ABwE :A Practical Approach to Available Bandwidth Estimation

Jiri Navratil and R. Les. Cottrell

Stanford Linear Accelerator Center (SLAC), 2575 Sand Hill Road, Menlo Park, California 94025

***Abstract*—We report on a new tool developed for monitoring available bandwidth in the range from several Mbps to 1000 Mbps. The tool is based on the packet pair dispersion technique. It has been in experimental use at SLAC for several months and it has been demonstrated at iGrid2002 and SC2002. It can be used for detecting bandwidth changes caused by routing or congestions. We have studied the influence of cross traffic and the behavior of routers on the delay of the probing packets. The paper presents part of these analyses together with results from monitoring of 12 paths in US, Japan and in Europe.**

***Index Terms*—available bandwidth, dispersion technique, high-throughput, measurement tools, monitoring, networks, tuning, TCP.**

I. INTRODUCTION

The High Energy and Nuclear Physics (HENP) community is increasingly dependent on networking as the need for international cooperation grows in this field. The main research laboratories in the US and in Europe and many of the large US and European universities are currently connected with the links with 622 Mbps or more. The HENP physicists need to transfer huge amounts of data both between major sites such as SLAC in Menlo Park California, IN2P3 in Lyon France, and CERN in Geneva Switzerland, as well as between the major collaborating sites and home institutes spread over the world. Our main task is to provide the physicists with the infrastructure and reliable access to the network. An integral part of this is our monitoring activity that provides information about the current status of the network from SLAC to many places in the US, Japan and Europe.

We have several monitoring systems in operation. Some of these are based on active monitoring tools such as ping, iperf or bbcp. Others are monitoring network devices via SNMP tools and using passive monitoring of network flows. Each system has certain advantages and covers several parameters of interest. Network administrators and users need to know the Round Trip Time (RTT), losses, routing path, level of utilization of critical links and also the available bandwidth to our partners. Currently, we have such information in limited sampling periods. Missing was a tool that could be used in a continuous mode during 24 hours and could quickly detect and report unusual changes in the available bandwidth. Besides quickly detecting changes, such a tool must not be very intrusive on the network, since we want to use it on many paths, and must not overload the monitoring host's entry point. We concluded such demands can be met by a tool based on packet dispersion techniques. There are several tools using this technique [10], [13], [14], [17]. We evaluated several, but practically all of them failed on the high capacity paths (> 155 Mbps). Some of them failed methodologically (pathload), or they took too long (pathrate), some of them are just technically too complicated (pathchirp), and some of them are very inaccurate (pipechar).

Packet-pair dispersion techniques have been described in many papers [4], [5], [6], [10], [12], [13], [14], [17] and recently there are several groups that are doing active research in this field. The basic principal is simple: one sends packet probes to the final destination and measures the inter-packet delay between the packets as they arrive at the destination. Each group is trying to find ways to create improved configurations of probing packets to gain better information for the analysis. However, if the method is too complicated, it usually needs much time for evaluation. This is the case for the tools that use mathematical modeling. Some other methods are trying to fill the queue at the "potential bottleneck node" and find the parameters of this queue [8], [14]. They appear to have had good results on slow speed paths. However, according to our observations, it is very difficult to fill queue on the paths with a capacity of 622 Mbps or higher. And if successful, they would probably lose the non-intrusiveness advantage. For these reasons, we developed our own simple and effective tool that is also based on the dispersion technique.

II. BASIC PRINCIPALS

A. Characterization of traffic with respect to dispersion techniques

Our method is based on the simplest way of probing. We are using only packet pairs with a fixed size and initial delay between packets. We send several (typically 20) such closely spaced probes. The evaluation of the observed packet pair delays is based on detailed technical analysis of the problems that we can expect to meet in the routers and other network devices, and many experimental observations that we made during the development phase. We work as long as possible in the discrete domain of the measurements' delays, and attempt to classify the measurements according to the properties based on our observations. Later we can make final evaluations according to these classes. Such evaluations are only possible if we know the detailed characteristics and behavior of the paths and the character of the cross traffic on the path.

