A New Approach to Managing a Best Effort IP WAN Service

Jim Martin
Department of Computer Science
Clemson University
Box 340974
Clemson, SC 29634-0974

Abstract **--The network standards organizations have left the assessment component of network management to the network administrator. While there are good reasons for doing so, we feel that this has unnecessarily left a void between network service management and the end user. Our research is motivated by this position. We present a network monitoring and assessment algorithm that bridges the gap that exists between readily available network performance statistics and the subsequent assessment which must decide if the network is meeting user and organizational requirements. We focus on the network planning activities surrounding an organization's IP-based WAN (i.e., a best effort IP VPN or a private IP service). We propose a method that incorporates two fundamental concepts: first, the underlying performance metrics must be user oriented; second, the overall assessment of a WAN service is more meaningful if measured performance results drive an economic model that provides an estimate of the financial implications of network performance. While a contribution of this paper is to present our preliminary work on an 'organizationally oriented' method to manage an IP WAN service, the more important objective is to identify the issues surrounding the development and application of quality of experience assessment techniques to the network planning process.**

Keywords—network management, network provisioning, user oriented performance metrics

I. INTRODUCTION

Network management involves multiple activities including traffic engineering, network engineering, network monitoring and network planning. Traffic engineering attempts to better manage traffic on an existing infrastructure while network engineering establishes capacity in the form of link additions or updates as needed. Both of these activities require metric data that is generated by a monitoring process in order to make a decision (i.e., an assessment) as to whether the network is meeting requirements. Network planning requires an assessment of the current level of provisioning and must also take into account observed growth trends. Each of the network management frameworks specified by the IETF, the Open Group and the DMTF specifically leave the assessment decision to the organization [1,2,3].

Due to the high link speeds and the large number of relatively low bandwidth flows, planning high speed ISP backbones is manageable [4,5]. Growth is usually steady and predictable such that performance problems are rare. It is now taken for granted that most large service provider's networks are sufficiently (over)provisioned such that the bottlenecks are the low speed access link used by their customers [6]. With this motivation, we focus on the more challenging task of network planning of the corporate WAN where a low speed access link can be overwhelmed by traffic from a single user. The WAN represents a significant cost to an organization that includes monthly access charges as well as the cost of lost employee productivity if the WAN is not correctly provisioned. The required bandwidth of a WAN depends on many parameters (e.g., number of users, application characterization) but cost tends to limit the service to what is needed (i.e., over provisioning is not feasible).

Monitoring network performance at the ingress and egress points of a WAN service allows the corporation to validate that the WAN is correctly provisioned and also that the service is meeting stated performance objectives (i.e., negotiated SLAs). Fundamental to performance monitoring and assessment is the choice of performance metrics. A metric is defined as a quantity that is related to the performance and reliability of a network [1]. A metric requires a well defined measurement methodology that includes the relevant timescales. For network planning purposes, the relevant timescales for the metrics and subsequent assessment are on the order of 1 month. This reflects provisioning times associated with IP WAN services and also correctly reflects the amount of time required to obtain an accurate view of the workloads that operate over a WAN.

We are interested in validating the provisioning of end-to-end IP WAN services. Figure 1 illustrates a possible corporate network scenario. Branch offices interconnect with a corporate network via an IP VPN or a private IP best effort service. The service endpoints are defined by the routers located at the branch and corporate networks (e.g.., between routers R-1 and C-1). We assume that the provider's backbone and the corporate network access link (i.e., between C-1 and SP-1) are well provisioned such that the bottleneck for the service is the branch office access link. With this assumption, the simplest assessment algorithm based on access link utilization is:

Assessment algorithm 1: Once the average utilization of the access link exceeds 50% over a 30 day time period, the WAN is considered under-provisioned.

The obvious problem with this assessment is that it is not end-to-end. A second problem is that the large time scales make it difficult to accurately assess the impact of utilization on the end user experience and more importantly on the organization as a whole. For example, if the link is over utilized during business hours but under utilized during off hours, this assessment algorithm will not identify a problem. A network administrator might therefore monitor the utilization on an hourly basis during business hours, but then the problem becomes how to translate link utilization into an assessment of the end user quality of experience?

To more robustly assess WAN performance, active probing methods that monitor the end-to-end service can be employed. Assessment 1 can be extended with ping-based metrics as follows[1]:

Assessment Algorithm 2: The WAN is adequately provisioned as long as the average latency and loss over a 30 day period are less than 120 milliseconds and 1% respectively. In addition, the average link utilization must be less than 50% over the same 30 day period.

