Operating a Network Link at 100%

Changhyun Lee[1], DK Lee[1], Yung Yi[2], and Sue Moon[1]

[1] Department of Computer Science, KAIST, South Korea
[2] Department of Electrical Engineering, KAIST, South Korea

Abstract. Internet speed at the edge is increasing fast with the spread of fiber-based broadband technology. The appearance of bandwidth-consuming applications, such as peer-to-peer file sharing and video streaming, has made traffic growth a serious concern like never before. Network operators fear congestion at their links and try to keep them underutilized while no concrete report exists about performance degradation at highly utilized links until today. In this paper, we reveal the degree of performance degradation at a 100% utilized link using the packet-level traces collected at our campus network link. The link has been fully utilized during the peak hours for more than three years. We have found that per-flow loss rate at our border router is surprisingly low, but $30 \sim 50$ msec delay is added. The increase in delay results in overall RTT increase and degrades user satisfaction for domestic web flows. Comparison of two busy traces shows that the same 100% utilization can result in different amount of performance loss according to the traffic conditions. This paper stands as a good reference to the network administrators facing future congestion in their networks.

1 Introduction

Video-driven emerging services, such as YouTube, IPTV, and other streaming media, are driving traffic growth in the Internet today. Explosive market expansion of smart phones is also adding much strain not only on the cellular network infrastructure but increasingly on the IP backbone networks. Such growth represents insatiable demand for bandwidth and some forecast IP traffic to grow four-fold from 2009 to 2014 [1]. Network service providers provision their networks and plan for future capacity based on such forecasts, but they cannot always succeed in avoiding occasional hot spots in their networks. However, traffic patterns in a network are usually confidential and few reports on hot spots are available to general public. Beheshti *et al.* report that one of the links in Level 3 Communications' operational backbone network was once utilized up to 96% [5]. A trans-Pacific link in Japan was fully utilized until 2006[1]. Choi *et al.* have reported on a link on the Sprint backbone operating above 80% and likely causing a few moments of congestion [8].

Korea Advanced Institute of Science and Technology connects its internal network to the Internet via multiple 1 Gbps links. One of them is to SK Broadband, one of the top three ISPs in Korea, and its link is the most utilized of all. The link to SK Broadband has experienced *persistent* congestion in the past few years. The measurement on

[1] The packet traces at Samplepoint-B from 2003/04 to 2004/10 and from 2005/09 to 2006/06 in the MAWI working group traffic archive at http://mawi.wide.ad.jp/mawi show full utilization.

our campus network tells us that the link has experienced 100% utilization during the peak hours for more than *three years*! To the best of our knowledge, our work is the first to investigate a 100% utilized link. Even at 100% utilization the link has no rate limiting or filtering turned on. However, the operational cost of a 1 Gbps dedicated link is typically in the order of thousands of US dollars a month and a capacity upgrade is not always easy. Also the empirical evidence demonstrates that persistent congestion, although itself pathological, does not always incur pathological performance–we still get by daily web chores over the congested link!

In this paper we report on the persistent congestion in our network and analyze its impact on end-to-end performance. The questions we raise are: (i) *how much performance degradation does the fully-utilized link bring?*; (ii) *how badly does it affect the end-to-end performance?*; and (iii) *how tolerable is the degraded performance?* Based on the passive measurements on our campus network link we present quantitative answers to the above three questions. Per-flow loss rate at our border router is surprisingly low, mostly under 0.1% even at 100% utilization, but $30 \sim 50$ ms delay is added. The increase in delay results in overall RTT increase and degrades user satisfaction for domestic web flows. Flows destined to countries outside China, Japan, and Korea suffer less for both web surfing and bulk file transfer, but they account for less than 5% of total traffic. Comparison of two busy traces shows that the same 100% utilization can result in different performance degradation according to the traffic conditions.

