Network Performance Monitoring at Small Time Scales

Konstantina Papagiannaki[§], Rene Cruz[‡], Christophe Diot[†]

[§]Sprint ATL
Burlingame, CA
dina@sprintlabs.com

[‡]Electrical and Computer Engineering Department
University of California San Diego, CA
cruz@ece.ucsd.edu

[†]Intel Research
Cambridge, UK
christophe.diot@intel.com

ABSTRACT

SNMP statistics are usually collected over intervals of 5 minutes and correspond to average activity of IP links and network elements for the duration of the interval. Nevertheless, reports of traffic performance across periods of minutes can mask out performance degradation due to short-lived events, such as micro-congestion episodes, that manifest themselves at smaller time scales. In this paper we perform a measurement study of packet traces collected inside the Sprint IP network to identify the time scales over which micro-congestion episodes occur. We characterize these episodes with respect to their amplitude, frequency and duration. We define a new performance metric that could be easily computed by a router and reported every 5 minutes through SNMP to shed light into the micro-behavior of the carried traffic. We show that the proposed performance metric is well suited to track the time scales over which micro-congestion episodes occur, and may be useful for a variety of network provisioning tasks.

Categories and Subject Descriptors

C.2.3 [Computer-Communication Networks]: Network Operations - Network Monitoring

General Terms

Algorithms, Management, Measurement, Performance, Design

Keywords

Performance Monitoring, Congestion Detection, Internet measurement

1. MOTIVATION

Large-scale IP networks feature hundreds of routers and thousands of links. Monitoring of the network-wide state is usually achieved through the use of the Simple Network Management Protocol (SNMP)[4]. Routers are configured to collect measurements, in the form of counters, that they report to a specific location at frequent intervals. Those intervals are typically set to 5 minutes. Specific reasons that guide this decision have to do with the size of the network, i.e. the interval has to be large enough so that it allows the polling of the entire network during a single interval. On the other hand, this interval should not be too small so as to avoid overloading the polled network elements and the collection station itself.

Once a network operator decides on the time interval at which network elements are polled, he/she has to decide on the actual Management Information Base (MIB) values that each network element will have to report on. Usually these MIB values are link utilization, packet drops on a per link basis, the CPU utilization of the router itself, etc. In order to avoid router overload and limit the amount of SNMP traffic through the network, network operators usually select a small number of metrics to be polled by the Network Management Station and base their network provisioning decisions on those specific metrics.

Consequently, it is not uncommon for a network operator to collect link utilization measurements and infer delay performance based on the collected information. In fact each network provider usually aims for network-wide link utilizations that do not exceed the "acceptable utilization levels". Those levels are specific to each network provider and are frequently set around 50%. If each link in the network is utilized less than 50% it is capable of carrying the traffic of any equal-capacity link in its neighborhood in case of failure. Moreover, it has been shown in previous analytical work that links utilized less than 50% introduce minimal queueing delays [1].

However, link utilization is typically reported as a 5-minute average value. Network operators use this 5-minute average value as a provisioning tool to ensure that delays through the network are acceptable. In this work we show that traffic counters collected every 5 minutes mask out micro-congestion episodes that occur at time scales of milliseconds. We define a micro-congestion episode as a period of time when packets experience increased delays due to increased volume of traffic on a link.

We perform an empirical study of packet traces collected inside the Sprint Tier-1 IP backbone network to identify the time scales over which micro-congestion episodes manifest themselves. We characterize micro-congestion episodes with respect to their amplitude, frequency and duration. Our approach relies on the analysis of link utilization at multiple

IMC'03, October 27–29, 2003, Miami Beach, Florida, USA.

time scales simultaneously. We show that this type of analysis can provide insight into the time scales at which congestion occurs, while offering information that cannot be recovered by similar analysis at a single timescale. We propose a new performance metric that could be reported through SNMP every 5 minutes and capture network performance at small time scales. This type of measurements could have several applications within an operational setting.

