Estimating User-Perceived Web QoS by Means of Passive Measurement

Abstract: To enforce SLA management, SLA metrics and quality assessment method for network services or application services must be established in advance. In this paper two metrics for Web application service intrinsic performance, average round-trip time and delivery speed are defined taking both implementation mechanism of Web applications and user access behavior into account. In three cases, Web traffic trace for a specific Web site was recorded at client-side to calculate above two metrics; meanwhile the user-perceived service quality of the Web site evaluated subjectively was taken down. Linear regress analysis of these observation data indicates that user-perceived quality of Web application service can be objectively estimated by a linear regression function that uses two metrics as independent variables.

Keywords: Service Level Agreement (SLA); end-to-end SLA; Web application service performance metrics; round-trip time; delivery speed; Perceived Service Quality

1 INTRODUCTION

Rapid development of Internet technology and application makes service management for the Internet increasingly critical. The best effort service mode of the Internet has been gradually replaced by the differentiate service mode. At present, some Service Providers (SP) have commenced providing stated services to their customers against a contract called Service Level Agreement (SLA) that specifies measurable quality parameters. If the practical service provided cannot satisfy requirements of the contract, the SP will compensate according to the agreement. As a formal negotiated agreement, a SLA should contain a number of objective and measurable parameters that the SP guarantees to their customers. Values of these parameters may be reported to the customer as proof that the SP has met their commitments. SLAs can cover many aspects of the relationship between the Customer and the SP, such as performance of services, customer care, billing, service provisioning, etc [1,2]. Two parties of a SLA may be an end user and a SP, or a SP and another SP. An end user may be a single user or an organization. A service provider may be Network Operator, Internet Service Provider (ISP), Application Service Provider (ASP), etc. A SLA can assist SPs to improve customer relations and provides a vehicle for potential differentiation from their competitors. Although researches on SLA for IP services have made some progress, especially in SLA for network services, there still exist many issues without common understandings, such as the definition of end-to-end performance metrics for end-to-end SLA between an end user and an ASP (called as EASLA).

According to quality management science, Total Quality Management covers some basic viewpoints such as “Pursuit of quality is of primacy”, “Customers are paramount to all others”, and “Show with relevant data”[3]. “Pursuit of quality is of primacy” is not meant to pursuit the highest quality, however, it means that quality should be combined with cost, and distinct level quality should be delivered to users with diverse requirements. The second viewpoint means that quality delivered should conform to user’s requirements for QoS for the sake of user satisfaction. The third viewpoint means that besides qualitative analysis, quantitative analysis should be used whenever possible to avoid subjectivity. For EASLA, user satisfaction derives from the comparison between user-perceived QoS and user’s requirements (i.e. selected service level). Therefore, service levels should be related to user’s perception, otherwise they have no sense and it is impossible to make users satisfied. Although mainly determined by service performance, user-perceived QoS is subjective and deniable. Therefore, to objectively, quantitatively estimate perceived service quality is crucial for establishment of rational service levels, for EASLA compliance verification, or for judgment of the degree of user satisfaction. Thus we must attempt to define applicable metrics to measure perceived service quality based on Web traffic collected passively. Service performance consists of service result (technical quality) and service procedure (functional quality) [4]. With regard to ordinary Web application services that transfer non-stream data such as text, image and simple animation from Web servers to Web clients, technical quality can be divided into information quality, which is out of the range of EASLA, and information delivery quality. Functional quality, service- and technology-independent, relates to organization efficiency that can be omitted for Web application service.

The intentions of this paper are to find service-intrinsic performance metrics that can be used to estimate perceived quality of Web application service by user, and how to use them to estimate. Metrics defined in this paper can be used as SLA metrics for such EASLA in which the ASP is a Web ASP, and

the end user is a single user. If the end user is an organization, these metrics should be updated slightly. The rest of the paper is organized as follows. Section 2, gives characteristics of Web application service. We define two Web service-intrinsic performance metrics in section 3. In section 4, through passive measurements of Web traffic traces in three cases, we have demonstrated that user-perceived quality of Web application service can be estimated by above two metrics in a linear regression function. We finally provide summary and conclusion in section 5.