The remainder of this paper is structured as follows. Section 2 describes the measurement setup and Section 3 the traffic mix. In Section 4 we quantify the performance degradation in terms of loss and delay. In Section 5 we study the impact of increased delay and loss on the throughtput of web flows and bulk transfers. We present related work in Section 6 and conclude with future work in Section 7.

2 When and Where Do We See 100% Utilization?

Our campus network is connected to SK Broadband ISP with a 1 Gbps link, over which most daily traffic passes through to reach hosts outside KAIST. Figure 1 illustrates the campus network topology and the two packet capturing points, *Core* and *Border*. We have installed four Endace GIGEMONs equipped with DAG 4.3GE network monitoring cards [2] to capture packet-level traces to and from our campus network; each GIGEMON's clock is synchronized to the GPS signal.

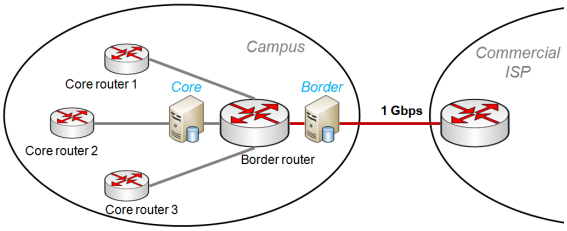
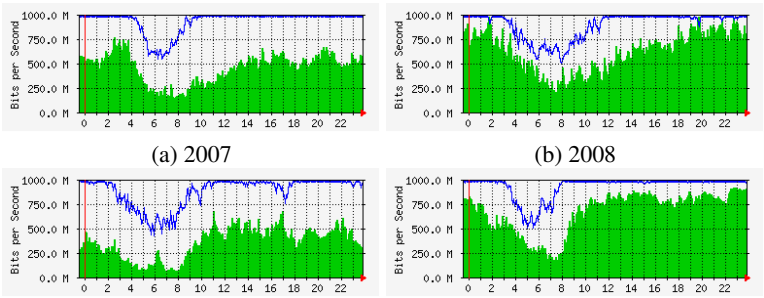


Fig. 1. Network topology on campus

The main observation, key to this work, is that the outgoing 1 Gbps link between the campus and the commercial ISP has been fully utilized during the peak hours for more than three years. The link utilization plotted by Multi Router Traffic Grapher (MRTG) on one day of July from 2007 to 2010 are in Figure 2. The solid lines and the colored region represent the utilizations of the uplink and the downlink, respectively. We see that the uplink lines stay at 100% most of the time. To the best of our knowledge, such long-lasting *persistent* congestion has never been reported in the literature.



(a) 2007 (b) 2008

(c) 2009 (d) 2010

Fig. 2. Link utilization of one day in July from 2007 to 2010; solid line is for uplink and colored region is for downlink. The time on x-axis is local time.

We have collected packet headers for one hour during the 100% utilized period from 2pm on March 24th (*trace-full1*) and September 8th in 2010 (*trace-full2*). We have also collected a one-hour long packet trace from 6am on August 31st in 2010 (*trace-dawn*) for comparison. As we see in Figure 2 the link utilization drops from 100% to around 60% during the few hours in the early morning. *Trace-dawn* has 65.6% of utilization and the number of flows is only half of those from full utilization. We summarize the trace-related details in Table 1.

Table 1. Details of packet traces

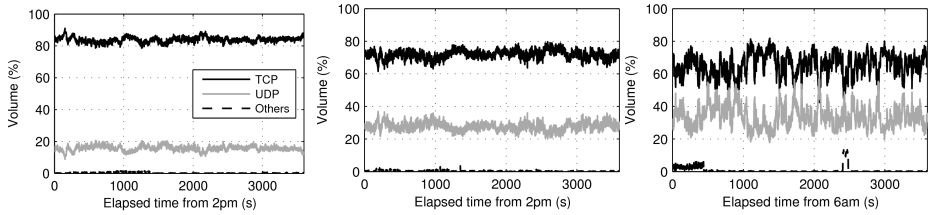
Trace name	Time of collection	Duration	Utilization	# of flows
trace-full1	2010/03/24 14:00	1 hour	100.0%	9,387,474
trace-full2	2010/09/08 14:00	1 hour	100.0%	9,687,043
trace-dawn	2010/08/31 06:00	1 hour	65.6%	4,391,860

