtained with the "constant load" approach fluctuates throughout the day between 1 KBytes/sec and 2.5 KBytes/sec. As already shown in Section II, the west coast trace exhibits high utilization levels during the peak period (Figure 1). A burst which is not accompanied by similar bursty behavior in the number of active network prefixes (Figure 2). This implies that the burst in link utilization is not due to new network prefixes becoming active, but rather due to existing network prefixes receiving traffic at higher rates. Given this rise in the bandwidths achieved by the already active network prefixes, it is expected that now the threshold value isolating the flows contributing 80% of the total traffic is much higher.

The "aest" approach provides a lower estimate for the threshold. The reason why the threshold is lower‘ in the "aest" approach is because many of the flows in the peak period have now been positioned in the tail of the bandwidth distribution.

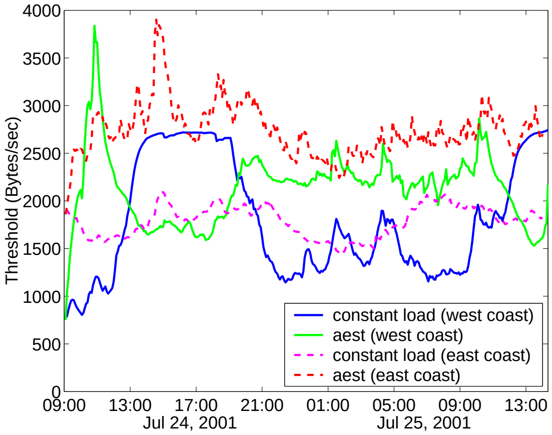

Fig. 5. Derived thresholds $\hat{v}(n)$ for *0.8-constant load*, and *aest*.

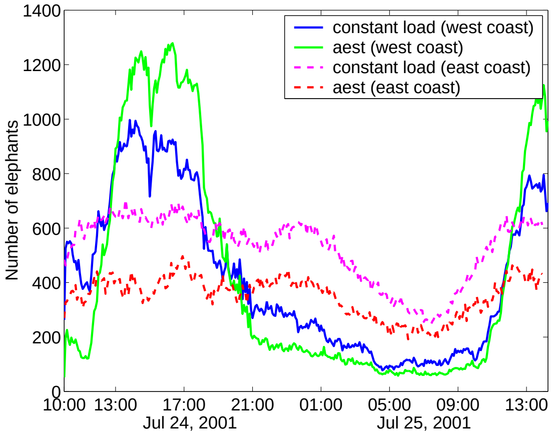

Fig. 6. Number of elephant flows for *0.8-constant load*, and *aest*.

The effect of the burstiness in the utilization of the west coast trace is that the number of elephants identified with both techniques is much larger. Given that the "aest" threshold is much lower than the "constant load" one, the number of elephants identified with "aest" is much higher, thus accounting for as much as 90% of the total traffic on the link.

For the smoother east coast trace, where the trend in the link utilization is accompanied by similar trend in the number of active network prefixes, the threshold values calculated with both techniques behave similarly. The threshold values are slightly higher for the "aest" approach, accounting for less elephants, and smaller fraction of the total traffic. It is important to notice, though, that for the east coast trace, the total amount of traffic apportioned to elephants

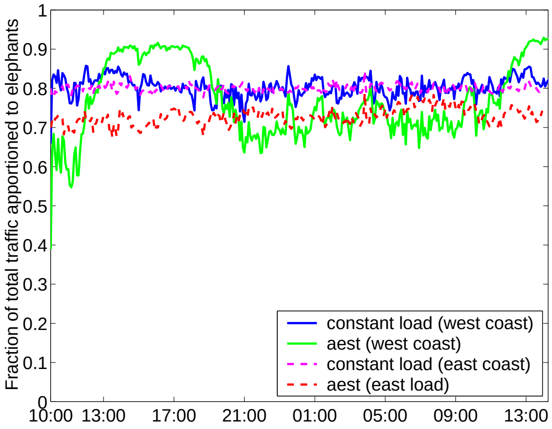

Fig. 7. Fraction of total traffic accounted to elephants for *0.8-constant load*, and *aest*.

is aproximately constant with time, and equal to 80% for the “constant load” (imposed by the technique itself), and 70% for the “aest” approach respectively.

