DOE/MICS Progress Report

Project Title: *"Bandwidth Estimation: Measurement Methodologies and Applications"*

Project Website: http://www.caida.org/projects/bwest/

Check one: **[X] SciDAC Project**
[ ] MICS/Base Project

Lead PI: **K.C.Claffy** **Institution:** **CAIDA-SDSC-UCSD**
Co-PI: **C.Dovrolis** **Institution:** **University of Delaware**

Check one:

[ ]	1st Quarterly Report	Date:
[ ]	2nd Quarterly Report	Date:
[ ]	3rd Quarterly Report	Date:
[ ]	4th Quarterly Report	Date:
[X]	Annual Report	Date: June 10, 2002

PROJECT ABSTRACT

The ability for an application to adapt its behavior to changing network conditions depends on the underlying bandwidth estimation mechanism that the application or transport protocol uses. As such, accurate bandwidth estimation algorithms and tools can benefit a large class of data-intensive and distributed scientific applications. However, existing tools and methodologies for measuring network bandwidth metrics, (such as capacity, available bandwidth, and throughput) are mostly ineffective across real Internet infrastructures.

We propose to improve existing bandwidth estimation techniques and tools, and to test and integrate them into DOE and other network infrastructures. This effort will overcome limitations of existing algorithms whose estimates degrade as the distance from the probing host increases. Existing VPS (Variable Packet Size) and PTD (Packet Train Dispersion) probing techniques will be studied as well as novel algorithms and methodologies as they become available. As we improve algorithms, we will incorporate this knowledge into an integrated tool suite that offers insights into both hop-by-hop and end-to-end network characteristics. We will also investigate mechanisms for incorporating bandwidth measurement methodologies into applications or operating systems, so that the applications quickly reach the highest throughput a path can provide. Finally, we will examine ways in which routing protocols, traffic engineering, and network resource management systems can use accurate bandwidth estimation techniques in order to improve overall network efficiency.

Summary of accomplishments by individual institution or PI

Main accomplishments at CAIDA:

- Developed methodology and testbed capabilities for identifying the strengths and weaknesses of bandwidth estimation tools using both simulated and real high-speed (e.g., greater than OC3) traffic. CAIDA has been utilizing unique opportunities available through *SDSC's CalNGI Reference Test Lab*. Our test configuration can be totally isolated, or connected under controlled conditions to high speed networks (e.g., ESnet, Internet2, CalREN).

- Specified test scenarios, installed bandwidth estimation tools on end node test boxes, acquired and configured testlab resources for generating controlled loads of simulated traffic of different types. Collected and analyzed baseline data while learning specific impacts of different configuration setups.

- Analyzed traffic characteristics of several OC48 bidirectional traces, each having a duration of approximately one hour. Traces were recorded at a *Metromedia Fiber Network (MFN)* backbone in San Jose. We performed several analyses including summary statistics, distribution of traffic as stratified by application, details on fragmentation of packets and source/destination pairs for packets. Results such as these will be used to refine test scenarios.

- Sponsored a *Bandwidth Estimation Workshop* at CAIDA/SDSC, inviting DOE researchers working on bandwidth estimation testing to discuss 1) progress to date on bandwidth estimations (bwest) tools and their validation/evaluation, 2) bwest middleware strategies and infrastructure, and 3) data sharing/correlation options.

Main accomplishments at University of Delaware:

- Developed an original measurement methodology, called *Self-Loading Periodic Streams (SLoPS)*, for the estimation of *end-to-end available bandwidth*. To the extent of our knowledge, SLoPS is the first measurement methodology that can achieve this objective in a non-intrusive and accurate manner.

- Published a paper entitled *"End-to-end available bandwidth estimation: measurement methodology, dynamics, and relation with TCP throughput"* [1], at the **ACM SIGCOMM 2002** conference in Pittsburgh, PA. It is noted that the acceptance ratio at the SIGCOMM conference this year was 8.3%, the lowest ever in the history of the conference. The paper details the SLoPS methodology.