The capturing point *Core* has generated two traces for each direction, and the point *Border* does the same; we have four packet traces in total for each collection period. In the following sections, we use different pairs of the four traces to analyze different performance metrics. For example, we exploit uplink traces from *Core* and *Border* to calculate the single-hop queueing delay and the single-hop packet loss rate. The uplink and downlink traces from *Core* are used to calculate flows' round trip times (RTTs). We monitored only one out of three core routers on campus, and thus only a part of the packets collected at *Border* are from *Core*. We note that, although incomplete, about

30% of traffic at *Border* comes from *Core*, and this is a significantly high sampling rate sufficient to represent the overall performance at *Border*.

3 Traffic Mix

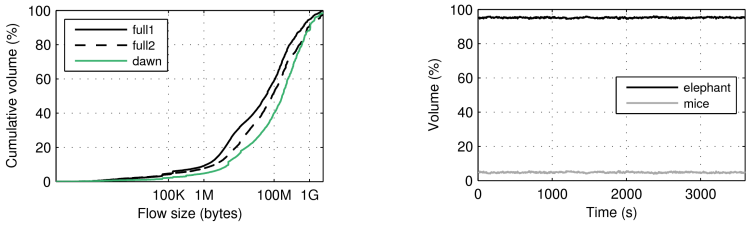
We first examine the traffic composition by the protocol in the collected traces. As shown in Figures 3(a) and (b), TCP traffic dominates when the 1 Gbps link is busy. The average percentages of TCP and UDP in *trace-full1* are 83.9% and 15.7%, respectively. The portion of UDP increases to 27.7% in *trace-full2* and 33.7% in *trace-dawn*. Although TCP is still larger in volume than UDP, the percentage of UDP is much larger than $2.0 \sim 8.5\%$ reported by previous work [7] [11]. We leave the detailed breakdown of UDP traffic as our future work. The dominance of TCP traffic indicates that most flows are responsive to congestion occurring in their paths.



(a) *trace-full1* (b) *trace-full2* (c) *trace-dawn*

Fig. 3. Protocol breakdown of the collected traces

In order to examine user-level performance later, we now group TCP packets into flows. Figure 4(a) shows the cumulative volume of flows. Flows larger than 100 KBytes take up 95.3% of the total volume in *trace-full1*, 95.8% in *trace-full2* and 97.2% in *trace-dawn*. We call those flows *elephant* flows and those smaller than 100 KBytes *mice* flows. In Figure 4(b) we plot the total volume in *trace-full1* contributed by elephant and mice flows in one second intervals and confirm that mice flows are evenly distributed over time. The other two traces exhibit the same pattern and we omit the graphs from them.



(a) Volume CDF versus flow size (b) Timeseries of elephant and mice flows

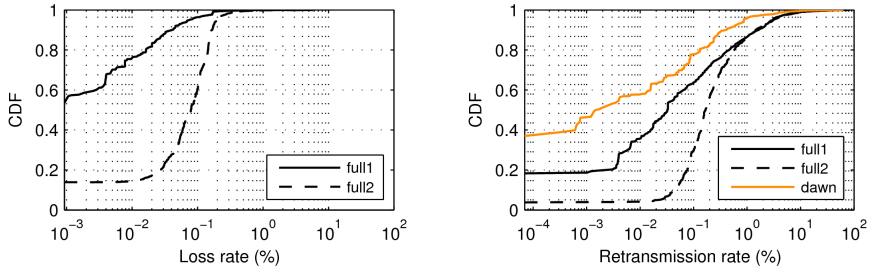
Fig. 4. Traffic volume by the flow size

4 Impact of Congestion on Packet Loss and Delay

In this section we explore the degree of degradation in single-hop and end-to-end performance brought on during the full utilization hours in comparison to the low utilization period. We begin with the analysis on loss and delay. In Section 3 we have observed that TCP flows, more specifically those larger than 100 KBytes, consume most of bandwidth. We thus focus on the delay and loss of elephant TCP flows in the remainder of this paper.