We understand a complete path as a cascade of queuing devices with FIFO strategy. Further, we have to assume that probing packets (PP) can be separated by cross traffic packets (CT) in any place on the path. Separation of the first probing packet (PP1) from the second probing packet (PP2) of a pair can happen even where there is no real bottleneck or congestion. The time delay (T_d) between PP1 and PP2 will grow discretely because it is caused by CT packets with particular lengths and finally it will contain the delay caused

by all the CT packets inserted between the PPs in any hops along the path. In many cases the process that causes T_d to increase is linear, but not always.

To understand what values T_d can achieve, we have to have more knowledge about the traffic statistics, especially the packet distributions. Several works in this field have been published [3], [7], [12] and all show very common features. However, each destination can have different packet distributions. Information on the packet distributions can be extracted and studied from the net flows which are provided by many routers. We used the Cisco Netflow data gathered at several sites (CERN, CESNET, SLAC) to obtain the packet distributions close to the end points. One of the parameters that we needed was the average packet length. For most flows that we have looked at, the average packet length is in the range 500–700 bytes but in some cases we have found values close to 1480 bytes. We have also found that the computed average packet length is very "fragile" because it is very often only a mean from the extreme values. Statistics from flows indicate that the traffic is dominated by very short flows (often with just one packet and this packet is usually very short). The fact that such a statistic has a very long tail can misjudge the role of short packets in the traffic spectrum.

Since we were not satisfied with the statistics from the flows, we studied the packet distribution analysis from tcpdump. The results from such studies were more plausible and useful for our methods. The distribution function of packet length shows two sets of maxima. There is a set of maxima (with values like 28, 34, 42 Bytes) in the region at the lower end of the possible packet size. In addition there is a second region with maxima at the higher end of the possible packet size range. The probability of short packets (< 100 bytes) is high but much lower than the probability of long packets. Also the variability of these regions is different. The range of packet sizes of the maxima for short packets is quite large and in the range 28 – 100 bytes while the location of the maxima for large packets is always very sharp (1470 -1485 bytes).

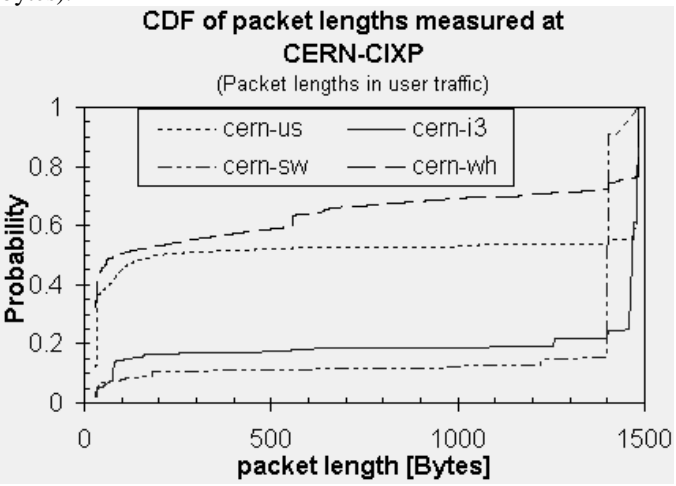

Figure 1: Packet Length distributions in CERN-CIXP

Fig. 1 shows the Cumulative Distribution Function (CDF) of packet lengths for several sites seen from the CERN-CIXP routing center (Geneva, Switzerland) which serves as a routing node for several other organizations in the region. The curves tell us the probability of the sizes of packets we can expect in the path. Fig. 1 shows that two of the four curves have very similar profiles. The curves (noted as CERN-i3 and CERN-sw) are profiles from CERN to IN2P3 and from CERN to SWITCH (Swiss scientific network) and via this into GEANT (European scientific networks). In these paths packets with lengths close to the Maximum Transfer Unit (MTU) of 1500 bytes dominate. These close-to-MTU sized packets represent a heavy load of file transfer type traffic. The curve marked as CERN-wh has a quite different distribution function. This is a path into an international organization that is also located in Geneva but doesn't belong to the HENP community. The curve is very similar to the profile which we found in the path to many universities and also corresponds with the results measured at the Internet exchange point [7]. The curve (noted as CERN-us) is for the path from CERN to the US. It is a mix of web style of traffic with heavy file transfer.