The remainder of the paper is structured as follows. In Section 2 we describe the data analyzed throughout the paper. In Sections 3 and 4 we present our methodology for the characterization of micro-congestion episodes with respect to their amplitude, frequency and duration along with results obtained for OC-3 and OC-12 link measurements. We validate our approach in Section 5. This work should be viewed as a first step toward the identification of micro-congestion episodes and their impact. Thus, in Section 6 we discuss our results and describe the work we intend to carry out in the future. Conclusions are drawn in Section 7.

2. COLLECTED DATA

Complete characterization of micro-congestion episodes at small time scales needs to capture their amplitude, frequency and duration. To perform this type of operation we need detailed packet traces with accurate timing information. We install optical splitters on selected links inside the Sprint IP backbone network and capture the first 40 bytes of each packet seen on each unidirectional IP link. Each record is GPS timestamped. We call the data sets collected *packet traces*.

In this work, we use four packet traces collected inside the network from OC-3 and OC-12 capacity links on August 9th, 2000 and September 5th, 2001. These packet traces are selected because they correspond to pairs of monitored links that attach to the same router inside the network. We call each pair of packet traces a *data set* for the remainder of the paper. We identify those packets that are common to both incoming and outgoing link and call them *matches* [3]. In Table 1 we provide information about the packet traces collected and the resulting *matched* data sets.

Set	Link	Speed	#pkts	Avg. Util.	#matches
1	`in1`	OC-3	$793 \cdot 10^6$	70 Mbps	$3 \cdot 10^6$
	`out1`	OC-3	$567 \cdot 10^6$	60 Mbps	
2	`in2`	OC-12	$1{,}3 \cdot 10^9$	150 Mbps	$18 \cdot 10^6$
	`out2`	OC-12	$1{,}1 \cdot 10^9$	250 Mbps	

Table 1: Details of traces

The routers participating in our measurements are usually characterized as "virtual output queued switches" and delays experienced by packets through them are not due to output queueing alone. Their architecture is briefly described in [2]. We compute the "single-hop delay" for the matches as the difference between the departure and arrival timestamp of a packet at the router [3]. These delay measurements completely characterize the delay performance experienced by packets on the router path between the monitored input and output link. Nevertheless, the output link receives traffic from other input interfaces, and the input link sends packets to other non-monitored output links. In previous work we have established that the collected single-hop delay measurements approximate a random sample for the delays experienced by any packet on the monitored output link [2]. Thus, they are representative with respect to the amount of time packets need to transit the monitored router and get transmitted from that specific output link.

3. AMPLITUDE OF AN EPISODE

The total amount of traffic flowing on a link computed over intervals of different duration can exhibit great variations. Traffic burstiness is likely to demonstrate itself through high throughput values at time scales of milliseconds, while throughput measurements collected over intervals on the order of minutes are likely to be within the provider's "acceptable utilization levels". In this section we look into the appropriate values for the timescale τ, over which throughput measurements should be collected in order to reveal performance degradation due to traffic burstiness at small time scales.

The decision on the appropriate value of τ is not trivial. If the value of τ is too small, then high throughput values may relate to the simple transmission of a small number of packets back to back. For instance consider the following simple scenario. A TCP connection sends a full window of packets that arrive at the router's incoming interface back to back. In the absence of crosstraffic, these packets will appear back to back at the output link. If the measurement interval contains those TCP packets alone, the measured throughput is 100% but packets have experienced no queueing delay since they arrived at an equal-speed link and faced no crosstraffic. Consequently, high output throughput values in this case do not correspond to increased delays through the router.

On the other hand, if the value of τ is too large, it may have no relevance to the time scale at which queues are building up inside the router. The maximum time scale of relevance for the definition of micro-congestion episodes should be related to the total amount of buffering available to packets while transiting the router. This latter time scale is useful when one wants to analyze the loss characteristics of the network, but it may be too great when one is interested in bounding the delay through each node. In reality even a small queue buildup of 1 ms may be considered prohibitive if it occurs frequently in space and/or time.

3.1 Methodology

We detect the amplitude of a micro-congestion episode through its impact on the delay experienced by packets through a single node. For each interval n of duration τ we collect the delay measurements for the packets that departed within that interval {D(n)}, and compute link throughput B as the total amount of traffic transmitted in that same interval. We associate the link throughput value B with the set of delay measurements $D(B) = \{D(n)\}$. We proceed for the remainder of the trace. If a later interval n' is characterized by the same throughput value B we augment set $D(B)$ with the new delay measurements $\{D(n')\}$. At the end of the trace we have associated each link throughput *level* (quantized per 1 Mbps) with a delay distribution.