From the previous discussion it can be easily seen that in cases of bursty link utilization, maintaining a constant amount of traffic in the elephant class leads to a volatile number of elephant flows, and equaly bursty threshold values. Moreover, in such cases, the performance achieved using the two suggested techniques is much different. In the case of smooth traffic the two approaches achieve comparable results.

B. *Time Behavior of Classification*

As discussed in Section III-C, we the flows to retain their classification as long as possible. We say “as long as possible” because on the one hand we want to avoid updating state too frequently, but on the other hand we don’t want to keep flows in a particular state if their bandwidth changes significantly. To assess this issue of the persistency of the classification with time, we examine our last two metrics, namely, the autocorrelation value at lag 1, and the average holding times in the elephant state. We compute these values for the peak hours, which is the period during which persistency is perhaps mostly needed. The peak hours are between 12pm and 5pm EDT.

Table II shows that the $I_i(n)$ processes achieve high values of autocorrelation for both approaches and for both traces. What has to be noted is that the values presented in Table II are highly biased towards the performance of the mice flows. The total number of destination prefixes that become active at some point throughout the duration of the west coast trace is 30,241 out of which the elephants vary between 454 and 994 (depending on the time interval n). In other words, only a small number of the autocorrelation values correspond to elephant flows. This is one of the reasons we were motivated to look at the holding times of the classification process.

	East Coast		West Coast	
	AEST	constant load	AEST	constant load
min.	0.8611	0.8687	0.8497	0.8583
5%	0.9789	0.9786	0.9737	0.9769
avg.	0.9827	0.9823	0.9814	0.9819
95%	0.9833	0.9833	0.9833	0.9833
max.	0.9913	0.9913	0.9906	0.9912

TABLE II
AUTOCORRELATION VALUES AT LAG 1 FOR PEAK HOURS.

	East Coast		West Coast	
	AEST	constant load	AEST	constant load
min.	1	1	1	1
20%	1	1	1	1
50%	1.5	1.5	2	2
avg.	3.72	4.1	8.63	8.09
80%	3.5	3.5	12	10.5
max.	60	60	60	60

TABLE III
AVERAGE HOLDING TIMES IN THE ELEPHANT STATE FOR PEAK HOURS (IN 5-MINUTES SLOTS).

We compute the average holding times for each flow in the elephant state for the five hour busy period. The distribution of the average holding times is shown in Figure 8. The x-axis values represent time slots of 5 minutes each (so a value of 12 corresponds to 1 hour). Table III presents the basic statistics for the average holding times in the elephant state. We see from this table that the average holding time is approximately 20 minutes for the east coast trace, and 40 minutes for the west coast trace. Moreover, more than 80% of the flows have average holding times less than half the maximum.

It is clear from Table III and Figure 8 that few flows remain elephants “forever” (i.e., the duration of the whole peak period). It is natural to ask how well past behavior can predict future behavior. In particular we would like to know what is the likelihood that an elephant that has been an elephant for a certain amount of time will remain an elephant “forever”. To measure this we calculate the probability that an elephant will remain an elephant for the whole duration of the peak period, given that it has stayed in the

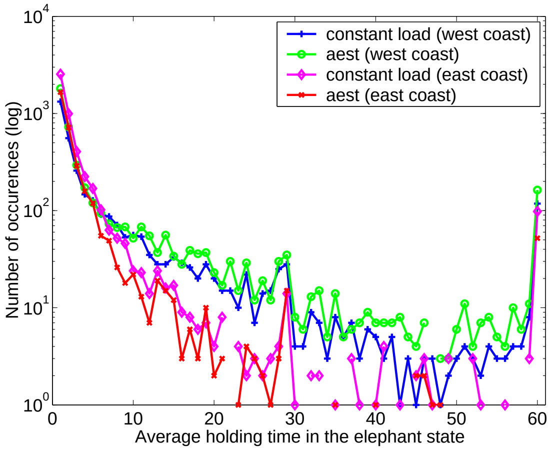

Fig. 8. Average holding time in the elephant state for peak hours (in 5-minutes slots).

elephant state for a given number of intervals. In Figure 9, we present those probabilities for both traces and for both approaches.