From the distribution of this function, we can see how big a role large packets play in the utilization of the path and the fact that short packets are not a very significant contributor to the overall traffic volume, even if their frequency is high. From this analysis we can presume that in most cases we will find more long packets than short packets in the queues of devices on the path. It also means that we can even expect that the delay T_d that we will measure is mostly caused by one or more long packets. We should also realize that one long packet will cause a delay that corresponds to 20–35 short packets and so it will be always more visible and detectable. According to our observations, the delay caused by short packets will in most cases create only a dispersion from the average value, that is itself defined by the transfer of long packets.

B. Paths characteristics and its influence on T_d

Most of today's network backbones are created in a similar fashion to those we are using (e.g. ESnet, Abilene, CalREN, GEANT). The capacity of these backbone links is usually much higher than the "last-mile" connection to the scientific laboratories or universities. Thus the bottlenecks are usually found at the site border connections, the LAN, or at the connection to the end-host itself.

The traffic on the back-bone links is composed of heavily aggregated traffic from many sources. The link utilization factor is thus relatively stable over the long term. However, the sources and relative contributions from the many flows changes instant to instant in any of the hops. Most of the CT may only share a common link with the PPs for a few steps of the total path. We believe that heavy aggregation would have a positive influence on the validity of the average packet length which will be probably very similar on many hops.

We don't know the utilization factor (ρ) of the interfaces in routing nodes. Lower values are not generally very interesting for our analysis since they will not result in congestion. To understand the effects of the router load on inter-packet delays, let us use as an example a utilization factor of $\rho=0.6$. From queuing theory for an M/M/1 type queue [2], [22], the average number of items (packets) in the queue $E(N)$ for this value of the utilization factor will be 1.5 packets. This means that PPs will sometimes arrive at an empty queue and sometimes at a

queue with one or more packets. The utilization factor also says that the probability that such a queue will be empty $p(0) = (1 - \rho) = 0.4$. PPs that arrive at an empty queue will experience no delay at this step. Generally $p(n) = \rho^n(1 - \rho)$. The probability that there will be at least one packet in the queue is $p(1) = 0.24$, and the probability that the PP will arrive at a queue with 2 packets in the queue is $p(2) = 0.144$, etc. This can also be interpreted as follows: 8 from 20 independent PP measurements can detect an arrival at an empty queue. This means that we would observe 8 similar inter-packet delays and 12 different inter-packet delays. Or more generally, when observing the frequency histograms of the PP inter-packet delays we will observe a peak identifying the inter-packet delay which belongs to ρ. If ρ is low, we will see only T_d which corresponds to the transfer of our PP. If ρ is high, new peaks will be formed since the inter-packet delay will be created by one, two or more CT packets inserted between our PPs. If this is true then we should find inter-packet delays that are very close to the time needed for data transfer of $E(N)$ packets (including PP). As expected, we found that T_d very often contains a certain repeating factor (2, 3, 4 etc.) due to the integer number of queued CT packets. And if we select our packet length close to the length of commonly used packets, the multiplicative factor will be more visible. Our results contain two interesting factors, the most probable bottleneck of the path (when ρ is small) and the level of queuing in other cases. We know that the long packets are more frequent and "visible", so finally we will see T_d corresponding to the queuing level of long packets. The short packets will cause only dispersion of T_d. It is clearly visible in the series of the independent measurements (see Figs. 3 and 4). We will refer to this multiplier value as a queuing delay factor (QDF). It is obvious that the frequency of QDF=1 is more frequent than QDF=2 etc. From large statistics *{more analysis and samples will be presented in full paper}* we can see that the QDF has an exponential character supporting this theory, see the peaks in Fig.2. Our results about T_d characteristics are in accord with the results from passive monitoring of packet inter-arrivals [11]. The QDF is the reason why they have seen the multiplicative bursts.