These results describe the delay behavior experienced by packets for specific levels of output link utilization. One could also consider this relationship as the relationship between the intensity of an output burst and the probability of increased delay due to crosstraffic. If the link utilization is high and the delays measured are high then measurements

reveal buildup of queues inside the router. If the link utilization if high and the delays experienced are small, then measurements simply reveal packets going through the router at line rate.

3.2 Results

We perform the analysis described above on our two data sets. We measure link throughput at intervals of duration τ equal to 1ms, 10 ms, 100 ms, 1 sec and 5 minutes and investigate the relationship between the delays experienced by packets while transiting the monitored router and the output link utilization.

3.2.1 Results for OC-3 links

In our analysis each level of utilization B is accompanied by a delay distribution. For the results presented in this section we characterize the obtained distribution using the minimum, average, 90th and 99th percentile[1]. All statistics are computed for the levels of link utilization that feature at least 100 delay measurements. Our results are summarized in Fig. 1 and Fig. 2.

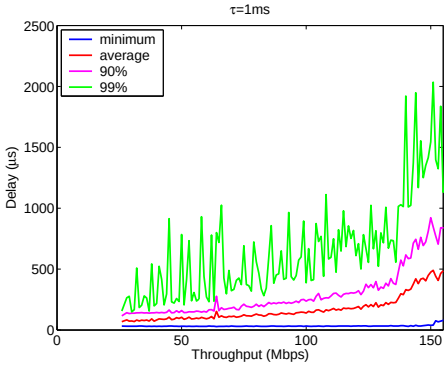

Figure 1: Output link utilization vs. single-hop delay ($\tau = 1ms$, OC-3).

Fig. 1 presents the delay statistics measured for each level of output link utilization for `set1` when τ is equal to 1ms. Link throughput measured over intervals of 1 ms duration approaches the link capacity (155 Mbps) and is accompanied by delays as large as 2 ms through the monitored router. The minimum delay reported for each level of output link utilization is approximately the same across the entire range of possible utilizations. This observation indicates that packets may experience minimal delays during a highly utilized interval. Nevertheless, there is a significant number of packets that are likely to spend milliseconds of delay through a single node.

Consequently, measurements of output link utilization at 1 ms intervals are able to reveal the amplitude of micro-congestion episodes. In Fig. 2 we investigate whether such a relationship persists with higher values of τ. Fig. 2 presents the relationship between the 99th percentile of the delay distribution and the output link utilization for different values of τ. Output link utilization measured at 1 ms and 10 ms

[1]We do not use the maximum delay values because `set1` contains outlier delay measurements that are not related to normal router operation [2].

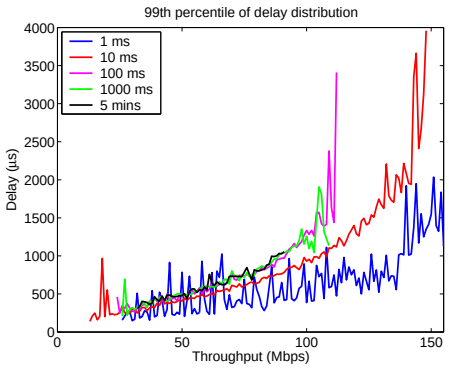

Figure 2: Output link utilization vs. 99th delay percentile (OC-3).

may reach the link capacity despite the fact that a 5 minute average value hardly reaches 60%. In addition, the 99th percentile of the delay distribution reaches higher values when τ is equal to 10 ms and significantly drops for values of τ greater than 1 sec.

Our delay measurements for `set1` occasionally exceed 1 ms with a maximum value of 4 ms. As a result, when τ is equal to 1 ms, there will be packets in our data set that remain inside the router for multiple 1 ms intervals. Since we report delay measurements only when the packets are seen at the output link, delay measurements lag behind the utilization measurements. This is the reason why the 99th percentiles reported at $\tau = 1$ ms are smaller than measurements at $\tau = 10$ ms. This is an artifact of our methodology since we rely on packets exiting from the monitored output link to measure the delay experienced through the router. Passive measurements on the router itself could correctly reveal the amplitude of the episode also at the 1 ms time scale.