Figure 9 shows that for the bursty west coast trace the predictability of the elephants is very limited. Flows that remain elephants for 36 intervals (i.e. 3 hours) during the peak period have a probability of less than 0.2 of remaining in that state for the whole 5 hours. For the east coast trace elephants show a higher level of predictability, and if they remain elephants for 3 hours, they have a probability greater than 0.5 of remaining elephants for the whole duration. Therefore, Figure 9 provides another indication that network prefix flows are very volatile, and once persistency is the most desirable property of the classification, simply classifying flows based on their bandwidth at each interval is not sufficient. In practical settings such small probabilities of successfully identifying elephants in the future has limited applicability.

These results have implications for other methods proposed in the literature. For example, in [5] the authors create state for each flow that exceeds their threshold, and they keep that state for the lifetime of the application. That wouldn't make sense for our data since less than 5% of the flows classified as elephants in any instant remained elephants for the entire 5-hour period. We conclude that flows need to be reclassified fairly frequently.

We saw in the previous subsection that in order to maintain a consistent amount of load in the elephant class, the number of elephants is quite volatile. The fact that average holding times are short is another indicator of the volatility of elephants. If elephants are indeed volatile, then they could need to be reclassified often if the goal is to maintain a consistent load in the elephant class.

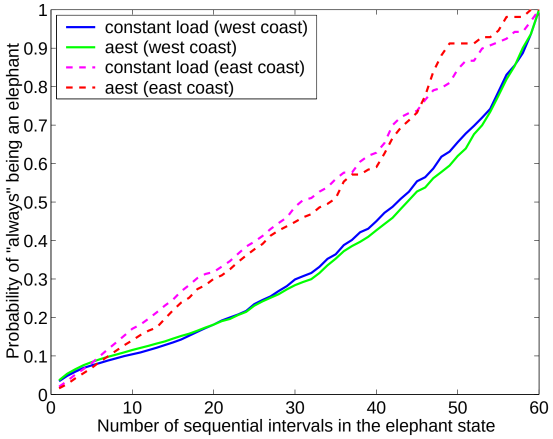

Fig. 9. Probability of "always" remain an elephant for *0.8-constant load*, and *aest*.

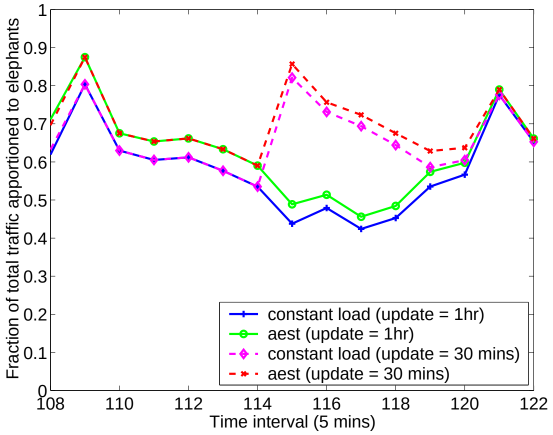

Fig. 10. Effect of infrequent classification on the load of the elephant class.

Until now we have considered reclassifying flows every 5 minutes. In order to evaluate other time scales for reclassification, we consider two other cases, namely reclassifying every 30 minutes and 1 hour. Even though we reclassifying at these intervals, we continue to calculate the separation threshold values every 5 minutes. Figure 10 shows the amount of traffic apportioned to elephants using both techniques with these longer reclassification times. This helps us to understand how fast the list of flows belonging to the elephant class becomes stale. Regardless of the method used for classification, we see that when flows are updated at longer intervals, the fraction of total traffic apportioned to the elephant class may drop dramatically, in particular it can fall as low as 40% within 25 minutes from the last classification interval. Flows classified as elephants at time slot 109 (Figure 10) experience lower bandwidths in the next intervals, while flows that may experience high loads cannot be placed in the elephant class, until the new update period. Moreover if we update too slowly, we can experience a significant jump in the elephant load at the time of reclassification (e.g., at time interval 115). Note that not updating the list of elephants sufficiently often may endanger the efficacy of the traffic engineering application because it may no longer be giving differential treatment to the largest flows.