- Implemented a new bandwidth estimation tool, called **pathload**, based on the SLoPS methodology. *Pathload* has been evaluated by comparing its output with SNMP utilization data from the path routers. An alpha release of the code will be made in the summer of 2002. Pre-alpha versions of the code have been sent to selected groups (CAIDA, SLAC, and LANL).

- Published a paper entitled *"Pathload: A Measurement Tool for End-to-End Available Bandwidth"* [2], at the **Passive and Active Measurements (PAM) 2002** conference in Fort Collins, CO. The paper describes *pathload* in detail.

- Released a new version of our *end-to-end capacity* measurement tool **pathrate**. The latest version (2.2.1) is able to measure the capacity of high-bandwidth paths, up to the speed of an OC-12 link (622Mbps). We are currently working on extending the measurement scope of *pathrate* to the Gigabit range. *Pathrate* is based on the packet dispersion techniques that we studied in [3, 4].

- Investigated the effect of Layer-2 switches on per-hop capacity estimation tools, such as **pathchar**, **pchar**, and **clink**. We showed that Layer-2 switches can cause consistent underestimation errors in such tools. A paper describing this work has been submitted to the ACM Internet Measurement Workshop in May 2002 [5].

ACCOMPLISHMENTS BY INDIVIDUAL PI

Perform overall coordination and integration of project pieces (UCSD/CAIDA, Task-1): CAIDA PI kc claffy worked extensively with co-PI Constantinos Dovrolis and Program Manager Thomas Ndousse to plan project activities and interactions with related DOE efforts, especially with Guojun Jin (LBNL), Les Cottrell (SLAC), Robert Baraniuk (Rice University), and Robert Nowak (Rice University).

Establish and maintain database of actual bandwidth characteristics of measured links (UCSD/CAIDA, Task-2): Using its extensive list of contacts in the commercial ISP community, CAIDA began collecting physical capacity information for paths serving skitter monitor boxes in its macroscopic topology measurement project. Capacities and contact information are stored in a MySQL database.

Calibrate tools and techniques for bandwidth estimation of Internet paths, considering their ability to measure bidirectional stability, link capacity, link available bandwidth, and path available bandwidth (UCSD/CAIDA, Task-3): CAIDA and U-Delaware have jointly defined test scenarios for comparing and contrasting different bandwidth estimation tools under controlled conditions.

Initial attempts to baseline tool measurements against constant bitrate fixed size UDP packets yielded surprising results, where VPS based tools *pathchar, pchar, and pipechar* returned measurements of 23-33Mbps on a known 100Mbps link. We came to realize that a hidden Layer 2 switch existing as part of the Extreme Networks Summit 5i router was adding a hop to the end-to-end path, distorting its VPS based capacity calculation. Changing to a Juniper M20 router having no intrinsic Layer 2 switch improved the accuracy of the measurement, but only to approximately 50Mbps. Remaining inaccuracies in baseline measurements for VPS tools are consistent with limitations described in [5].

Baseline measurements of *pathrate* were also problematic: While pathrate exhibits greater than 95% accuracy at very small traffic loads, this accuracy quickly falls off for traffic loads greater than 40%. After discussions with pathrate developers, we determined that our baseline test scenario itself was providing unrealistic traffic that interfered with pathrate's measurement methodology. We intend to repeat tests with a more realistic mix of variable sized packet streams.

Initial testing of TCP throughput tool *iperf* yielded results near zero. Upon examination, we discovered that the end-node machines had been configured as half-duplex, while the next-hop switch