Given that the 99th percentile of the delay measurements across the entire trace does not exceed 4 ms, which is correctly measured at 10 ms intervals, both time scales of 1 ms and 10 ms could accurately reveal the magnitude of micro-congestion episodes for `set1`. The 99th percentile of the delay distribution reported when τ =5 minutes is 1 ms, which is significantly smaller than 4 ms. Thus, computing the delay percentile for longer periods of time is likely to hide increased delays that persist for smaller periods of time.

Based on the relationship between delay and output link utilization, a network operator could select the "acceptable" link utilization levels that relate to acceptable delays through each node. For instance, if one considers 1 ms of delay through a single node to be prohibitive then links should not be allowed to exceed 70% utilization at a 10 ms timescale, and 55% utilization at 100 ms timescale (Fig. 2). In that case, 99% of the packets going through the router will experience single-hop delays less than 1 ms.

3.2.2 Results for OC-12 links

We continue our analysis using the second data set consisting of two OC-12 links. We apply our methodology on `set2` and display the relationship between experienced delays and output link utilization in Fig. 3 and Fig. 4.

The output link utilization for the OC-12 set is much lower

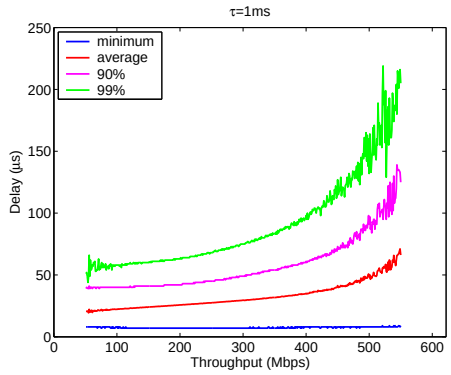

Figure 3: Output link utilization vs. single-hop delay ($\tau = 1ms$, OC-12).

than the output utilization for the OC-3 set even when measured over 1 ms time intervals. As a consequence, and due to the higher link rate, delay measurements for the second data set are much lower. The maximum delay in our data set is 500 μs. In addition, due to the increased line rate the differences between small and large time scale output link utilization measurements are far more significant.

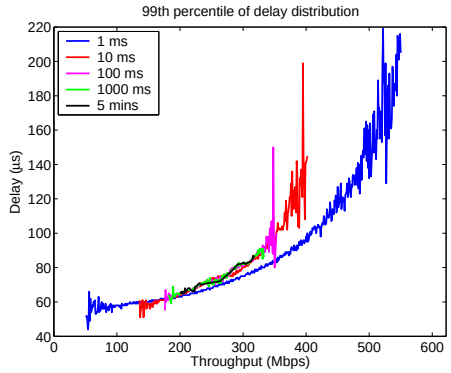

Figure 4: Output link utilization vs. 99th delay percentile (OC-12).

The output link utilization computed over 5 minute intervals hardly exceeds 300 Mbps (i.e. 40%). Nevertheless, the output link utilization measured over 1 ms intervals reaches 97%. The amplitude of the micro-congestion episodes is better revealed when measurements are taken at 1 ms intervals. The 99th percentile of the delays experienced by packets during the "overloaded" 1 ms time intervals is 250 μs, while the respective statistic reported over 5 minute intervals is 70 μs. The fact that the 99th delay percentile significantly drops for 10 ms and 100 ms time intervals indicates that the micro-congestion episodes occur infrequently, thus affecting a small number of packets in total.

4. FREQUENCY AND DURATION

With the previous analysis we showed that micro-congestion episodes can be better revealed when measurements are taken at intervals of duration equal to or below 10 ms. Given specific delay values that are considered prohibitive one can identify the link utilization levels that should not be exceeded for bounded delay performance for each set. These thresholds are likely to be defined by the network operator in accordance to the Service Level Agreements (SLAs) in place. Nevertheless, the impact of higher output link utilization onto the delay experienced by packets is not a function of the amplitude of the micro-congestion episodes alone, but also the frequency of these episodes and their duration.