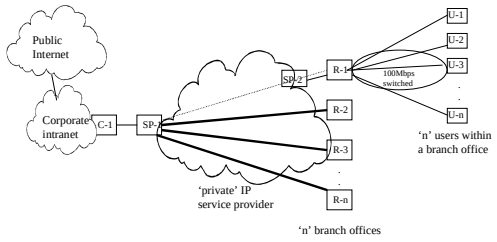

Figure 1. Example corporate network scenario

Primarily due to the large time scales, this form of assessment is 'service provider' oriented rather than user oriented and would not reliably validate provisioning decisions [7]. Many organization's apply their own monitor process and track ping latency and loss on an hourly basis. The difficulty is in correlating the observed statistics to the quality experienced by end users. For example, it is difficult to know the impact a network has on end user's network application sessions when the average ping RTT statistic reaches 120 milliseconds and or if the loss rate is .5 percent.

There are a number of network management products that allow an organization to monitor a network using application response time metrics. To the best of our knowledge, these products leave it up to the end user to craft specific metrics and subsequent assessment algorithms. Thus the burden of meaningful service level management falls onto the organization. The challenge associated with managing the WAN from a corporation's perspective lies in the gap that exists between readily available statistics and the subsequent network performance assessment that must determine the impact of WAN performance on end users and more importantly on the organization. Although ad hoc methods abound, there does not exist a standard and reliable method to assess the provisioning of an IP-based WAN service. We believe that there is the need to develop quality of experience guidelines for business applications (e.g., web applications) and to translate these guidelines into measurable network engineering criteria.

In this paper, we present a new approach to validating the provisioning of an IP based WAN service. We propose a method that incorporates two fundamental concepts. First, the underlying performance assessment should be based on metrics that assess the quality of experience associated with end users. Second, the overall assessment (i.e., the decision that confirms provisioning decisions) of a WAN is much more meaningful if the measured metric values drive an economic model which provides an estimate of the financial implications of observed network performance on the organization.

The remainder of this paper is organized as follows. First we present a web assessment algorithm. Then we present a comprehensive assessment algorithm that absorbs corporate policy into the network planning process. We describe the difficulties surrounding application level performance assessment. We conclude by identifying related work and future directions.

II. USER ORIENTED PERFORMANCE ASSESSMENT

In this section, we present a web oriented performance metric and assessment algorithm. Even though a more robust analysis and subsequent refinement of the algorithm is underway, we thought it useful to present our preliminary ideas and results.

A. The WRT metric

We define the Web Response Time (WRT) metric as follows. A client periodically issues an HTTP request for a web object from a web server. The client and server are positioned at the end points of the WAN or IP service that is to be monitored (i.e., client and server nodes located close to routers R-1 and C-1 respectively shown in Figure 1). A WRT 'sample' is the amount of time from when the client issues the request to when the entire web object has

[1] Assessment 2 is similar to the network performance component of a Service Level Agreement an ISP might provide for an IP VPN service [8].

been successfully received by the client. The client accumulates response time samples and periodically assesses network performance based on the web response time data.

It is important to make clear that the assessment is end-to-end with respect to the WAN service but not from the user's perspective. Using the example scenario portrayed in Figure 1, a corporate user located at a branch office is surfing the Internet and might experience congestion at some point over the path that is outside of the service provider's WAN network. This is outside the scope of the assessment. The intent is to establish a web-based probe process that can be used by an assessment algorithm to determine if network conditions are negatively impacting web application sessions that are being transported over the WAN. The web server that is utilized by the metric should run on an unloaded machine as delays created by the test machine will add error to the network assessment.

An application level performance metric offers several advantages over traditional ping based metrics. First, the WRT metric seamlessly incorporates the impact of loss and latency dynamics on the application into the performance assessment. Second, it is much more natural to translate WRT metric results into an assessment of how the network is impacting end users compared to an assessment based on loss or latency metrics. Studies have found that users become frustrated if web pages are not displayed within 11 seconds [9]. It has been suggested in [10] that a service (i.e., an application response time) must be predictable. The authors show that an end user that is conditioned to a certain level of performance is easily annoyed by even infrequent periodic lapses of quality. We conjecture that a quality of experience assessment must account for both the average performance as well as the variation in service levels.