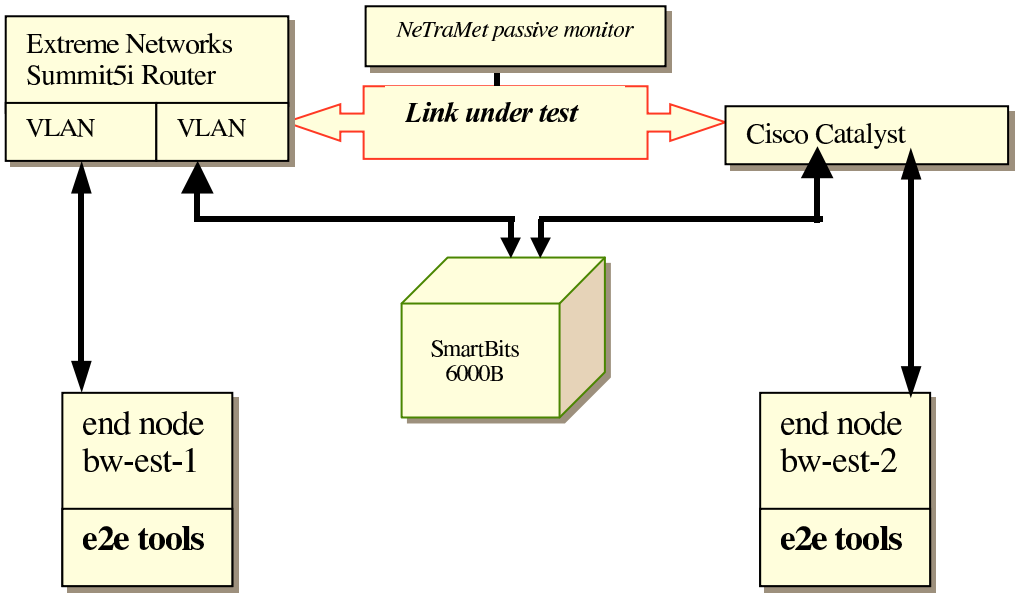

Figure 1: CalNGI Reference Test Lab. 2-hop configuration with SmartBits traffic generator.

was configured as full-duplex. After correcting the duplex mismatch, iperf available throughput measurements decreased linearly in direct proportion to an increase in generated cross-traffic.

Build a single-source testbed for implementing a bandwidth measurement methodology (UCSD/CAIDA, Task-4) and Begin to build an all-paths testbed for implementing a bandwidth measurement methodology (UCSD/CAIDA, Task-5): Figure 1 shows the initial hardware configuration on which we tested six bandwidth estimation tools: iperf, pathchar, pathrate, pchar, pipechar, and Sprobe. Boxes bw-est-1 and bw-est-2 are the end nodes on which bandwidth estimation tools are installed. Each end node has an IP address in a different subnet. The Extreme Networks Summit 5i router defines one end point of the link under test, and is configured with two VLANs, one for each subnet. For our 2-hop test scenario, a Cisco Catalyst 2900 switch establishes the other end of the link under traffic load. In future tests, a Juniper M20 router will be substituted for the switch, and multi-hop paths will be configured using available interfaces on both routers. The NeTraMet box, a Real-Time Flow Measurement (RTFM) meter, is placed to passively monitor the link under test. In this way, we can either verify simulated traffic generated by the SmartBits 6000B, or characterize real traffic occurring in the test environment.

Collaborate with U-Delaware group in the development and testing of pathload. CAIDA will be primarily responsible for porting the code to different OS platforms and optimizing the code (UCSD/CAIDA, Task-6): CAIDA has just begun to work with *pathload*, but expects to fully test this tool over the summer.

Create a front-end GUI for pathload. This interface will generate graphical views of available bandwidth in a path, similar to time-series graphs created by MRTG and RRDtool (UCSD/CAIDA, Task-7): CAIDA has begun to modify its *skping* RTT reporting

tool to provide graphic elements of a front-end GUI for any bandwidth estimation tool. (skping is associated with CAIDA's large macroscopic topology and performance analysis project based on deployment of *skitter* monitors.) In addition, a series of scripts developed to perform tests and collect data will contribute to the design of the user interface. Both of these efforts are expected to merge well with work done by U-Delaware to collect and present MRTG data from the tight link on the path in conjunction with pathload results.

Disseminate research results at relevant scientific conferences (UCSD/CAIDA, Task-8): By the end of the summer, CAIDA plans to submit papers to SIGCOMM or PAM 2003 conferences. One paper will detail our testing methodology. Another paper will present our analysis of bandwidth estimation tool testing data.

Development of available bandwidth estimation methodology (U-Delaware, Task-1): Researchers have been working on end-to-end measurement algorithms for available bandwidth over the last 15 years. From Keshav's *packet pair* [6] to Carter and Crovella's *cprobe* [7], the objective was to measure end-to-end available bandwidth *accurately*, *quickly*, and without affecting the traffic in the path, i.e., *non-intrusively*. Previously proposed measurement methodologies, as those published in [7, 8, 9], cannot measure end-to-end available bandwidth, either because they are based on packet-train dispersion (see results in [3]), or because they assume that a multi-hop path behaves as a single queue.