4.1 Packet loss

We examine the single-hop loss rate of the elephant TCP flows at our congested link. From the flows appearing both at *Core* and *Border*, we pick elephant TCP flows with SYN and FIN packets within the collection period. Existence of SYN packets improves the accuracy of RTT estimation, as we use the three-way handshake for our RTT estimation. For those flows we use IP and TCP headers of each packet collected at the capturing points *Core* and *Border* and detect loss, if any, through the border router as in [13].



(a) Single-hop loss rate (b) Estimated global loss rate

Fig. 5. Single-hop loss rate and estimated global loss rate (volume-weighted CDF)

Figure 5(a) shows the cumulative distribution of loss rates *weighted by the flow's size*: the cumulative distribution function on the y-axis represents the proportion in the total traffic volume as in Figure 4. Throughout this paper, we use this weighted CDF for most of the analysis so that we can capture the performance of elephant flows.

Because no loss is observed in *trace-dawn*, we do not show the loss rates in Figure 5(a). The maximum loss rate of flows reaches 5.77% for *trace-full1* and 5.71% for *trace-full2*. Flows taking up 53.1% of the total TCP traffic have experienced no loss during the collection period in *trace-full1*, whereas a much lower ratio of 13.8% in *trace-full2*. The performance degradation even at the same 100% utilization varies. *Trace-full1* and *trace-full2* differ mostly in the region between no loss and 1% loss. In the former 3.6% of traffic has loss rate greater than 0.1%, while in the latter the percentage rises to 39.5%. Apparently flows in *trace-full2* suffer higher loss. Here the utilization level alone is the sole indicator of performance degradation. In the future, we

plan to identify the main cause for such performance difference between the two fully utilized traces. Yet still 99.3% of *trace-full1* and 95.4% of *trace-full2* experience a loss rate less than 0.2%.

The full loss rate a flow experiences end-to-end is equal to or higher than what we measure at the border router. The loss rate in Figure 5(a) is the lower bound. It is not straightforward to measure the end-to-end loss rate for a TCP flow without direct access to both the source and the destination. Consider the following example. Let us consider a bundle of packets in flight en route to the destination. The first packet in the bundle is dropped at a hop and the second packet at a later hop. The sender may retransmit the entire bundle based on the detection of the first packet loss without the knowledge about the second packet loss. By monitoring the entire bundle being retransmitted at the hop of the first loss, one may not be able to tell if the second packet was dropped or not.

For us to examine the end-to-end loss performance we analyze the retransmission rate seen at the capturing point *Core*. A retransmission rate is calculated based on the number of duplicate TCP sequence numbers. There could be loss between the source and the border router, and the retransmission rate we observe is equal to or lower than what the source sees. However we expect the loss in the campus local area network to be extremely small and refer to the retransmission rate at the border router as end-to-end. We plot the retransmission rates for the three traces in Figure 5(b). We use logscale in the x-axis and cannot plot the case of 0% retransmitted packets. In case of *trace-dawn* 28.9% of traffic has no retransmission. In case of *trace-full1* and *trace-full2*, 18.3% and 3.8% of traffic has no retransmission, respectively. As in the case of single-hop loss, *trace-full2* has worse retransmission rates than *trace-full1*.