2 CHARACTERISTICS OF WEB APPLICATION SERVICE

2.1 Information Exchange

Ordinary Web applications mainly provide all sorts of information, and generate elastic traffic. There are also a few Web servers that provide live audio or video, and generate stream traffic. We only consider the former. A user must first provide a Web address (URL) by clicking a hyperlink, a favorite, or a button, or by keyboard input, in order to access Web resource. Then, the browser will communicate with relevant Web servers including the Web server S_0 given in the URL and other ones involved, to get Web pages direct or indirect specified by the URL. We call above procedure “an access” that will allow us to speak clearly below.

During an access, if a Web address contains host name, the browser will ask a name resolver process to look up the requested host name. The browser, after receiving a positive DNS reply, sets up a TCP connection with the Web server through "three-way handshake", during which two parties notify own TCP initial sequence number and receiver maximum segment size. On the first TCP connection, the Web browser sends the first HTTP request (Get or Post method), in response to which S_0 answers with the HTML document H_0 of the first Web page. Besides text information, H_0 specifies the layout of the first Web page, i.e., how other elements (that may be stored in other Web servers) such as images, flashes and even other HTML documents are referenced. As H_0 encapsulated in TCP segments can only be received gradually, the browser knows gradually what else elements exist and then tries to establish TCP connections with relevant Web servers, on which each element is loaded by separate HTTP request and the corresponding HTTP response. A TCP connection is generally used to transfer only one element, but it can be used to transfer several elements serially when Web server supports HTTP keep-alive. Thus, other elements may be transferred within the first TCP connection besides H_0. The transfers of the first Web page and other Web pages are similar. There normally exist 2~4 simultaneous TCP connections downloading HTML documents and other elements of a set of Web pages for an access.

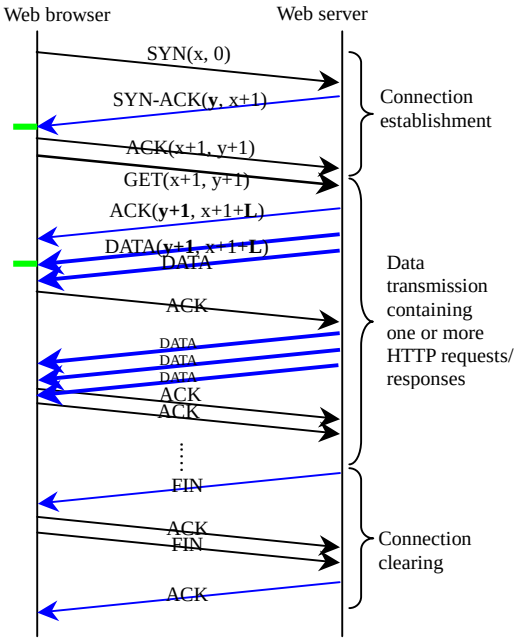

Figure 1. HTTP/TCP connection

As a Web client, a Web browser exchanges information with Web servers based on HTTP/TCP connections. HTTP requests and HTTP responses are embedded in TCP segments, as showed in Figure 1. On an established TCP connection, one or more Web elements may be transferred, after that the connection is terminated or closed. A Web server, after receiving a HTTP request, often answers with a pure ACK (acknowledgement) followed by DATA segments carrying a HTTP response, or just responds with DATA segments. While the Web client may transmit a pure ACK after it has received two DATA segments or a DATA segment with PUSH flag set. Data transmission of the Web server is governed by the minimum of the congestion window (cwnd) and the receiver's advertised window (rwnd) [5]. The initial value of cwnd may be less than or equal to 2*SMSS (sender maximum segment size). At the beginning of a transfer, TCP increments cwnd by at most SMSS bytes for each ACK received that acknowledges new data, until congestion is observed (3 duplicate ACKs are received or retransmission timer expires) or cwnd exceeds ssthresh (slow start threshold). In Figure 1, a majority of segments contain no HTTP message, thus belong to control packets. Partial TCP sequence numbers and acknowledgment numbers are given in the parentheses, where the L denotes the length of the first HTTP request.