4.1 Methodology

To characterize the frequency of micro-congestion episodes, we analyze output link utilization at different time scales and for different values of the "acceptable utilization levels" which we call *thresholds* (denoted as th). If a micro-congestion episode is persistent, then its effect should be significant across multiple time scales. For instance, if a link experiences a micro-congestion episode that lasts more than 1 ms, then its effect will "leak" into the higher analyzed time scale, in our case 10 ms. Thus, one way to identify the way micro-congestion episodes persist in time is by analyzing link utilization at multiple time scales *simultaneously*.

Our requirements for the description of micro-congestion episodes are that they can be reported in a concise form at time intervals on the order of minutes. In addition, the statistics that network elements need to compute should be lightweight so as to avoid overload at the network nodes. In order to characterize the frequency of micro-congestion episodes we introduce a new SNMP metric FrOverload defined as follows.

FrOverload(τ, th): the fraction of intervals of duration τ that report a utilization level greater than th within each 5-minute interval.

Observing the statistics collected for a specific time scale τ across different values of th reveals the *amplitude* of the episode. Observing the statistics collected for specific values of th across multiple time scales reveals the *duration* of the episode. For instance, if the number of congested intervals at the smallest time scale is high and the number of congested intervals at the next coarser time scale is low, we can conclude that the congested intervals at the smallest time scale are not sequential and thus their impact on network performance should be limited. Lastly, the number of intervals exceeding specific values of th at a specific time scale reveals the *frequency* of the micro-congestion episodes of the respective amplitude (determined by th).

4.2 Results

4.2.1 OC-3 results

We analyze the output packet trace `out2` and compute link throughput at the time scales of 1 ms, 10 ms, and 100 ms (since these time scales are the ones that can better reveal the amplitude of an episode as shown above). Then, for different values of th, we count the number of intervals within each 5-minute interval in the trace, that exceed the specified threshold values. We present our results for the three time scales and three different values of th in Fig. 5.

We observe that when th is set to 50%, then approximately all the 100 ms intervals until 3 pm can be characterized as "congested"[2]. Nevertheless, not all 1 ms intervals exceed 50%. Therefore, at 1 ms intervals utilization may be

[2]Given that approximately all 100 ms intervals are utilized

less than 50%. The fact that utilization measured at 10 ms still exceeds 50% indicates that when 1 ms intervals exceed 50% utilization they exceed it significantly. Indeed, when $th = 70\%$ there are still 30% of the 1 ms intervals that get characterized as "congested". Such a finding verifies our intuition that throughput measured at the finest time scale can greatly fluctuate with time.

When the threshold value is set to 60%, the highest fraction of "congested" intervals is reported by the 1 ms timescale, but does not equally affect the measurements at higher time scales. Consequently, utilization measurements collected at 1 ms intervals are likely to exceed 60% but do not persist in time and result in lower values at coarser time scales. Once the threshold is set to 70% the difference in the fraction of "congested" intervals across time scales is even greater.

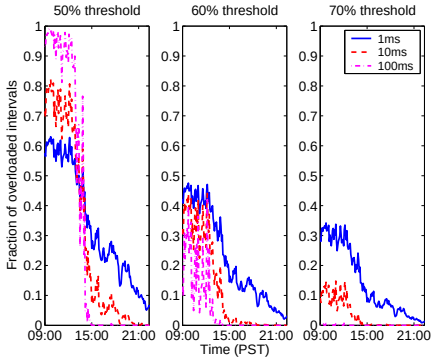

Figure 5: Fraction of overloaded intervals across time for different values of th and τ (OC-3).

4.2.2 OC-12 results

In Fig. 6 we present the results for the OC-12 data set for τ equal to 1ms, 10 ms, and 100 ms and th equal to 50%, 60% and 70%. We observe that the time scales of 10 and 100 ms report a greater number for the fraction of overloaded intervals only when $th = 50\%$. The largest number of "congested" intervals occurs toward the end of the trace after 12pm local time. Link utilization measurements at 1 ms intervals rarely exceed 70% (their fraction is less than 0.05) and only marginally exceed 60% (their fraction reaches 0.2 at most).