We propose the following algorithm (referred to as the WRT assessment):[2]

> *Assessment algorithm 3: The link is considered correctly provisioned if the 95'th percentile of the most recent web response times (i.e.,samples no older than 2 hours) is less than 2 seconds. The web request specifies a single 19Kbyte object.*

The assessment algorithm has several parameters. The *tolerance* (95% by default) is intended to reflect that the network service is best effort IP and that web browsing (or an interactive web application) is an elastic application which does not have stringent service quality requirements. The metric *tolerance* is a convenient tuning knob for the algorithm. If the WAN service offers multiple performance levels (i.e., a differentiated service offering), multiple WRT management sessions can be active, each set with a different tolerance level (e.g., the higher priority traffic might have a tolerance of 99.5% while the best effort traffic might have a tolerance of 95%). The *timescale* parameter (2 hours by default) defines the time scale associated with the assessment. The intent is to match the assessment period to a reasonable duration of the human-to-computer interaction associated with the application. A time interval of 2 hours was chosen as this represents a reasonable web browsing duration. The *threshold* parameter is the metric threshold and directly correlates measured performance to the assessment. This parameter is the most troublesome to specify. By default it is 2 seconds, however increasing or lowering it will loosen or tighten performance objectives. Finally, the *size* parameter specifies the amount of data sent by the server to the client. The default size parameter is 19Kbytes which is a reasonable choice given that a single web object can range in size from 100 bytes to 100000 bytes (mean of 1500 bytes) and that the number of web objects per page can exceed 20 (mean of 9) [11,12][3].

The algorithm is designed based on the following rationale. Assume that a user is browsing the Web for a period of 2 hours. At the end of a 2 hour time interval, we would like to know if the user was dissatisfied with the browsing experience (and to what level). We partition time into 2 hour intervals. During an interval, we periodically obtain a web response time sample using a fixed sized web object defined by the transfer size parameter. If more than 5% of the samples exceed 2 seconds (in an interval), we assume that end users were negatively impacted by network performance.

Besides the web nature of the assessment, there are two other significant differences between it and assessment 2: first the time scale is very small; second there is no traffic constraint (i.e., a maximum utilization). In order for the assessment to help a network administrator validate that a WAN service is sufficiently provisioned, the time scale must match that of the capacity planning process. One possible modification to the assessment is to assume that a link is considered correctly provisioned only if a small number of time intervals (i.e., a threshold) do not meet the WRT assessment criteria over a 30 day period The exact number of time intervals that do not meet performance objectives should track the desired tolerance level. For example, a 95% tolerance translates to 18 time intervals (over a 30 day period) that can be in violation. A further

[2] The proposed algorithm is preliminary and is provided simply to facilitate the presentation of our ideas.

[3] This aspect of the metric is under evaluation. Many factors come into play when defining the response time component of a web metric including the size and number of web objects to pull, the use of multiple and/or persistent connections and TCP configuration parameters.

improvement presented in the next section incorporates additional organizational dimensions into the assessment.

III. ORGANIZATIONAL ASSESSMENT

The decision of whether an IP-based WAN service is correctly provisioned must include an assessment of the financial impact to the organization caused by poor WAN performance. There must be a utility function that maps value (or cost) to network performance. The utility function might depend on certain parameters allowing aspects of corporate policy to drive the provisioning decision. For example, the utility (or the value) of good network performance might vary over time (i.e., it might be more important to the organization that the WAN perform well during business hours rather than on the weekends). Further, the utility might vary base on the type of flow (i.e., good performance has more value to certain applications and users).

To generate the utility function we must extend the web assessment algorithm defined in the previous section to assess the *level* that network performance is impacting the end user rather than a binary assessment. We define a second threshold for the WRT metric, the *unavailability_threshold*, which reflects the point at which the assessment considers the WAN to be effectively unusable for web users. In order for the algorithm to seamlessly integrate network availability, we must also define a timeout period for WRT samples. If a web response probe exceeds 10 seconds, the algorithm assumes a timeout and inserts a sample value of 10 seconds into the results. A service provider generally provides an availability component to an SLA (e.g., 99% uptime as measured by a ping-based metric). In our method, network availability is defined by the *tolerance* and the *unavailability_threshold* parameters. For example, using a default tolerance of 95%, if the WRT metric samples are 10 seconds for more than roughly 6 minutes in any interval (i.e, 5% of 2 hours), the assessment considers the network to be unavailable for web browsing applications.