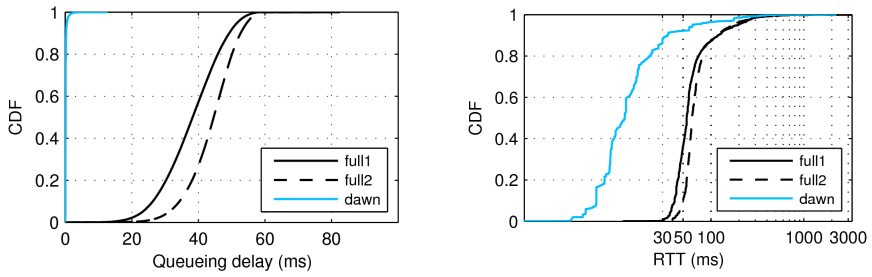
We count those flows that experience no loss at our border router, but have retransmitted packets. They account for 34.9% in *trace-full1* and 9.4% in *trace-full2* of total TCP traffic. For them the bottleneck exists at some other points in the network and our link is not their bottleneck. That is, even at 100% utilization our link is not always the bottleneck for all the flows.

Here we have shown the loss rates of only TCP flows, and we note that UDP flows in our traces can have higher loss rates than elephant TCP flows since the TCP congestion control algorithm reduces loss rates by throttling packet sending rates.

4.2 Delay

We now study delay, where we aim at examining the impact of the local delay added by our fully utilized link on the RTT of the whole path. To calculate the single-hop delay, we subtract the timestamp of each packet at the capturing point *Core* from the timestamp of the same packet captured at *Border*. We calculate the single-hop delay for each packet in the flows from each of the three traces and plot the distributions in Figure 6(a). *Trace-dawn* has almost no queueing delay at our border router. Note that the median queueing delay of *trace-full1* and *trace-full2* is 38.3 msec and 44.6 msec, respectively, and the delay variation is strong as most delays oscillate from 20 msec to 60 msec. Such high queueing delay badly affects user experience, which we will show in the next section.

To infer RTT of each flow from bi-directional packet traces collected in the middle of path, we adopt techniques by Jaiswal *et al.* [10]. Their tool keeps track of the



(a) Single-hop delay for each packet

(b) Round trip time (volume-weighted CDF)

Fig. 6. Single-hop delay and round trip time

TCP congestion window and gives RTT samples for each ack and data packet pair. Figure 6(b) shows the average per-flow RTT distribution weighted by the flow size. We note that the large queueing delay at the router adds significant delay to RTT for both *trace-full1* and *trace-full2*.

5 Impact of Congestion on Application Performance

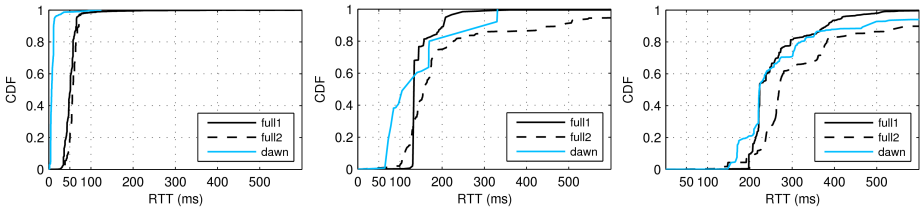
We have so far investigated the impact of the network congestion measured on our campus on the performance degradation in terms of per-flow end-to-end delays and packet losses. We now turn our attention to an application-specific view and examine the impact of the fully-utilized link on the user-perceived performance.

5.1 Web flows

In this subsection, we consider web flows and examine the variation in their RTTs caused by the 100% utilized link. As port-based classification of web traffic is known to be fairly accurate [12], we pick the flows whose TCP source port number is 80 and assume all the resulting flows are web flows. We then divide those flows into three geographic regional cases, domestic, China and Japan, and other countries. Each case includes the flows that have destination addresses located in the region. Our mapping of an IP address to a country is based on MaxMind's GeoIP [3].

In Figure 7, we plot RTT distributions of web flows for different network conditions. For all three regional cases, we observe that *trace-full1* and *trace-full2* have larger RTTs than *trace-dawn*. In section 4, we have observed that the median of the border router's single-hop delay at the border router is 38.3 msec in *trace-full1* (44.6 msec in *trace-full2*) when its link is fully utilized, and our observations in Figure 7 conform to such queuing delay increase.