We got sufficient client based traces through capturing Web traffic to and from SINA and SOHU Web sites at the side of the Web client. The results below have been obtained from simple analysis of those traces.

(1) For upstream traffic, control packets probably account for around 80% of total outbound packets; the average IP total length for all outbound packets is usually less than 100 bytes.

(2) For downstream traffic, control packets account for 30%~60% of total inbound packets, which mainly depends on factors such as the sizes of Web elements, whether one or multiple elements are loaded on a TCP connection, and whether elements have already been stored in Web client's cache. Control packets and large DATA packets (IP total length>1400 bytes) go over 70% or so, which depends on path MTU in addition to above factors. The average IP total length for all inbound packets usually exceeds 500 bytes.

(3) The number of inbound packets is approximately equal to that of outbound packets when new Web pages are accessed. Possible explanation is that each Web element is loaded on separate TCP connection, and moreover, DATA packets of each TCP connection amount to 2~5.

(4) Almost all spaces between any two successive inbound packets during an access are less than 3s, and the average space ranges from tens to more than 100ms for broadband access or LAN access in campuses; while almost all less than 6s and from nearly 200ms to nearly 400ms respectively for 56K dial-up access.

2.2 User Access Behavior Analysis

User behavior of accessing Web is influenced by many factors such as charge manner, personal reading ability and characteristic, information capacity of Web pages, and so on. We analyze user access behavior based on Web traffic collected at the user link.

When charged fixedly per month or based on traffic, in general, people surf the WWW relaxedly to search information interested in. Of Course, users may frequently convert to other hyperlinks to start another "access", when they find Web pages being downloaded uninteresting. Statistics for broadband or LAN access show that packets carrying distinct sets of Web pages resulting from adjacent accesses generally arrive with at least 3s gap. Seldom occur speedy and continuous two accesses, whose inbound packets are so mixed that two accesses cannot be delimited in terms of time. However, two adjacent accesses frequently cannot be delimited by arrival spaces between adjacent inbound packets for 56K dial-up access.

When charged based on usage time, users may utilize the WWW more economically, and even read offline to save time. Thus there is more probability that speedy and continuous two accesses occur.

To sum up, access behavior for a single user to visit a Web site can be divided into three cases. First is an isolated access, which can be easily delimited by arrival time of inbound packets. Packets belonging to different such accesses arrive with about at least 3s gap for broadband or LAN access, about over 6s for dial-up access. Second is a series of accesses, during which inbound packets of any two adjacent accesses arrive separately or with a little overlap in terms of time. We term such a series of accesses and an isolated access a weak burst access. Third is continuous multiple accesses, during which inbound packets of any two adjacent accesses arrive with more overlap. We term such multiple accesses a strong burst access. With the improvement in Internet access and billing, more and more user access behavior will belong to case 1 and 2. Burst accesses can be effectively distinguished from each other according to arrival time of their inbound packets.

3 WEB SERVICE-INTRINSIC PERFORMANCE METRICS DEFINITION

Two metrics easy to measure are defined herein to characterize information delivery quality most important for perceived Web QoS present to a single user. They are average round-trip time (unit: ms) and delivery speed (unit: number of packets per second). Both can be measured based upon a single measurement point that should be set on the user host or other host on the same broadcast link with the user host.

(1) Average round-trip time (RTT) is defined as the average of all TCP round-trip times for Web traffic between a user and a specified Web site during a period of time. TCP RTT is defined as the time it takes for the SYN-ACK packet and for the first DATA packet on the same TCP connection generated by a Web server in answer to the first HTTP request to reach the Web client respectively. The two packets have the same source address, destination address, and protocol (6 for TCP) in IP header, as well as the same source port (e.g. 80) and destination port in TCP header, where the source address refers to one of Web servers of the Web site, and the destination address refers to the user host. Furthermore, the TCP sequence numbers of the former packet is one less than that of the latter.