A *productivity_loss* function is needed to quantify the relationship between the performance metric statistics and the loss in productivity experienced by end users. We set the range of loss from 0 to 1. A value of 0 implies the network is not causing any problems, a value of 1 implies that the network impact is so severe that the employee might as well go home. The *productivity_loss* function requires an understanding of the threshold at which performance begins to impact users and the point at which the network is unusable. It is possible to generate a *productivity_loss* function based on any of the three assessment algorithms described previously. Figure 2-1 and 2-2 illustrate that it might be difficult to select meaningful threshold points for the *productivity_loss* functions based on utilization and latency. We focus on the WRT assessment algorithm as it offers the best chance of generating an accurate assessment over a range of network configurations and dynamics.

Based on preliminary analysis using informal surveys of web users under controlled settings, we set the WRT metric thresholds to 2 and 10 respectively. In other words, once the WRT metric exceeds a value of 2, we assume that the WAN service is negatively impacting end users. The level of degradation increases up to the 10 second threshold which represents the point that the network is unusable. Although the WRT assessment facilitates setting the assessment algorithm's threshold points, an organization will still have to select a specific *productivity_loss* function. One approach is to make the *productivity_loss* function increase linearly as the WRT metric increases from 2 seconds to 10 seconds (i.e., as illustrated in Figure 2-3). An alternative is to make the function increase exponentially, possibly approaching a maximum less than 1 (i.e., .5 as illustrated in Figure 2-4). The assumption behind the latter function is that as the WAN approaches the unavailable state, users shift to other work that does not require the network. Therefore, when the network is unavailable, user productivity is at most 50% reduced.

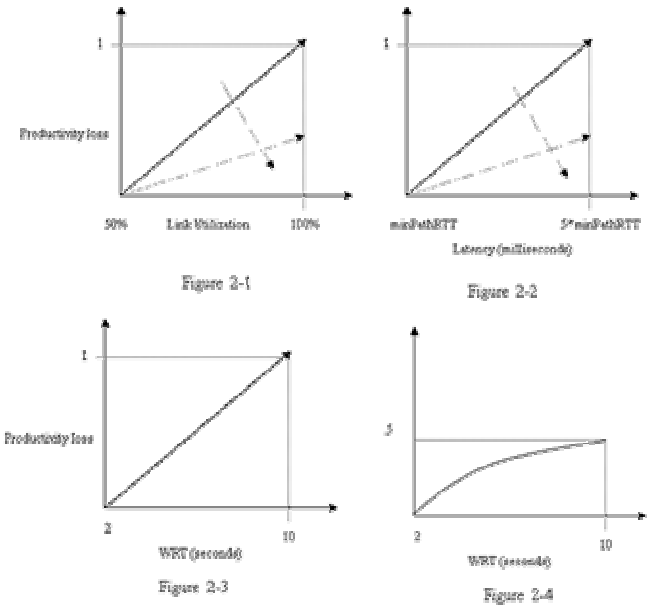

Figure 2. Productivity Loss Functions

The assessment is intended to process a large amount of WRT metric result data (e.g., 1 month of data). We divide the total assessment period into 2 hour intervals and compute the *productivity_loss* value for each. We use the loss function illustrated in Figure 2-3. The output of this process is an array (the *productivity_loss_array*) that is

indexed by the sequential numbering of time intervals (i.e, the first element represents the first 2 hours of the overall assessment period, the last element in the array represents the final 2 hour interval). Each array element also contains a *weight* that allows the organization to specify an indication of the importance of this time interval's *productivity_loss* value with respect to other intervals. If all time intervals are equally important, each array element's *weight* value is set to 1. If time intervals that occur outside of business hours are not important in the assessment, the weight for these intervals is set to 0.

A *cost_function* algorithm estimates the financial cost caused by poor WAN performance. The algorithm incorporates the simple linear *productivity_loss* function illustrated in Figure 2-3. The *cost_function* is defined as follows:

Cost = cost_function(productivity_loss_array, averageCostPerUser, averageNumberOfUsers)

In addition to the *productivity_loss_array*, the model requires the following parameters:

- *averageCostPerUser* : This is an estimate of the average cost per active user of the WAN. This cost can reflect the level of dependence the average active user has on the network by saying, for example, that the average cost will be 20% of the total cost associated with an employee.
- *averageNumberOfUsers*. This estimates the average number of active employees that utilize the WAN. As an example, for a given employee population of 1000, perhaps only 500 employees utilize the WAN (on average) during a given day.