Consequently, the number of micro-congestion episodes in the OC-12 data set is significantly smaller but not negligible. In addition, given that the amplitude of these episodes hardly reaches 60% we anticipate that their impact on the performance observed by packets is going to be limited.

5. VALIDATION

In summary, we conclude that by looking at time scales below 100 ms and across different values of th we can characterize the amplitude of the micro-congestion episode. If the throughput measured at 10 ms is greater than 60% on an OC-3 link then according to Fig. 2 we expect delays to be significant. By observing the results for the same value of th and across time scales we can identify the duration

above 50%, any other coarser time scale will provide us with similar results and thus is omitted.

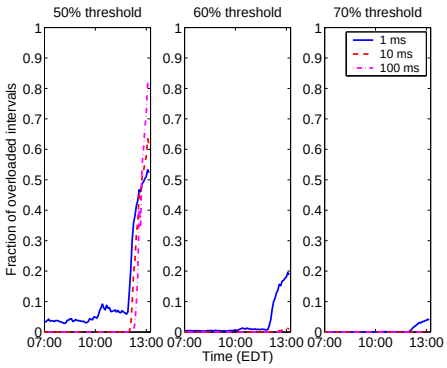

Figure 6: Fraction of overloaded intervals across time for different values of th and τ (OC-12).

and frequency of the episodes. For instance, if high values of threshold are accompanied by high fraction of overloaded intervals for the smallest time scale *alone*, we have an indication that the micro-congestion episodes do not last long.

Our proposed SNMP metric will be useful within an operational context if it is able to reveal congestion through time in a way that actually relates to degradation in performance. To address this specific issue we look into the delays experienced by packets for our two data sets during the entire duration of the trace. In Fig. 7 and Fig. 8 we present the delay statistics collected for each 5-minute interval in the trace for `set1` and `set2` respectively.

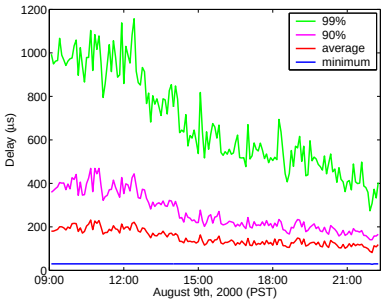

Figure 7: Delay statistics computed over 5 minute intervals across time (OC-3).

Fig. 7 shows that packets going through the monitored router may experience delays that reach 1 ms. Nevertheless this behavior is only observed for the first part of the trace until 3 pm. After 3pm the measured delays show a significant drop. From an operational standpoint it would be beneficial if one knew about the periods in time when packets experience larger delays and what their amplitudes are.

Returning to Fig. 5 we notice that until 3 pm all three time scales report utilization levels greater than 50%. In addition, until 3 pm 30% of the 1 ms measurements in each 5 minute interval exceed 70% utilization. Therefore, micro-congestion episodes for the first 6 hours in the trace are significant in terms of amplitude and persist in time so that they affect the time scale of 10 ms and 100 ms. Simple

observation of the collected statistics at the 100 ms time scale alone when $th = 50\%$ would lead to the same findings; micro-congestion in the first 6 hours that may lead to delays in the order of 1 ms according to Fig. 2. Nevertheless, collection of statistics at multiple time scales allows the network operator to investigate the reasons behind these episodes. If the reason is micro-congestion then utilization measurements at small time scales should exhibit significant fluctuations. Consequently, not all 1 ms intervals should report utilization levels greater than 50% and some small fraction of those should report significantly higher utilization levels, as is the case for `set1`. On the other hand, if 100% of the 1 ms intervals report utilization levels greater than 50% then there is an indication of persistent congestion that should be further researched.

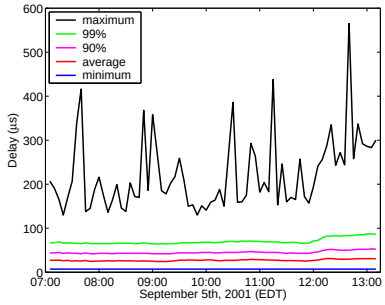

Figure 8: Delay statistics computed over 5 minute intervals across time (OC-12).