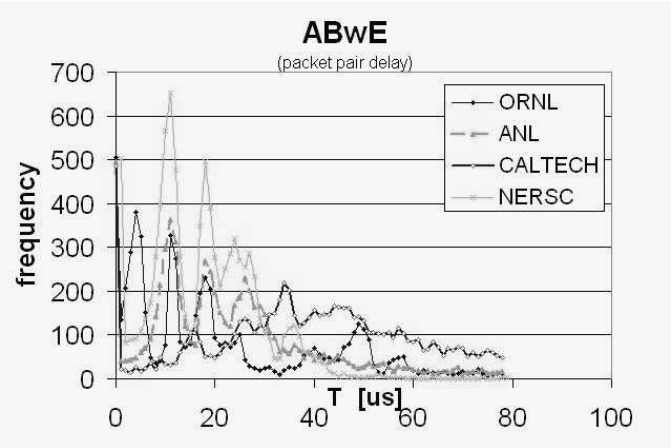

Figure 2: The QDF for 4 independent paths

III. SAMPLES OF PP DELAYS

We demonstrate the above conclusions for 2 different situations in graphs where the x-axis is the relative time (or the packet pair index) and the y-axis is the inter-packet delay. Fig. 4 shows T_d data from four measurement paths from SLAC to Europe. The bunch of data values (from the left) belong to a host at CERN, IN2P3, DL (Daresbury Lab., UK) and at CESNET (Prague). Hosts at CERN and in DL were located on 100Mbps LANs that comprised the narrow link in the path. The host in IN2P3 and CESNET had a 1000Mbps network interface. (The time to transmit our PP – 1450 bytes long at 100 Mbps is 116 μs, at 155 Mbps is 75 μs and for 622 Mbps is about 20 μs.)

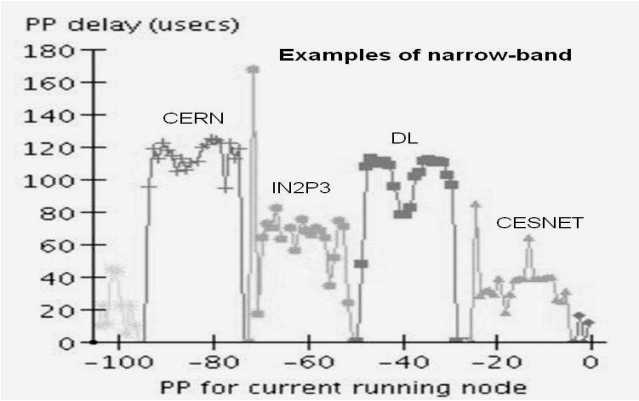

Figure 3: PP data -Narrow band limitations

From Fig.3 it can be seen that nearly all values for CERN and Daresbury are in the range between 110-120 μs. We can say that this was a potential bottleneck (or "narrow link" as defined by [13]) but both paths were relatively empty (our packets passed the paths without extra delay). The IN2P3 data illustrates a similar situation. The difference was in the place of the potential bottleneck. In case of IN2P3 there is a bottleneck caused by the 155Mbps line between CERN and IN2P3. The path to CESNET has no classical single narrow link. The delay for such a host can be limited in many places along the path by the utilization factor on the routers. The delays for CESNET in Fig. 3 are those observed across a path with multiple 622 Mbits/s links. These examples demonstrated that we can detect the potential bottlenecks and also determine their capacity.

Fig. 4 demonstrates a higher load on the path and illustrates the QDF. This is clearly visible on the last bunch of data which belongs to the host located in Canada (TRIUMF). The bottleneck is on the Canadian gateway to ESnet. From the figure you can see the groups of values belong to different QDFs.. In this figure are also data from previously described nodes (CERN, IN2P3 and DL) measured on a different day from Fig. 3. The scale of the y-axis is automatically selected according to the largest value, so the data for these hosts are not so readable but in the case of CERN the behavior is similar to that in Fig. 3. Data for the second bunch (IN2P3) alternate between 2 levels, one belonging to no queuing status and the second to strong queuing (QDF = 5).