C. Other Single-Feature Classification Techniques

The two metrics for the number of elephants identified, and the fraction of total traffic apportioned to the elephant class are clearly related. The target of our second classification approach is to maintain a constant elephant load. In this section, we report results about simple classification approaches that could be selected and would target at maintaining a constant number of elephants or a threshold consistently equal to a specific fraction of the link utilization or capacity. This section should not be conceived as a complete evaluation of the techniques described, but a simple illustration of how alternative techniques would perform.

A simple way to identify the high-volume flows in a network is to collect the bandwidth distribution for the destination prefixes in each interval, and to classify the top N flows at each interval as elephants. For our west coast trace, we set N equal to 100 (approximately the number of flows consistently in the elephant class for the 5-peak hours), and measure the fraction of total traffic apportioned to the elephant class[1]. For comparison reasons, we also calculate the threshold value as the bandwidth of the smallest elephant. This is the bandwidth values flows would have to exceed to be classified as elephants in the threshold-based approaches.

Our results indicate that with such a technique the elephant load may fluctuate between 35 and 94% of the total traffic. This points to an important tradeoff. We saw in Figures 6 and 7 that methods that succeed in keeping the fraction of total traffic in the elephant class reasonably consistent yield large fluctuations in the number of elephants. In this alternate approach that focuses on keeping the number of elephants constant, we observe large fluctuations in the load due to elephant flows. It thus appears difficult to keep both the elephant load and the number of elephant from experiencing fluctuations simultaneously.

The corresponding threshold range for this method was 1 to 12 KBytes/sec depending on the time of day, exhibiting similar busy periods, as the utilization of the respective link. An important result of this classification approach is that only 11 flows remained elephants for the whole 5-hour peak period, while 97% of the other elephant flows featured holding times of less than 24 intervals (i.e. 2 hours). During this 5-hour period, approximately 1800 flows became elephants at some point (in both traces). Since only 11 of these remained elephants for the entire period, the probability that a given flow in the set of top 100 at any given interval remains among the top 100 flows for a very long duration is minimal.

We present results when the threshold value is set equal to 1% of the utilization of the monitored link, as proposed in [5] (1% of the capacity of our OC-12 links was never exceeded by a single flow in our traces). Once we set the threshold value in such a way, then the number of elephants identified is limited between 3 and 23 flows, and the amount of traffic they correspond to fluctuates between 10 and 60% of the total traffic. The maximum fraction of traffic is apportioned to the elephant class in the off-peak hours, when the utilization is much lower, and therefore easier to exceed. Such a classification process offers much lower autocorrelation values than any other classification technique, while the maximum holding time at the elephant class is 60 5-minute intervals, achieved by only 2 flows. All the other elephant flows achieve holding times of less than 6 intervals (i.e. 36 minutes). The statistics for the average holding times achieved with the four single-feature classification methods, described, are summarized in Table IV.

	Average holding time statistics					
	min.	20%	50%	avg.	80%	max.
AEST	1	1	2	8.63	12	60
constant load	1	1	2	8.09	10.5	60
constant number	1	1	1.2	3.47	2.75	60
1% utilization	1	1	1	4.12	3	60

TABLE IV

STATISTICS FOR THE HOLDING TIMES AT THE ELEPHANT STATE FOR ALL FOUR SINGLE-FEATURE CLASSIFICATION METHODS (WEST COAST IN 5-MINUTE SLOTS).

D. Other Timescales for Measurement and Classification

Throughout the paper we have considered measurement and classification intervals of $\tau = 5$ minutes each. As stated in Section III, we chose 5 minutes for practical reasons because this corresponds to the interval at which SNMP information is normally collected. In order to assess the utility of this time scale we also considered $\tau = 1$ and $\tau = 60$ minutes.

[1] The nature of our results did not change for different values of N.

Table V presents the statistics for autocorrelation values achieved at lag 1, calculated over the five peak hours. These values are much smaller than the ones in Table II. At $\tau = 60$ we see more of a distinction between the two schemes than we did when $\tau = 5$. We also calculate the main statistics for the average holding time at the elephant state, and we observe that the 80th percentile of the derived distribution is equal to 3 intervals (i.e. 3 hours).