We have developed an original end-to-end available bandwidth measurement methodology, called *Self-Loading Periodic Slops* or *SLoPS*. To the extent of our knowledge, SLoPS is the first measurement methodology that can estimate the end-to-end available bandwidth of a path in a non-intrusive and accurate manner. The basic idea in SLoPS is that the one-way delays of a periodic packet stream show an increasing trend when the stream's rate is higher than the available bandwidth. This is illustrated in Figure 2. The sender transmits a stream of K packets of size L, at a constant rate R. If R is larger than the available bandwidth in the path A, the one-way delays D_i $(i = 1, \ldots K)$ of the stream packets are increasing, due to the growing queue size at the tight link of the path. Otherwise, the one-way delays do no show an increasing trend.

SLoPS is based on an iterative measurement algorithm, in which the probing rate gradually converges to the available bandwidth in the path. An important feature of SLoPS is that, instead of reporting a single figure for the average available bandwidth in a time interval $(t_0, t_0 + \Theta)$, *it estimates the range in which the available bandwidth varies in* $(t_0, t_0 + \Theta)$*, when it is measured with an averaging timescale* $\tau \ll \Theta$. The timescales τ and Θ are related to two key SLoPS parameters, namely the "stream duration" and the "fleet duration". SLoPS is described in detail in a ACM SIGCOMM 2002 publication [1].

Implementation of *Pathload* (U-Delaware, Task-2): The SLoPS measurement methodology has been implemented in a tool called *pathload*. *Pathload* is described in detail in a paper that has been published at the 2002 Passive and Active Measurements (PAM) workshop [2]. We verified *pathload* running it in several Internet paths across the US and Europe, and comparing its output with MRTG graphs from the path's tight link. The available bandwidth in those paths varies between 5Mbps and 90Mbps. We are currently working on testing *pathload* in higher-bandwidth paths, to the range of OC-12 and Gigabit Ethernet speeds. Each *pathload* measurement currently takes 5-10 seconds. We are also working on reducing this latency even further.

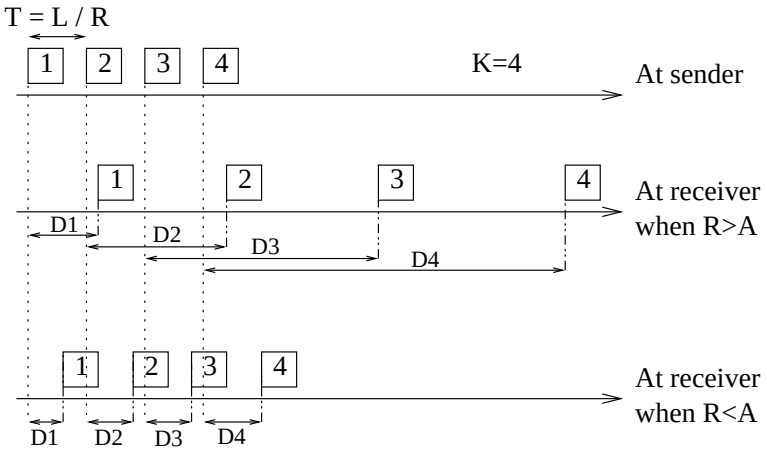

Figure 2: The basic idea in "Self-Loading Periodic Streams" (SLoPS) probing.

Pathload has been shown to be non-intrusive [1], meaning that its measurements do not cause increases in the network utilization, delays, or losses. We have used *pathload* to estimate the variability ('dynamics') of the available bandwidth in different paths and load conditions [1]. An important observation is that the available bandwidth variability increases as the utilization of the tight link increases. When that link operates in heavy-load conditions, the increase in the variability is dramatic.

An initial ("pre-alpha") release of the *pathload* source code was made in April 2002 to selected research groups (CAIDA, SLAC, and LANL). Based on feedback from those sites, we are currently working on fixing some bugs and improving the tool's output. A wide alpha release will be made in July 2002, announcing the tool at the IETF IPPM and the GGF Network Measurement working groups.