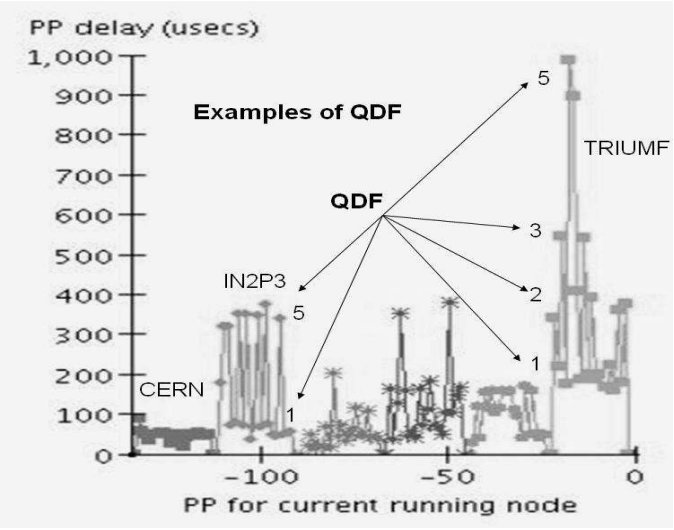

Figure 4: PP data sample with QDF

A. *Bandwidth Estimation*

In the final stage of the analysis we have to convert the T_d into bandwidth. The conversion from the time domain into bandwidth is a weak point of all dispersion techniques. In our methods, we use an empirically taken value for the CT packet length which could be expected on aggregated links. It gives us a value which we understand as a residual capacity of the path (sometimes referred to as the available bandwidth). To obtain the bandwidth which is achievable for TCP traffic we can normalize [23] our residual capacity with some TCP performance measurements (for example with Iperf).

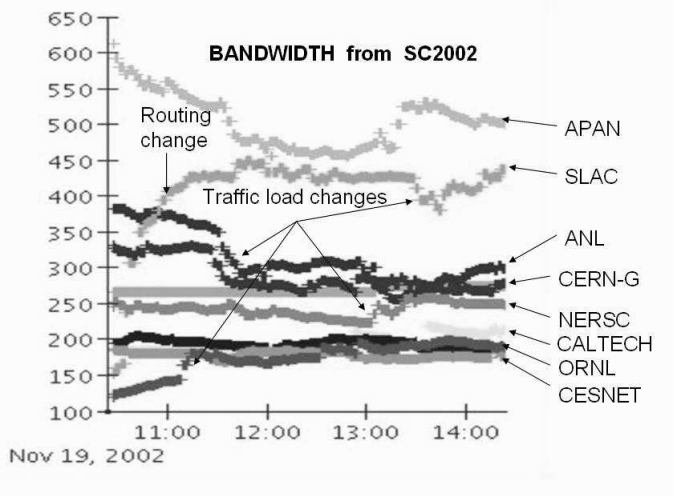

Figure 5: Bandwidth from SC2002 to "fast hosts" measured and visualized by ABwE.

We are running the ABwE monitoring at SLAC since August 2002 and demonstrated it at the iGrid2002 and SC2002 conferences. We usually work with two groups of remote hosts the "slow hosts" with throughputs up to 200 Mbits/s; and the "fast hosts" with speeds of hundreds of Mbits/s. Real-time time-series bandwidth plots for the groups are created using the UTH package [22]. Examples of a bandwidth time-series plot for the fast hosts are shown in Fig. 5. The bottom axis is the time in hours. The points associated with each host were distinguished by colors, and had a colored legend which would not reproduce well in black and white. So we have removed the legend and added labels to help the identification. The example in Fig.5 was measured from SC2002 in November 2002. The picture shows the long term monitoring capability of ABwE. During 4 hours we could see a dramatic change of bandwidth (150 to 450 Mbps) to SLAC caused by a routing change and several changes caused by heavy traffic. From previous examples and from Fig.5 it can be seen that ABwE can detect bandwidth in a wide range**.**

B. *Presentation of Results*

The ABwE tool is able to quickly show in real-time (within a couple of minutes) changes in the network performance (e.g. due to route changes or big changes in traffic and congestion). To make the graphs show trends more clearly, we displayed Exponentially Weighted Moving Averages (EWMA) [20] of the measurements, i.e. the current average avg_i is given by:

$avg_i = (1 - a) * y_i + a * avg_{i-1}$

where y_i is the current measurement, and a is a constant that is normally set to 0.9. This provides fairly heavy smoothing (Fig. 5). However, it has also the negative feature that changes would not be visible for several minutes (depending on the magnitude of the change). Therefore we are also using a version that presents both the actual value and the EWMA. This plot is noisier, but allows one to see changes more quickly. Fig. 6 shows an example of a sudden change in network performance (from 60 to 4 Mbps) for CERN displayed with and without EWMA. It can be seen that the sudden changes in throughput are seen within 2 minutes. The picture also shows the smoothing effect of EWMA with a strong a.