(2) Delivery speed is defined as the average of burst arrival rates for inbound packets in burst

accesses to a specified Web site during a period of time. Burst arrival rate is defined as the arrival rate for inbound packets during a burst access. Inbound packets attribute to a burst access have the following characteristics: the source IP address refers to any Web server of the Web site, the destination IP address refers to the user host, and the time space between any two adjacent packets is less than the threshold (e.g. 3s for broadband or LAN access, 6s for dial-up access). Denote arrival times for inbound packets in a burst access t_1, t_2, t_3, … t_n, where n is the number of inbound packets, we can compute the burst arrival rate by $(n-1)/(t_n-t_1)$. Actually there are a few scattered inbound packets with fairly long time space at the end of a burst access. Though hardly impacting user perception, they may be taken as generated by other burst accesses. Thus any burst access with n less than a lower limit should better be omitted.

4 INSTANTIATION AND ANALYSIS

We conducted experiments at three places in a campus network that is connected to the Internet via 1Gbps link: one is building A, which is connected to campus network center via 1Gbps fiber link, the second is building B, which is connected to campus network center via 155Mbps fiber link, and the third is an abode connected to the campus network center via dial-up link. In general, users often feel poor or ordinary when surfing WWW at the abode; fairly good at building B; and a little more better at building A. At three places, Web traffic trace consisting of about 800 samples (inbound packets) was recorded to calculate above two metrics while several users visited the SINA Web site on one host (thus actually a single user); meanwhile the user-perceived service quality of the Web site was evaluated subjectively. The measurement results are given in Table 1.

test places	Average RTT(ms) (x_1)	Delivery speed (number of packets per second) (x_2)	Subjective evaluation of Perceived Web QoS (y)	Estimate of y ($\hat{y}$)	Measurement time
Building A	123.98	48.22	9	8.9304	2003-9-27 Sat. 14:12:07.821~14:12: 48.229
	183.75	24.46	8	8.2108	2003-9-27 Sat. 12:36:24.943~12:45:08.516
Building B	259.311	18.206	8	7.9952	2003-10-15 Wed. 15:03:53.562~15:16:08.629
	188.174	12.084	8	7.8477	2003-10-15 Wed. 15:44:33.230~15:47:03.987
Abode	1829.051	5.148	7	6.9273	2003-10-12 Sun. 10:37:29.500~10:43:20.781
	1496.688	5.699	7	7.0888	2003-10-12 Sun. 10:44:25.750~10:49:48.531

Table 1. Average RTT and delivery speed measured on client-based Web traffic traces to the SINA Web site as well as subjective evaluations and regression estimates of user-perceived Web QoS

Next we analyze the relationship between two metrics and perceived Web QoS. In nature, perceived Web QoS is personal temporal feel while accessing a Web site. Due to the complexity of feel, it has no absolutely objective value, however it may have diverse subjective evaluations that may vary from person to person. As it is reasonable to suppose that subjective evaluations are normally distributed, the mean can be used as the unique subjective value. In addition, user-perceived Web QoS is a time variable, but may be stable during a few minutes. In the experiments, users objectively scored their true feel when accessing SINA Web site with a real number from 0 to 10 (10 means the perfect service quality). Suppose two metrics defined above are independent variables, and perceived Web QoS is dependent variable. We attempt to apply the multiple linear regression models on the observation data in Table 1:

$$y = X\beta + \varepsilon,\ where\ \varepsilon \sim N(0, \sigma^2 I),\ X = \begin{pmatrix} 1 & 123.98 & 48.22 \\ 1 & 183.75 & 24.46 \\ 1 & 259.311 & 18.206 \\ 1 & 188.174 & 12.084 \\ 1 & 1829.051 & 5.148 \\ 1 & 1496.688 & 5.699 \end{pmatrix}, \beta = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}, y = \begin{pmatrix} 9 \\ 8 \\ 8 \\ 8 \\ 7 \\ 7 \end{pmatrix} \quad (1)$$