Pathload testing and verification (U-Delaware, Task-3): We have verified *pathload* experimentally, comparing its output with the available bandwidth given by the MRTG graph of the path's tight link. The MRTG graphs are drawn in real-time based on SNMP utilization data maintained by the path routers.

Figure 3 shows the MRTG and *pathload* results for 12 independent runs in a path from a Univ-Oregon host to a Univ-Delaware host.[1] An interesting point about this path is that the tight link is different than the narrow link. The former is a 155Mbps POS OC-3 link, while the latter is a 100Mbps Fast Ethernet interface. The MRTG readings are given as 6Mbps ranges, due to the limited resolution of the graphs. Note that the *pathload* estimate falls within the MRTG range in 10 out of the 12 runs, while the deviations are marginal in the two other runs. These are fairly typical results for the accuracy of the tool in all the paths that we have experimented with.

An important result about per-hop capacity estimation tools (U-Delaware, additional task): Tools such as *pathchar*, *clink*, and *pchar* attempt to measure the capacity of every Layer-3 (L3) hop in a network path. These tools use the same underlying measurement methodology, that

[1] More information about the location of the measurements hosts and the underlying routes is given in [2].

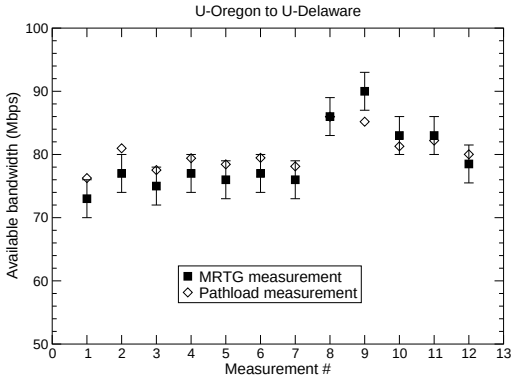

Figure 3: Accuracy of available bandwidth measurements using *pathload*.

we refer to as *Variable Packet Size (VPS) probing*. They key assumption in VPS is that each L3 hop along a path increases the delay of a packet by a 'transmission latency', which is the ratio of the packet size over that hop's capacity. Unfortunately, the capacity estimates that VPS tools produce are often wrong.

We have investigated the sources of these errors and showed that the presence of Layer-2 (L2) store-and-forward devices, such as switches and hubs, can have a detrimental effect on the accuracy of VPS tools. Specifically, an L2 device introduces additional transmission latency in a packet's delay, causing consistent underestimation in the corresponding L3 hop's capacity. We analyzed this negative effect, deriving the measured capacity of a hop as a function of the L2 link capacities in that hop. Experimental results in local, campus, and ISP networks verified our model, illustrating that L2 devices should be expected in networks of diverse scope and size. A paper describing this work has been submitted to the ACM Internet Measurement Workshop in May 2002 [5].

Tool	Capacity estimate
pathchar	49.0±1.5Mbps
clink	47.5±1.0Mbps
pchar	47.0±1.0Mbps
pipechar	93.5±3.0Mbps
pathrate	97.5±0.5Mbps
bprobe	95.5±2.0Mbps
Nominal capacity	100.0 Mbps

Table 1: Capacity estimates from *orion.pc.cis.udel.edu* to *sirius.pc.cis.udel.edu*.

To illustrate the problem that L2 switches cause to VPS tools, Table 1 shows the capacity estimates for a single-path from six different tools. The path connects two hosts that are located at a CS lab at Univ-Delaware through a HP4000 L2 Fast Ethernet switch. Both machines have Fast Ethernet network interfaces (100Mbps). Note that the VPS tools *pathchar*, *clink*, and *pchar* fail to measure the nominal capacity of this path (100Mbps), due to the presence of the Fast Ethernet L2 switch. The last three tools (*pipechar*, *pathrate*, and *bprobe*), on the other hand, are based on a different measurement methodology, namely the dispersion of packet pairs and trains, and they

provide a much more accurate capacity estimate for this path.