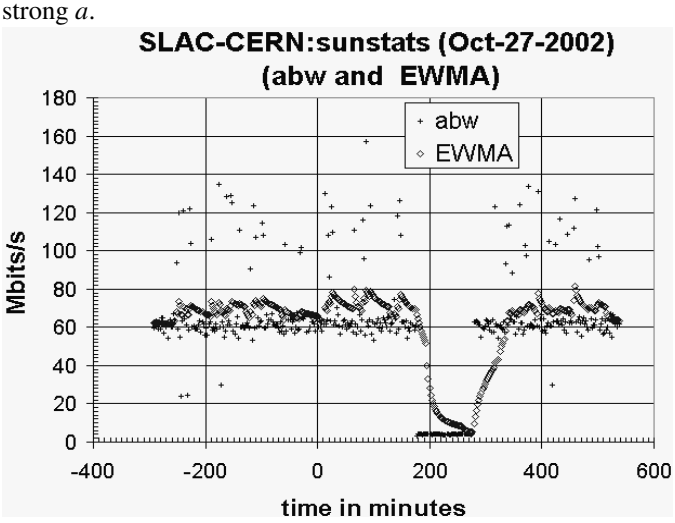

Figure 6: Unsmoothed ABWE (abw) measurements vs EWMA smoothed values.

Further investigation of this particular case showed that the first drop was caused by a change from the normal route from SLAC to CERN going via StarLight to a new route shared with commercial traffic going via BBNPlanet. The increase back to the normal bandwidth at 280 minutes was caused when the route to CERN was changed to going via GEANT.

C. *Iperf vs. ABWE*

We compared the SLAC iperf TCP throughputs averaged over 60 days from August 24 through October 26 2002 with the SLAC ABwE results. The correlation was strong (square of the correlation coefficient, $R^2 > 0.6$), see for example Fig.7. In Fig. 7 each point represents a single host with the x value being the 60 day average of the iperf TCP measurements, and the y value being the average ABwE for a four hour period on October 26, 2002 (the pattern of the ABwE values stays fairly consistent from weekday to weekday, so the actual choice of time period has little effect). The line is a least-squares fit to a straight line constrained to go through the origin. We also noted that if there were large (e.g. diurnal) variations in the iperf TCP measurements they also showed up in the ABwE measurements.

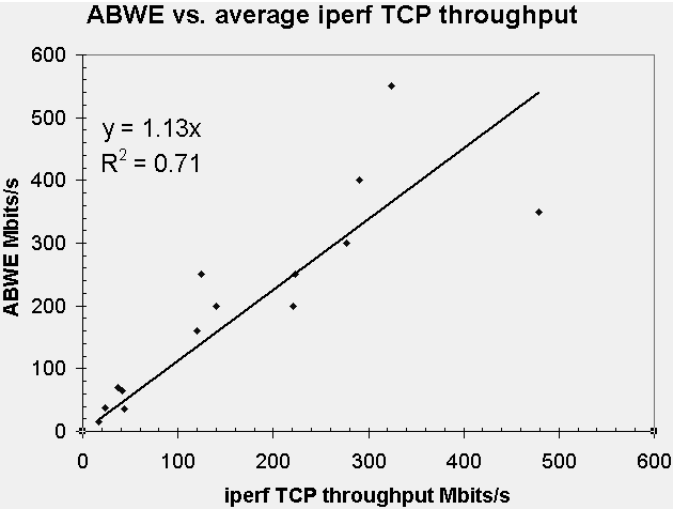

Figure 7: Correlation between ABWE measurements and the average iperf TCP throughputs.

This is encouraging, and we are studying it further. If it bears out, then we hope to be able to use the more heavyweight iperf TCP measurements to normalize the more frequent ABwE measurements. Besides providing low network impact (each bandwidth measurement needs only 40 packets/minute for one remote host) it is responsive to quick changes in performance, and is complementary to more heavyweight tools.

IV. Conclusions

We demonstrated a new tool for monitoring available path capacity in range from several Mbps to 1000 Mbps. The tool can be used in continuous mode and detects all substantial bandwidth changes caused by improper routing or by congestions. The usefulness of such a tool has been proven several times during last months when we discovered several dramatic routing changes which were corrected soon after we identified and reported them.