We have implemented the method described and we have deployed the tool in a corporate network allowing us to test and validate the assessment algorithm. The corporate network is actually similar to the example network shown in Figure 1. We deployed a client at a branch that connects to the corporate network using a frame relay network service defined by a peak rate of 1 Mbps and a committed information rate of 256 kbps. Figure 3 illustrates the results observed between a client located at a corporate branch located in North Carolina interacting with a corporate server located at a data center in Connecticut. The purpose of the deployment was to demonstrate the operation of the tool. The method that we have described can be applied to any environment (i.e., Frame Relay, ATM or an IP transport service).

After the assessment, data is gathered from the *reportServer* and processed using a tool that visualizes the performance and that performs the cost based network assessment. The output from the analysis tool is shown in Figure 3. The results of the assessment are shown in the lower text box. The assessment period (31 days) was divided into time intervals of 2 hours. The WRT assessment indicates that 40 intervals did not meet minimum performance objectives. The cost model was instructed to use an *averageCostPerUser* of $150000 (per year) and an *averageNumberOfUsers* of 25.[4] The productivity loss model illustrated in Figure 2-3 was used.

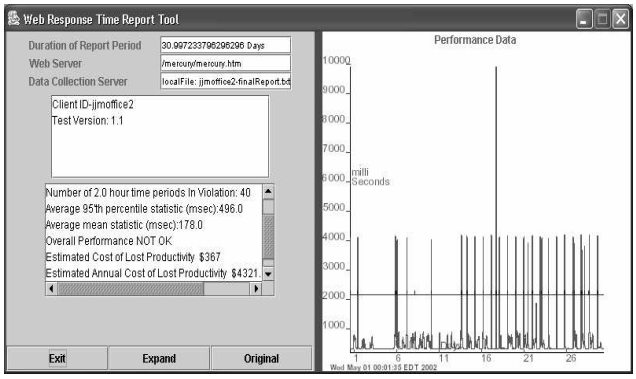

Figure 3: Tool Data Analysis Output

Expanding the results from Figure 3, Figure 4 illustrates the performance for 1 day. The graph on the right side of Figure 4 plots 4 curves (which are labeled): the WRT metric results, an average of the most recent web response time samples (the last 10 samples), and the moving window average of web response time samples (with a window of 2 hours). The horizontal line at *y=2* seconds indicates the *threshold* parameter setting. As we are still developing the assessment algorithm, we found it helpful to collect the two response time mean statistics in addition to the 95'th percentile.

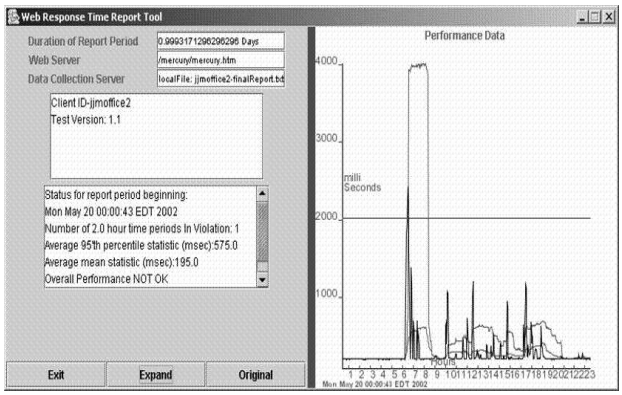

Figure 4: Expanded View of the Tool Data

Based on the results, we found that all of the 2 hour time intervals that were identified as being in violation of minimum performance requirements were caused by a scheduled data transfer that starts at 6:00AM and lasts until 6:30AM. For the entire assessment duration, the total cost due to lost productivity is $16000. If we focus on business

[4] The *averageNumberOfUsers* was estimated in this example as we did not have access to netflow data [13].

hours (9:00AM to 6:00PM), the cost of lost productivity drops significantly to about $4000 for an entire year which would not outweigh the cost of upgrading bandwidth. The results proved useful to the network administrators as they confirmed that the WAN is correctly provisioned.