In the domestic case, 92.2% of web flows experience RTTs less than 50 msec in the dawn, while only 36.2% (9.8% in *trace-full2*) have delays less than 50 msec during the fully utilized period. We observe similar trend in the case of China and Japan, but



(a) Domestic (b) China and Japan (c) Others

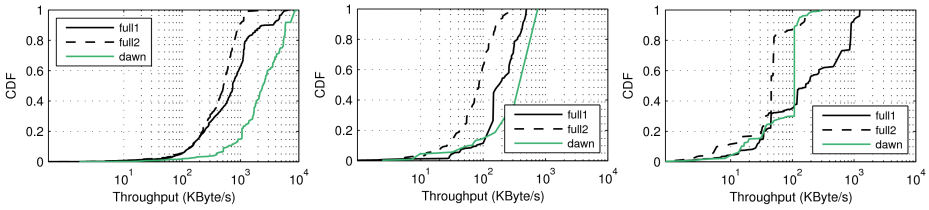
Fig. 7. RTT of domestic and foreign web flows for each trace (volume-weighted CDF)

the delay increase becomes less severe for the case of other countries. Most flows have RTTs larger than 100 msec regardless of the network condition.

Khirman *et al.*, have studied the effect of HTTP response time on users' cancelation decision of HTTP requests. They have reported that any additional improvement of response time in the $50 \sim 500$ msec range does not make much difference in user experience as the cancelation rate remains almost the same in that range; they have also found that additional delay improvement below 50 msec brings better user experience. According to these findings, our measurement shows that users in *trace-dawn* are more satisfied than those in the fully utilized traces when they connect to domestic Internet hosts. On the other hand, user experience for foreign flows stays similar for all the three traces because most RTTs fall between 50 msec and 500 msec regardless of the link utilization level.

5.2 Bulk transfer flows

We now examine the performance change of bulk transfer flows under full utilization. Bulk transfer flows may deliver high-definition pictures, videos, executables, etc. Different from the case of web flows for where we analyze the degradation in RTTs, we examine per-flow throughput that is a primary performance metric for the download completion time. We first identify bulk transfer flows as the flows larger than 1 MB from each trace and classify them into three geographic regional cases used in the web flow analysis. We summarize the results in Figure 8.



(a) Domestic (b) China and Japan (c) Others

Fig. 8. Throughput of domestic and foreign bulk transfers for each trace (volume-weighted CDF)

In the domestic case, 85.0% of bulk transfer flows have throughputs larger than 1 MByte/sec in *trace-dawn*. When the network is fully utilized, the performance degrades greatly, and only 36.6% (9.6% in *trace-full2*) of total volume have throughput larger than 1 MByte/sec in Figure 8(a). In Figure 8(c), the previous observation that *trace-dawn* has better throughput than the others disappears. We conjecture that our fully-loaded link has minor effect on the throughput of the overseas bulk transfers. There are other possible causes that limit a TCP flow's throughput (e.g, sender/receiver window, network congestion on other side) [16], and we plan to have the flows categorized according to each throughput-limiting factor in the future.

We are aware that comparing RTTs and throughputs from different traces may not be fair since source and destination hosts of flows can differ in each trace. We expect that the effect of the variation of hosts on campus should not be too serious because most hosts on campus are Windows-based and have the same 100 Mbps wired connection to the Internet.

6 Related Work

A few references exist that report on heavily utilized links in operational networks [5, 6, 8]. Link performance of varying utilization up to 100% has been studied in context of finding proper buffer size at routers. Most studies, however, have relied on simulation and testbed experiment results [4] [9] [14] [15]. Such experiments have limitations that the network scale and the generated traffic condition cannot be as same as the operational network. In our work, we report measurement results of 100% utilization at a real world network link with collected packet-level traces, so more detailed and accurate analysis are possible.