Therefore, it is evident that even at higher timescales, when flows' bandwidth would be expected to behave in a smoother fashion, volatility is the main property of the elephant flows. We notice that the autocorrelation at $\tau = 60$ is approximately 10% less than what it is at $\tau = 5$, and the number of intervals an elephant may remain as an elephant is much smaller than the one achieved when $\tau = 5$. Those remarks give us confidence in that the results presented so far are not an artifact of the timescale used but rather a property of the traffic itself.

We also consider the interval $\tau = 1$ minute. We find that the autocorrelation for $\tau = 1$ and $\tau = 5$ are almost identical, indicating that there is not any need to go lower than 5 minutes. This is only an initial examination into the appropriate time scale for such a classification; however it gives us confidence that we are using reasonable values.

	East-Coast τ = 60 mins	
	AEST	0.8-constant load
min.	0.5833	0.8333
5%	0.8333	0.8333
avg.	0.8318	0.8373
95%	0.8333	0.8571
max.	0.8889	0.8571

TABLE V

AUTOCOVARIANCE VALUES AT LAG 1 FOR THE PEAK 5 HOURS.

V. TWO FEATURE CLASSIFICATION

It is evident from the previous section that regardless of the criterion we use to define a separation threshold, we still see a lot of volatility in the elephants. What we want to avoid is reclassification of a flow if it transitions to the other side of the threshold for a short time. Large flows can have short transient periods in which their volume drops. Similarly, small flows can have short-lived bursts. Although many flows may spend most their time on the same side of the separation threshold, every so often they cross the threshold for just one or two time intervals. We do not want to reclassify flows due to short transient bursts or dips. This motivated us to include a second feature in the classification that would be based on a temporal component of an elephant's behavior. For traffic engineering purposes, it is important to be able to differentiate between flows that contribute large amounts of traffic yet may experience short-lived periods of reduced bandwidth, and flows that fall below the threshold and remain there for the rest of their lifetime. As a consequence, our reaction to transient drops below the identified threshold should exhibit sufficient latency.

In thermodynamics, "latent heat" is the heat energy required to change a substance from one state to another. In order to be able to identify the points in time, when elephant flows below the threshold change state into mice, we define a new metric, which we call "latent heat". An elephant is assumed to accumulate "latent heat" as long as it exceeds the identified threshold. Once it falls below the threshold, it starts losing "heat", and changes state (i.e. becomes a mouse) when its "latent heat" becomes lower than a preset value.

In order to achieve this, at each time interval, apart from the flow's bandwidth we calculate the distance between the bandwidth achieved and the corresponding threshold value, derived using the proposed approaches such as the "aest", and the "constant load". We define the "latent heat" of a flow as the sum of those distances in the past 12 timeslots, i.e. the previous hour.

$$LH_i(n) = \sum_{t=n-12}^{t=n} (X_i(t) - \hat{v}(t))$$

In each classification interval $n + 1$, we classify flows as follows.

1) if $X_i(n+1) > \hat{v}(n)$ and $LH_i(n) > 0, i \in C_1$
2) if $X_i(n+1) > \hat{v}(n)$ and $LH_i(n) < 0, i \in C_2$
3) if $X_i(n+1) < \hat{v}(n)$ and $LH_i(n) > 0, i \in C_1$
4) if $X_i(n+1) < \hat{v}(n)$ and $LH_i(n) < 0, i \in C_2$

With the first condition, we can track flows that systematically transmit over the identified threshold. If the flow happens to transmit under the threshold for a small number of intervals, then with condition (3) we make sure that it does not change its state. With the fourth condition, we classify as mice those flows that fall consistently below the threshold. If those flows happen to transmit over the threshold for a small amount of intervals, then they remain mice (condition 2). The "latent heat" metric takes into account how much beyond or below the threshold a flow may transmit and thus accounts for major changes in the flow's behavior. In other words, if a flow has been idle and starts transmitting at three times the identified threshold, then it will be classified as an elephant within at least 4 time intervals, assuming the threshold remains constant.