Future Challenges

One of the major tasks in the second year of this project will be *to tune and test our two bandwidth estimation tools,* **pathrate** *and* **pathload***, in Gigabit network paths.* Our objective is to demonstrate that the tools can accurately estimate up to 1Gbps by the end of 2002, and even higher bandwidth links (e.g., OC-48) in the later phases of the project. Even though the underlying measurement methodologies are expected to remain effective in that bandwidth range, we may need to deal with some hard practical issues, such as timestamping accuracy and resolution, operating system latencies, and batched interrupts.

A second major task in the second year of this project will be to *to incorporate our available bandwidth measurement methodology into transport protocols and applications.* We will also explore ways in which a bulk-transfer application can capture all the available bandwidth in a path. Possible strategies include the use of existing protocols in unconventional ways (e.g., multiple parallel TCP connections), modification of existing protocols (e.g., modify the TCP congestion avoidance algorithms for higher performance), and prototyping a UDP-based large file-transfer application. In pursuing these tasks, we will collaborate with Les Cottrell (SLAC) on the issue of parallel TCP connections, with the Net100 group in providing available bandwidth information to the TCP stack, and with the RADIAN group (LANL) in the design of innovative high-bandwidth transport protocols and techniques.

A third major task in the second year of this project will be *to make our two bandwidth estimation tools,* **pathrate** *and* **pathload***, user-friendly and accessible to users that do not necessarily have a networking background.* This task will be performed through a Web-based application that allows a user to monitor in real-time the capacity, available bandwidth, and other performance metrics of a path, such as round-trip delays, loss rate, and route stability. The resulting measurements will be visualized through an MRTG-like tool, and they will be archived for future processing through the MySQL database system. We are currently in the process of prototyping this tool, and we expect to complete it sometime during the second year of this project.

Finally, we hope to collaborate with DOE and ESnet personnel to design and implement bandwidth estimation tool validation tests appropriate for controlled deployment on high-speed production DOE networks. We are acutely aware of the sensitivity and costs associated with such testing. In particular, rigorous validation of any methodology will require read access to router SNMP counters for every hop along (both directions of) the path under test. We recognize the need to work within constraints imposed by DOE network managers and operators to minimize network performance impact while maximizing the utility and validity of the testing. The potential benefits of accurate bandwidth forecasting for the DOE user community demand efficient, yet rigorous testing.

Research Interactions

[1] - Interactions with other DOE network research projects

CAIDA has examined IEPM bandwidth data shared by Les Cottrell. We expect to continue to share data and would like to extend our collaboration include correlating our testing methodology, analysis and visualization techniques. CAIDA has also spent much time discussing potential collaboration activities with Guojun Jin (LBNL). We plan to evaluate Jin's new netest-2 tool as soon as

it becomes available.

CAIDA's work with SDSC's CalNGI Reference Test Lab will continue as it gives us an access point to high performance networks such as Internet2 and ESnet. We are looking for opportunities to conduct tool testing in environments of interest to the DOE community, and intend to contact Jim Leighton to discuss the potential options for testing tools on ESnet.

A *Bandwidth Estimation Workshop* was held at CAIDA/SDSC on June 5-7, 2002, inviting DOE researchers working on bandwidth estimation testing to discuss 1) progress to date on bandwidth estimations (bwest) tools and their validation/evaluation, 2) bwest middleware strategies and infrastructure, and 3) data sharing/correlation options. Visiting attendees included Les Cottrell (SLAC), Richard Hughes-Jones (University of Manchester and Grid Measurement Group), Guojun Jin (LBNL), Jiri Navratil (LBNL), and Martin Swaney (UCSB).

Over the last year, our U-Delaware group has collaborated with several other DOE researchers, including Les Cottrell (SLAC), Tom Dunigan (ORNL), Jin Guojun (LBNL), Wu-chen Feng (LANL), and Karsten Schwan (GATech). Specifically, we worked with Les Cottrell on improving *pathrate*, testing it over a large number of national and international paths, and comparing its results with other similar tools. Our collaboration with Tom Dunigan has also focused on *pathrate*. Tom has tested *pathrate* in several high-performance paths, limited by OC3 and OC12 links. These experiments are crucial for the verification and tuning of the tool, given that we do not currently have access to such high-performance links. Tom has also tested *pathrate* in paths that include traffic shaping devices. These collaborations will be greater in the second and third years of the project, when we pursue ways to apply bandwidth estimation in transport protocols and real-time applications.