7 Conclusions

In this paper, we have revealed the degree of performance degradation at a 100% utilized link using the packet-level traces; Our link has been fully utilized during the peak hours for more than three years, and this paper is the first report on such *persistent* congestion. We have observed that 100% utilization at 1 Gbps link can make more than half of TCP volume in the link suffer from packet loss, but the loss rate is not as high as expected; 95% of total TCP volume have single-hop loss rate less than 0.2%. The median single-hop queueing delay has also increased to about 40 msec when the link is busy. Comparing *trace-full1* and *trace-full2*, we confirm that even the same 100% utilization can have quite different amount of performance degradation according to traffic conditions. We plan to explore the main cause of this difference in the future. On the other hand, fully utilized link significantly worsens user satisfaction with increased RTT for domestic web flows while foreign flows suffer less. Bulk file transfers also experience severe throughput degradation. This paper stands as a good reference to the network administrators facing future congestion in their networks.

We have two future research directions from the measurement results in this paper. First, we plan to apply the small buffer schemes [4] [9] [14] to our network link to see whether it still works on a 100% utilized link in the real world. Second, we plan to

develop a method to estimate bandwidth demand in a congested link. When network operators want to upgrade the capacity of their links, predicting the exact potential bandwidth of the current traffic is important to make an informed decision.

Acknowledgements. This work was supported by the IT R&D program of MKE/KEIT [KI001878, "CASFI : High-Precision Measurement and Analysis Research"] and Korea Research Council of Fundamental Science and Technology.

References

1. Cisco Visual Networking Index: Forecast and Methodology 2009-2014 (White paper), http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf
2. Endace, http://www.endace.com
3. Maxmind's geoip country database, http://www.maxmind.com/app/country
4. Appenzeller, G., Keslassy, I., McKeown, N.: Sizing Router Buffers. In: Proc. ACM SIGCOMM (2004)
5. Beheshti, N., Ganjali, Y., Ghobadi, M., McKeown, N., Salmon, G.: Experimental Study of Router Buffer Sizing. In: Proc. ACM SIGCOMM IMC (2008)
6. Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., Cho, K.: Seven Years and One Day: Sketching the Evolution of Internet Traffic. In: Proc. IEEE INFOCOM (2009)
7. Cho, K., Fukuda, K., Esaki, H., Kato, A.: Observing Slow Crustal Movement in Residential User Traffic. In: Proc. ACM CoNEXT (2008)
8. Choi, B., Moon, S., Zhang, Z., Papagiannaki, K., Diot, C.: Analysis of Point-to-Point Packet Delay in an Operational Network. Comput. Netw. 51, 3812–3827 (2007)
9. Dhamdhere, A., Jiang, H., Dovrolis, C.: Buffer Sizing for Congested Internet Links. In: Proc. IEEE INFOCOM (2005)
10. Jaiswal, S., Iannaccone, G., Diot, C., Kurose, J., Towsley, D.: Inferring TCP Connection Characteristics Through Passive Measurements. In: Proc. IEEE INFOCOM (2004)
11. John, W., Tafvelin, S.: Analysis of Internet Backbone Traffic and Header Anomalies Observed. In: Proc. ACM SIGCOMM IMC (2007)
12. Kim, H., Claffy, K., Fomenkov, M., Barman, D., Faloutsos, M., Lee, K.: Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices. In: Proc. ACM CoNEXT (2008)
13. Papagiannaki, K., Moon, S., Fraleigh, C., Thiran, P., Tobagi, F., Diot, C.: Analysis of Measured Single-Hop Delay from an Operational Backbone Network. In: Proc. IEEE INFOCOM (2002)
14. Prasad, R., Dovrolis, C., Thottan, M.: Router Buffer Sizing Revisited: the Role of the Output/Input Capacity Ratio. In: Proc. ACM CoNEXT (2007)
15. Sommers, J., Barford, P., Greenberg, A., Willinger, W.: An SLA Perspective on the Router Buffer Sizing Problem. SIGMETRICS Perform. Eval. Rev. 35, 40–51 (2008)
16. Zhang, Y., Breslau, L., Paxson, V., Shenker, S.: On the Characteristics and Origins of Internet Flow Rates. In: Proc. ACM SIGCOMM (2002)