We classify flows based on the threshold values calculated using the “aest” approach, and their “latent heat” in time. The resulting number of elephants and fraction of total traffic apportioned to the elephant class is presented in Figures 11 and 12 for both traces. If we compare these figures with those of Figures 6 and 7, we see that with this classification approach we can keep the load reasonably consistent (Figure 12) while simultaneously experiencing less fluctuations in the elephant load (Figure 11). Hence our two-feature classification scheme using latent heat achieves a better tradeoff between reducing fluctuations in either elephant load or the number of elephants, than any of the single feature classification schemes.

The empirical probability density function for the average holding time in the elephant state is presented in Figure 13. Specific statistics are given in Table VI. There is no longer a peak at small holding time values. Recall that for single feature classification we have average holding times of 20 and 40 minutes (depending upon the trace), while here we see average holding times for both traces are 1 1/2 hours or longer. We believe that classification schemes such as this one, that avoid reclassification for short-term fluctuations, identifies the type of long-lived elephant flows that are good candidates for traffic engineering applications.

Average holding time statistics						
	min.	20%	50%	avg.	80%	max.
west coast	1	7	16.5	23.56	44	60
east coast	1	6	11.5	17.42	24	60

TABLE VI

STATISTICS FOR THE HOLDING TIMES AT THE ELEPHANT STATE FOR THE “LATENT HEAT” APPROACH ON THE AEST THRESHOLDS (IN 5-MINUTE SLOTS).

VI. Profile of Elephant Flows

In this section we examine the profile of the elephants. We considered those flows that were classified as elephants by our two feature classification method throughout the entire 5 hour peak period. As a preliminary assessment on the profile of the elephants, we examine their network prefix lengths and the type of organization they belong to.

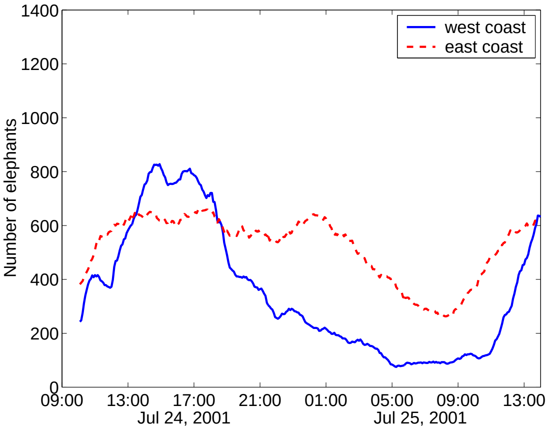

Fig. 11. Number of elephants derived using the two-feature classification method.

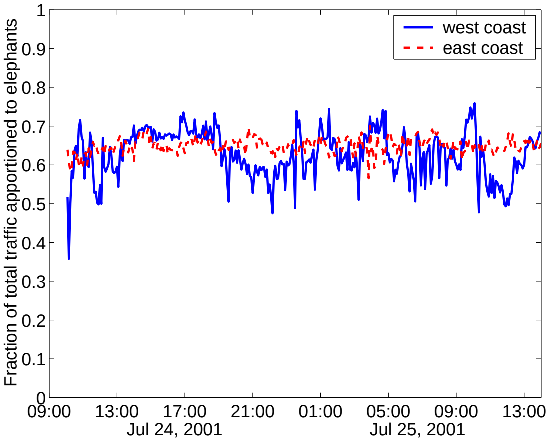

Fig. 12. Fraction of total traffic apportioned to elephants using the two-feature classification method.

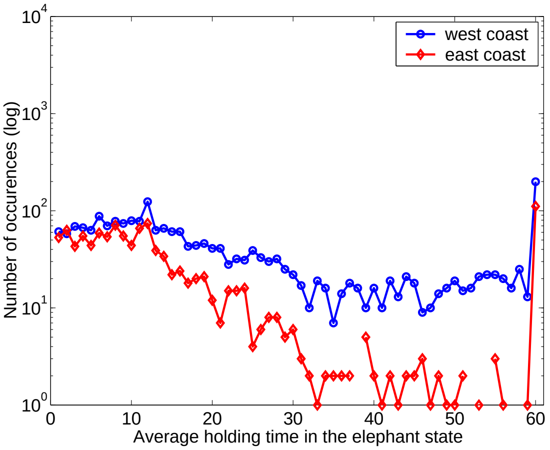

Fig. 13. Average holding times in the elephant state, when classifying based on the aest threshold, and the latent heat of the flow bandwidth in an hour.