IV. ASSESSMENT ISSUES

There are significant obstacles to application level network assessment. While the benefit is that application metrics are the best indicator of how complex network dynamics might impact the end user, if is difficult to find a metric that provides a consistent measure of the quality of experience across a variety of platforms and applications. One problem is that different browser sessions utilize different TCP stacks (i.e., TCP/Reno, TCP/Newreno) and different higher layer protocols. These differences can impact how network dynamics affects the end user quality of experience. For example, while most browsers and servers now support HTTP 1.1, they vary on their use of concurrent connections and pipelining [12]. Related to a browser's implementation of HTTP, browsers will also differ in their display approach. For example, one browser might begin to layout the GUI before all web objects have been pulled and another browser might wait until a larger number of objects have been downloaded.

A key contributor to end user perceived quality is the actual content of the web site. For a given network, one user might browse a site that has screens of text data while another user might browse a site with complex pages that include graphics and Java programs. If the WAN under observation shows signs of congestion, the first user might not notice the slowdown while the second user might become extremely annoyed. A further challenge is that the quality perceived by end users is subjective and consequently difficult to reliably quantify. For two web browser users experiencing the similar service levels (also assuming the same browser, content, etc.), there will be a range in the perceived quality. Although we have not proven this, we conjecture that the range of error associated with the WRT metric is much lower compared to an assessment based on one or more packet loss/latency metrics.

V. RELATED WORK

A fundamental premise behind our work is that first order performance metrics (i.e., packet latency and loss) are insufficient for assessing the impact of network performance on complex applications. Second order network dynamics are fundamental to end-to-end performance. It is well known that bursty loss patterns can significantly impact TCP's behavior [17,18]. Bursty loss patterns impact a multimedia forward error correction (FEC) algorithm's ability to recover from loss [19, 20]. Measurement studies confirm that packet loss over the Internet is correlated over time scales between 200-1000ms [21,22,23].

The use of application level metrics is not new. The IETF's Remote Network Monitoring (RMON) working group is working on extending current RMON capabilities to support application specific performance probes [24, 25]. While the exact methods are left to vendors, the focus is on passive monitoring rather than active probing techniques.

It has been established that the dynamics of loss and latency dominate the end user experience for voice and video applications [19]. The ITU has developed methods to solve the difficult problem of assessing the perceived quality of a voice call [14,15,16]. The Mean Opinion Score (MOS) provides the foundation for quantitative assessments of voice quality. Also of relevance to our work is the E-model which predicts the subjective quality of a telephone call based on its characterizing transmission parameters. Our work is similar in nature in that we drive a quality of experience model with web performance metrics to obtain an assessment of anticipated perceived quality. We then further refine the assessment by adding dimensions of corporate policy to validate the provisioning of IP WAN services.

The authors in [26] propose a model of TCP throughput based on easily measured metric loss and latency statistics. The goal is to convert network observations into user level performance metrics. They suggest that the bigger task for the future is to correlate the metric to application level performance. Although our approach estimates quality of experience based on an application metric, we share the same goals with those of [26].

There is a significant amount of research dealing with the provisioning of networks to meet statistical SLAs [27,28,29,30]. Most work in this space focuses on methods to provide levels of service and in general assumes admission control. Our assessment method can be the basis for SLAs in a network as long as traffic controls (e.g., a 50% link utilization constraint) are in place. Further, the assessment would be a natural fit in a measurement-based admission control scheme.

A 'hose' service interface for VPNs has been presented as a method to characterize aggregate traffic over a VPN pipe with performance guarantees [31]. The service provides a more efficient method to provision VPN's to meet performance objectives. The authors provide a technique to resize the pipes on demand based on online measurements. Our method is complementary to the hose concept and in fact could be an interesting addition to the realtime measurements that drive the resizing algorithm.

VI. CONCLUSIONS

The network standards community has left the assessment component of network management to the organization. While there are good reasons for doing so, we feel that this has left a gap between service management and the end user. With this motivation, we have proposed a method that directly connects the end user (i.e., the organization) to the provisioning of network services. The contributions of our work include: first the presentation of an assessment method that quantifies the impact of poor WAN performance on the end user's quality of experience; second a subsequent method to map the productivity loss associated with measured WAN performance to the financial cost incurred by the organization.

Although we have not robustly validated the accuracy of the web assessment algorithm, our work has helped us identify the challenges associated with application level assessment. The most basic issue is to determine the accuracy of a web assessment over a range of variables including widely varying end user performance expectations, different versions and configurations of TCP and different implementations and behaviors of the browser. In the future we plan to develop assessment algorithms that include other applications such as streaming and VoIP. We also would like to explore how application level assessment can be integrated into next generation network services which will seamlessly include appropriate service level agreements.