[2] - Interactions with DOE application communities

Our interactions with the DOE application communities have been limited so far, given that we have been focused on developing and testing bandwidth estimation tools. Interactions with DOE application communities are likely to increase once we are able to test tools against real traffic on DOE networks. Initially, such interactions will serve to calibrate and tune tool efficacy. In later phases of this project, we will focus on providing users with an easy to use, graphical interface to bandwidth estimation tools and path performance monitoring.

References

[1] M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput," in *Proceedings of ACM SIGCOMM*, Aug. 2002.

[2] M. Jain and C. Dovrolis, "Pathload: A measurement tool for end-to-end available bandwidth," in *Proceedings of Passive and Active Measurements (PAM) Workshop*, Mar. 2002.

[3] C. Dovrolis, P. Ramanathan, and D. Moore, "What do Packet Dispersion Techniques Measure?," in *Proceedings of IEEE INFOCOM*, pp. 905–914, Apr. 2001.

[4] C. Dovrolis, P. Ramanathan, and D. Moore, "Packet Dispersion Techniques and Capacity Estimation," tech. rep., University of Delaware, June 2002. Submitted for publication to the IEEE/ACM Transactions on Networking.

[5] R. S. Prasad, C. Dovrolis, and B. A. Mah, "The Effect of Layer-2 Store-and-Forward Devices on Per-Hop Capacity Estimation," tech. rep., University of Delaware, May 2002. Submitted for publication at the 2nd ACM Internet Measurement Workshop (IMW).

[6] S. Keshav, "A Control-Theoretic Approach to Flow Control," in *Proceedings of ACM SIGCOMM*, Sept. 1991.

[7] R. L. Carter and M. E. Crovella, "Measuring Bottleneck Link Speed in Packet-Switched Networks," *Performance Evaluation*, vol. 27,28, pp. 297–318, 1996.

[8] G. Jin, G. Yang, B. Crowley, and D. Agarwal, "Network Characterization Service (NCS)," in *Proceedings of 10th IEEE Symposium on High Performance Distributed Computing*, Aug. 2001.

[9] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk, "Multifractal Cross-Traffic Estimation," in *Proceedings ITC Specialist Seminar on IP Traffic Measurement, Modeling, and Management*, Sept. 2000.

Appendix (Attachments)

- Publications

  1. M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput," in *Proceedings of ACM SIGCOMM,* Aug. 2002.
  2. M. Jain and C. Dovrolis, "Pathload: A measurement tool for end-to-end available bandwidth," in *Proceedings of Passive and Active Measurements (PAM) Workshop,* Mar. 2002.
  3. R.S. Prasad, C. Dovrolis, and B.A. Mah, "The Effect of Layer-2 Store-and-Forward Devices on Per-Hop Capacity Estimation," tech. rep., University of Delaware. May 2002. Submitted for publication at the 2nd ACM Internet Measurement Workshop (IMW).
  4. C. Dovrolis, P. Ramanathan, and D. Moore, "What do Packet Dispersion Techniques Measure?," in *Proceedings of IEEE INFOCOM.* pp. 905-914, Apr. 2001.

- Presentations

  1. kc claffy "Bandwidth estimation: measurement methodologies and applications" (DOE SciDAC PI Kickoff Meeting Jan 17, 2002) http://www.caida.org/projects/bwest/presentations/bwest_kickoff/.
  2. kc claffy "Internet measurement: state of DeUnion" (Tarafica Bandwidth Seminar, Hong Kong, Nov 01) http://www.caida.org/outreach/presentations/bw0111/.

- Awards

- Others

  1. Meeting Agenda: "Bandwidth Estimation PI Meeting at SDSC" June 5-7, 2002 http://www.caida.org
  2. Poster: "Bandwidth Estimation Methodologies and Applications" http://www.caida.org/projects/bwest/presentations/bwest_kickoff/kickoff_poster_large.gif