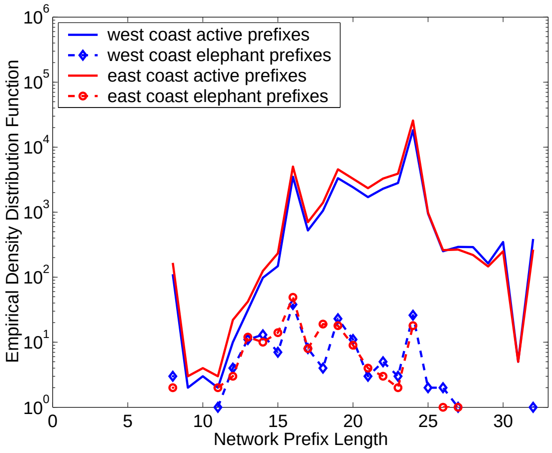

Fig. 14. Network prefix length distribution for the elephants.

One would expect that the network prefixes of the elephants would correspond to large domains advertised through BGP. As already mentioned, the monitored links are OC-12 links interconnecting backbone routers in a POP, and traffic is captured as it travels towards the core of the network. In Figure 14 we present the density function for the length of network prefixes that become active at some point in our trace. We observe that most of our active network prefixes come from /24 networks, followed by /16 and /18. Active networks with prefixes of /8, and /10 are limited to less than 100. In the same figure we present similar statistics for the network prefixes that get classified as elephants. The lengths for the elephant prefixes are contained in a smaller region, namely between /12 and /26. For the west-coast we even witness a /32 network always in the elephant class. This is a customer network using NAT, advertising a single IP address within the Sprint IP backbone.

Finding only two or three (depending upon the trace) /8 flows among our elephants is an unexpected result. Such a result cannot be generalized for the whole Internet because it is influenced by the way our network is engineered, and the way BGP policies impact traffic flow throughout our network. For instance, we compared the elephants identified on the west and the east coast, and we found that a few network prefixes were elephants on only one of the two coasts. This result was expected since BGP policies guide the entry and exit points of our network for traffic that is not destined for our customers.

In order to get a better understanding about the type of businesses contributing the most to the overall load, we classify elephants into the following categories: Tier-1, Tier-2, Tier-3 ISPs, corporations, universities, and government institutions. The definition of what constitutes a Tier-1 network is rather controversial. The most popular definition is that Tier-1 ISPs are those networks that have access to the global Internet Routing Table and do not purchase network capacity from other providers [1]. The second tier generally has a smaller national presence than a Tier 1 ISP and may lease part or all of its network from a Tier 1. Lastly, a Tier 3 ISP is typically a regional provider with no national backbone.

According to the above categorization, we profile the elephants in our network as shown in Table VII. The vast majority of the elephant network prefixes belongs to Tier-1, Tier-2, and Tier-3 ISP providers. We believe that profiling elephant flows is essential once one intends to apply specific traffic engineering policies on the traffic destined towards them.

	Tier-1	Tier-2	Tier-3	corp.	.edu	gov.
west coast	40	42	28	32	19	6
east coast	22	72	27	40	7	7

TABLE VII

PROFILE OF THE ELEPHANT NETWORK PREFIX FLOWS.

Notice that certain businesses may advertise more than one network prefixes for their network. Therefore, the results presented in Table VII do not correspond to unique businesses, but rather to the number of elephant network prefixes in each category.

VII. CONCLUSIONS

In this paper we considered classification schemes for network prefix flows that identify elephants by separating the flows into two classes called elephants and mice. We considered two single feature schemes in depth, and two others summarily. We compared these methods in terms of the number of elephants they yield, and the total load in the elephant class. We showed that single feature schemes that lead to persistency for one of these metrics, lead to fluctuations for the other metric. Hence there is a tradeoff between these two metrics.