REFERENCES

[1] V. Paxson, et. Al., "Framework for IP Performance Metrics", RFC 2330, May 1998.

[2] The Open Group, "Application Response Measurement Issue 3.0 Java Binding", 2001.

[3] Distributed Management Task Force, "CIM Metrics White Paper", version 2.6, June 2002.

[4] C. Fraleigh, F. Tobagi, C. Diot, "Provisioning IP Backbone Networks to Support Latency Sensitive Traffic", IEEE INFOCOM 2003.

[5] K. Papagiannaki, et. Al., "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models", IEEE INFOCOM2003.

[6] N. Taft, et.al., "Understanding Traffic Dynamics at a Backbone POP", Proceedings of the Workshop on Scalability and Traffic Control in IP Networks, SPIE 2001, August 2001.

[7] J. Martin, A. Nilsson, "On Service Level Agreements for IP Networks", IEEE Infocom 2002, June 2002.

[8] WorldCom/Uunet's VPN Total Access Edition, http://www.worldcom.com/us/legal/sla/servicessupported/vpn.xml.

[9] N. Bhatti, et. Al., "Integrating User-Perceived Quality into Web Server Design", Ninth International WWW Conference, May 2000. http://www9.org/w9cdrom/92/92.html.

[10] A. Bouch, A. Kuchinsky, N. Bhatti, "Quality is in the eye of the beholder: Meeting Users' Requirements for Internet Quality", Proceedings of the 2000 Conference on Human Factors in Computing Systems (CHI-00), April 2000.

[11] P. Barford, M. Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation", ACM SIGMETRICS, 1997.

[12] F. Smith, F. Campos, K. Jeffay, D. Ott, "What TCP/IP Protocol Headers Can Tell Us About the Web", ACM SIGMETRICS, June 2001.

[13] Cisco Corporation, Netflow Services and Applications (whitepaper), August 1999.

[14] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality", August 1996.

[15] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", February, 2001.

[16] ITU-T Recommendation G.107, "The E-model, a Computational Model for Use in Transmission Planning", December, 1998.

[17] M. Allman, S. Floyd, "Enhancing TCP's Loss Recovery Using Limited Transmit", Internet Draft, August, 2000, file: draft-ietf-tsvwg-limited-xmit-00.txt".

[18] S. Floyd, et. Al., "Equation Based Congestion Control for Unicast Applications", SIGCOMM2000.

[19] W. Jian, H. Schulzrinne, "Modeling of Packet Loss and Delay and Their Effect on Real-Time Mutlimedia Service Quality", NOSSDAV 2000, June 2000.

[20] S. Varadarajan, et. Al., "Error Spreading: A Perception-Driven Approach to Handling Error in Continuous Media Streaming",

[21] J. Bolot, "End-to-end Packet Delay and Loss Behavior in the Internet", ACM SIGCOMM93.

[22] V. Paxson, "End-to-end Internet Packet Dynamics", IEEE/ACM Transactions on Networking, June 1997.

[23] M. Yajnik, S. Moon, J. Kurose, D. Towsley, "Measurement and Modeling of the Temporal Dependence in Packet Loss", INFOCOM99, March 1999.

[24] R Dietz, "Transport Performance Metrics MIB", IETF draft: <draft-ietf-rmonmib-tpm-mib-07.txt>, August, 2002.

[25] S. Waldbusser, "Application Performance Measurement MIB", IETF draft: draft-ietf-rmonmib-apm-mib-07.txt, April, 2002.

[26] M. Goyal, R. Guerin, R. Rajan, "Predicting TCP Throughput From Non-invasive Network Sampling", IEEE Infocom 2002.

[27] J. Heinanen, Et Al., "Assured Forwarding PHB Group", RFC 2597, June 1999.

[28], R. Gibbens, "An Approach To Service Level Agreements with Differentiated Services", Philosophical Transactions of the Royal Society, Mathematical, Physical and Engineering Sciences, August 2000.

[29] E. Knightly, N. Shroff, "Admission Control for Statistical QoS: Theory and Practice", IEEE Network, March 1999.

[30] K. Nichols, V. Jacobson, L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet", April 1999.

[31] N. Duffield, et. Al., "Resource Management with Hoses: Point-to-Cloud Services for Virtual Private Networks", IEEE/ACM Transactions on Networking, Oct 2002.