Our results for the OC-12 data set[3] are presented in Fig. 8. Delays experienced by packets through the monitored router are under 100 μs and increase toward the end of the trace. Increase in the single-hop delay occurs after 12 pm. From Fig. 6 the last 2 hours are characterized by 60% and 80% of the 10 ms and 100 ms intervals respectively exceeding 50% utilization. Approximately 50% of the 1 ms intervals exceed similar levels of activity. The amplitude of the delay can be inferred by observing that 20% of the 1 ms intervals exceed 60% utilization for the same period of time. Nevertheless, given that only the 1 ms timescale reports such high output link utilization values, we can conjecture that the actual delay values experienced by packets are rather limited (Fig. 3).

6. DISCUSSION/FUTURE WORK

Our methodology is not specific to link utilization measurements. Observing different performance metrics for different levels of activity (th) and across multiple time scales could have further applications. For instance, one could also collect counters for the number of τ-intervals within a 5-minute interval when output queues exceed specific threshold values. Observation of the collected results across time scales can also give insight into how long and how frequently queues reach specific occupancy rates.

From our results it can be seen that the most appropriate values of τ should be less than 100 ms. In [5] the authors showed that Internet traffic exhibits different scaling behavior for time scales below 100 ms and above 100 ms. At longer time scales traffic can be characterized as long-range dependent, while at shorter time scales it appears almost uncorrelated or lightly correlated. In future work we would like to investigate the relationship between the work presented in [5] and our work.

Identification of the factors affecting the time scale over which micro-congestion episodes manifest themselves clearly calls for additional information. We are currently in the process of installing monitoring equipment on all the active links of a router inside the operational Sprint IP backbone network. Extensive monitoring of an operational router will not only allow us to identify the factors affecting micro-congestion episodes, but also possibly quantify the effect that different factors have on the measured behavior.

Our ultimate goal is to make recommendations about traffic metrics that would be useful in an operational setting for the provisioning of IP networks. Such metrics, like the one presented in this work, would provide priceless insight into the traffic dynamics and could greatly improve current best practices in the area of link dimensioning and network provisioning based on SNMP measurements.

7. CONCLUSIONS

We presented an experimental analysis for the characterization of congestion episodes that occur at time scales much smaller than what is normally observed by operators. We showed that currently available SNMP metrics collected over intervals of 5 minutes mask out the existence of such short-lived episodes and thus are inappropriate for the link/queue dimensioning of IP links. We presented a methodology for the identification of micro-congestion episodes that is capable of characterizing them with respect to their amplitude, duration and frequency. Our approach relies on the analysis of link utilization at multiple time scales simultaneously. We showed that this type of analysis can provide great insight into the dynamics of the traffic at small time scales and their impact on the delay experienced by packets.

The time scales at which congestion occurs may in fact vary, but appear to be below 100 ms. We proposed performance metrics that can be collected over 5-minute intervals shedding light into the behavior of Internet traffic at these time scales. Our data sets here are fairly limited, and a more comprehensive analysis is needed. However, our preliminary investigation suggests that collection of such data would provide much greater insight into the dynamic behavior of IP networks than is currently available, as well as what could be obtained by observing performance at a single time scale.

8. REFERENCES

[1] C. Fraleigh et al. Provisioning IP Backbone Networks to Support Latency Sensitive Traffic. In *IEEE Infocom*, San Francisco, March 2003.

[2] K. Papagiannaki. *Provisioning IP Backbone Networks Based on Measurements*. PhD thesis, University College London, March 2003.

[3] K. Papagiannaki et al. Measurement and Analysis of Single-Hop Delay on an IP Backbone Network. In IEEE JSAC, vol. 21, no. 6, August 2003.

[4] M. Schoffstall et al. A Simple Network Management Protocol (SNMP). RFC 1157, May 1990.

[5] Z.-L. Zhang et al. Small-Time Scaling behaviors of internet backbone traffic: An Empirical Study. In *IEEE Infocom*, San Francisco, March 2003.

[3]In this figure we also display the maximum delay since it's not affected by router idiosyncrasies as shown in [2].