We plug in the estimate values for the β coefficients calculated through using MATLAB, then get the linear regression function:

$$\hat{y} = 7.5773 - 0.0004x_1 + 0.0292x_2 \quad (2)$$

Next we should verify two hypotheses in (1) and (2). Firstly verify whether y and x_1,x_2 indeed have linear relationship. b_1=b_2=0 means that y and x_1,x_2 have no relationship, thus we verify whether the hypothesis

$$H_0: b_1=b_2=0$$

is true [6]. Suppose H_0 is true, then

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{m}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m-1}} \sim F(m, n-m-1)$$

For given level α, if $F > F_{1-\alpha}(m, n-m-1)$, then H_0 should be rejected. For our observation data, m=2, n=6, $F = (2.7477/2)/(0.0857/(6-2-1)) = 48.1$, and $F_{1-0.01}(2,3) = 30.8165$ for α=0.01. H_0 is false because $F > F_{1-0.01}(2,3)$, that is, the above regression function is significant at the level 0.01, and moreover the regression effect is high significant [6].

Secondly, even though there is linear relationship between y and x_1,x_2, it does not mean all independent variables influence the random variable y significantly. To verify whether x_i influences y significantly, we should verify whether m+1 hypotheses

$$H_{0j} : b_j = 0,\ j = 0,1,2$$

are true [6]. Suppose the hypothesis H_{0j} (j=0,1, or 2) is true, then

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-m-1),\ j=0,1, \text{ or } 2$$

where c_{jj} is the (j+1)th element on the diagonal in matrix $(X'X)^{-1}$, and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m-1}$. For given level α, if $|t_j| > t_{\alpha/2}(\text{freedom degree is } n-m-1)$, then H_{0j} should be rejected. For our observation data, m=2, n=6, n-m-1=3, t_0=36.806, t_1=-3.2806, t_2=4.5861, and $t_{\alpha/2}$=3.1824 for α=0.05. Obviously, all H_{0j} should be rejected, that is, both metrics influence perceived Web QoS significantly at the level 0.05.

5 CONCLUSION

In this paper we analyze characteristics of Web traffic as well as user access behavior, based on which two metrics for Web service intrinsic performance, average round-trip time and delivery speed, are defined. The novel aspect of this paper is that the latter metric is not TCP connection-based, but based on weak burst accesses. We consider that TCP connections contained in each weak burst access together determine user-perceived Web QoS rather than separate TCP connections. Through our experiments, we have verified that perceived Web QoS can be objectively estimated by above two metrics in multiple linear regression function.

6 REFERENCES

[1] GB 917. SLA Management Handbook Public Evaluation/Version 1.5[EB/OL]. http://www.tmforum.org. 2001-6-1/2002-10-10

[2] TMF 701 Performance Reporting Concepts & Definitions Public Version 2.0[EB/OL]. http://www.tmforum.org. 2001-11-1/2002-10-10.

[3] Li Xiaochun, Zeng Yao. Quality Management Science [M]. Beijing: Beijing University of Posts and Telecommunications Press, 2002(in Chinese). 27~29,50~52.

[4] Wei Fuxiang. Active Relationship among the Customer's Perceived Service Quality, Customer's Satisfaction and Customer'S Loyalty [J]. Modern Finance & Economics, 2001(in Chinese), 21(7): 39~42.

[5] M. Allman, V. Paxson, W. Stevens. RFC2581: TCP Congestion Control [S]. http://www.ietf.org/. 1999–4-7/2004-3-10.

[6] Zhu Daoyuan, Wu Chengou, Qin Weiliang. Multivariate Statistical Analysis and SAS Software [M]. Nanjing: Southeast University Press, 1999(in Chinese). 258~269.