The mix of a low network intrusive packet pair bandwidth measurement tool (ABwE) with a more intrusive, user-centric TCP throughput tool (iperf) appears promising to provide low-impact short term (updates/minute) real-time measurements with good fidelity.

Acknowledgment

We gratefully acknowledge the assistance Tony Johnson for his UTH plotting program. We thank Olivier Martin and others at CERN for providing access to measurement hosts there, and all the administrators and contacts of the remote hosts for providing the hosts and responding to questions and requests.

References

[1] L.Kleinrock. Sequential processing machines (S.P.M.) analyzed with a queuing theory model. J. ACM 13,2 (Apr.1966), 179-193
[2] L.Kleinrock. Queuing theory – Vol.2 Wiley,1976
[3] V.Paxson and S.Floyd, "Wide-area traffic: the failure of Poisson modeling." In Proc of ACM SIGCOMM, London, September 1994, pp 257-268
[4] R.L. Carter and M.E.Crovella, "Measuring Bottleneck Link Speed in Packet-Switched Networks," Performance Evaluation, vol. 27,28, pp.297–318, 1996
[5] V. Jacobson, "pathchar: A Tool to Infer Characteristics of Internet Paths," ftp://ftp.ee.lbl.gov/pathchar/ Apr. 1997.
[6] V. Paxson, "End-to-End Internet Packet Dynamics," IEEE/ACM Transaction on Networking, vol. 7, no. 3, pp. 277–292, June 1999
[7] S. McCreary and K. C. Claffy, "Trends in Wide Area IP Traffic Patterns," Tech. Rep., CAIDA, Feb. 2000.
[8] B. Melander, M. Bjorkman, and P. Gunningberg. A new end-to-end probing and analysis method for estimating bandwidth bottlenecks. Global Internet Symposium, 2000
[9] Y.Zhang and L.Qui, "Understanding the end-to-end performance impact of RED in heterogeneous environment", Cornell CS Technical Report 2000-1802,July 2000
[10] V. J. Ribeiro, R. H. Riedi, M. S. Crouse, and R. G. Baraniuk Multiscale queuing analysis of long-range-dependent network traffic. Proc. IEEE INFOCOM, March 2000
[11] D. Katabi, CH.Blake "Inferring Congestion Sharing and Path Characteristics from Packet Interarrival Times" ,MIT-LCS-TR 828,Dec2001, Available at http://www.lcs.mit.edu/publications/pubs/pdf/MIT-LCS-TR-828.pdf
[12] A.Klemm, C.Lindemann, M.Lohman, "Traffic Modelling and Characterization for UMTS Networks". Proc. Of the Globecom, Internet Performance Symposium, San Antonio TX, November 2001
[13] C. Dovrolis, D. Moore, and P. Ramanathan. What do packet dispersion techniques measure? Proc. of INFOCOM, April 2001
[14] M. Jain and C. Dovrolis. Pathload: A measurement tool for end-to-end available bandwidth. In Proc. of the 3rd Passive and Active Measurements Workshop, March 2002
[15] M. Jain, C. Dovrolis, "End-to-End Available Bandwidth Methodology, Dynamics, and Relation with TCP Throughput". ACM SIGCOMM 2002, Pittsburgh PA, August 2002
[16] R. L. Cottrell. "Network Monitoring Tools". Available at http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2001/host-req.html
[17] Network Characterization Service (NCS). Available http://www-didc.lbl.gov/pipechar/
[18] Iperf. For more information, see http://dast.nlanr.net/Projects/Iperf/
[19] Netdyn: Network measurements tool. For more information, see http://www.cs.umd.edu/esuman/netdyn/
[20] "Dealing with the measurement noise-EWMA Filter" .Available at http://lorien.ncl.ac.uk/ming/filter/filewma.htm
[21] I.Adam and J.Resing "Stochastic Methods for Design and Planning", Dept. of Mathematics and Computer Science Eidhoven University of Technology, The Nederlands, see http://www.win.tue.nl/math/onderwijs/2V050/week2.pdf
[22] "UTH". Available at http://jas.freehep.org/documentation/UTH/
[23] M. Swany and R. Wolski, "Multivariate Resource Performance Forecasting in the Network Weather Service (2002)", SC2002, Nov. 2002.