We evaluated metrics based on the temporal behavior of elephants. We illustrated the volatility of elephants three ways: 1) we found that flows remain in the elephant state for surprisingly short periods of time; 2) the number of elephants varies considerably over different time intervals; and 3) if elephants are not reclassified frequently, the total load in the elephant class can change substantially.

Elephant flows that are not persistent for sufficient amounts of time, such as periods of time larger than routing protocol convergence times, may not be useful for traffic engineering applications. Therefore, classification techniques should make sure that flows are reclassified only when necessary and not when they exhibit short-term transitions above or below the threshold.

For that reason, we proposed a new classification mechanism based on two features, a flow's bandwidth and its "latent heat". We calculate the "latent heat" of a flow as follows. We measure the difference between a flow's bandwidth and the separation threshold at each time interval, and sum a sequence of these differences over multiple time intervals. This measure smoothes out short-term transitions because an elephant's latent heat will remain positive despite a brief transition below the threshold. Similarly, a mouse's latent heat will remain negative despite a brief transition above the separation threshold. We show that with this two feature classification scheme we achieve a better tradeoff between our first two metrics. The elephant load remains reasonably consistent while the number of elephants fluctuates less than in single-feature classification schemes. More importantly, the propsoed two-feature scheme yields elephants that remain elephants for much longer periods. With single feature classification elephants remained elephants on average between 20-40 minutes, while using the two feature classification approach, elephants remained elephants for 1 1/2 to 2 hours.

In profiling the elephants from both our traces, we found the elephants' prefixes generally lie within a limited range of network prefix lengths between /16 and /26. Surprisingly, we found few large prefixes such as /8 and /10. Although there were a few corporations, government agencies and universities among our elephants, we learned that the vast majority of them are Tier-2 and Tier-3 ISPs.

Overall, we conclude that while single feature classification schemes are attractive in their simplicity, they are probably insufficient for most traffic engineering applications. While we believe that our two feature classification scheme is more attractive to traffic engineering applications, we also conclude that identifying elephants is not straightforward despite their heavy-hitter nature. This is important because the idea of isolating elephants for traffic engineering purposes has been widely proposed, but there has been no prior effort on assessing the feasibility and issues involved in doing so.

References

[1] D. Allen. The impact of peering on isp performance: What's best for you? *Network Magazine*, 2001.

[2] S. Bhattacharyya, C. Diot, J. Jetcheva, and N Taft. POP-level and Access-Link-Level Traffic Dynamics in a Tier-1 POP. *ACM SIGCOMM Internet Measurement Workshop*, November 2001.

[3] M. Crovella. Performance evaluation with heavy tailed distributions. In *Lecture Notes in Computer Science 1786*, March 2000.

[4] M. Crovella and M. Taqqu. Estimating the Heavy Tail Index from Scaling Properties. *Methodology and Computing in Applied Probability*, 1999.

[5] C. Estan and G. Varghese. New Directions in Traffic Measurement and Accounting. *ACM SIGCOMM Internet Measurement Workshop*, August 2001.

[6] W. Fang and L. Peterson. Inter-AS Traffic Patterns and Their Implications. *Proc. IEEE Globecom*, December 1999. Brazil.

[7] S. Floyd. Simulation is Crucial. *IEEE Spectrum*, January 2001.

[8] C. Fraleigh, C. Diot, B. Lyles, S. Moon, P. Owezarski, D. Papagiannaki, and F. Tobagi. Design and deployment of a passive monitoring infrastructure. In *Proceedings of Passive and Active Measurement Workshop*, Amsterdam, April 2001.

[9] S. Sarvotham, R. Riedi, and R. Baraniuk. Connection-Level Analysis and Modeling of Network Traffic. *ACM SIGCOMM Internet Measurement Workshop*, August 2001.

[10] A. Shaikh, J. Rexford, and K. Shin. Load-Sensitive Routing of Long-Lived IP Flows. *Proc. ACM SIGCOMM*, September 1999.

[11] J. W. Steward. *BGP4: Inter-Domain Routing in the Internet*. Addison Wesley, 1999.

[12] S. Uhlig and O. Bonaventure. On the Cost of Using MPLS for Interdomain Traffic. *Quality of Future Internet Services*, September 2000. Berlin.