Mean Waiting Time of an End-user in the Multiple Web Access Environment

Yong-Jin Lee
Department of Technology Education
Korea National University of Education
Cheongwon, Korea
e-mail: lyj@knue.ac.kr

Abstract— **Mean response time for single user and mean waiting time for multiple users are important measures of Quality of Service (QoS) in accessing a web server. This paper presents analytical models to find the mean response time and the mean waiting time for web service using Hyper Text Transfer protocol (HTTP) over Stream Control Transmission Protocol (SCTP). The proposed response and waiting time model assumes the multiple packet losses and a narrowband network, where fast retransmission is not possible due to small window. Our experiments validate the accuracy of the proposed model. It is shown that the differences between the results from the model and those from the experiments are very small on average. We also find that the mean waiting time for HTTP over SCTP is less than that for HTTP over TCP. The model can be used for dimensioning of the network link bandwidth to satisfy the QoS of end users.**

Keywords-mean waiting time; multiple web access; QoS

I. INTRODUCTION

TCP [1] provides a single streamed and strictly ordered delivery of data, which increases the users' perceived latency. SCTP [2,3] was proposed as a new transport layer protocol which has multi-streaming capability to transmit several independent streams of chunks (or messages) in parallel. When a packet loss occurs in a stream, it affects the relevant stream only.

Typically, response time is affected by data size and transmission time according to transmission rate of link as well as by congestion control mechanism. The congestion control mechanism of SCTP is similar with window-based one of TCP. Their common functions are slow-start, congestion avoidance, timeout, and fast retransmission.

Previous related works on analytical models of data transmission delay over TCP are as following: Padhye [4] considered large amount of data transmission on steady state over TCP. Most of TCP connections for HTTP data transmission, however, are short for small amount of data instead of large one in current internet environment. Connection setup or slow-start time dominates the performance of web in this environment. Noticing this phenomenon, Cardwell [5] extended the above steady state model but he did not consider delay of TCP after time-out. Jiong [6] enhanced the Cardwell's model by considering slow-start time after timeout of retransmission. However, since the above models assumed wideband network, they cannot be applied to the narrowband network environment, which this paper considers. That is because the narrowband network environment does not allow fast retransmission of data due to the very small size of window [7]. Furthermore, the previous studies are limited to single user cases, where the response time is a good measure of the end-to-end delay experienced by a user.

Chang et al. [8] studied the performance of File Transfer Protocol (FTP) over SCTP, and Lu [9] analyzed the performance of Session Initiated Protocol (SIP) over SCTP. Fei Ge [10] presents a simple closed-form formula to estimate the HTTP latency over FAST TCP, taking into account the network parameters such as packet size, link capacity, and propagation delay. Eklund et al. [11] developed a model that predicts the transfer times of SCTP messages during slow start. However, mean waiting time model for HTTP over SCTP in multiple users' environment has not yet been presented.

The motivation of this paper is to study the case of multiple users accessing a server, where the waiting and turnaround times depend on the server load. In such a case, the response time may not be a good measure of end-to-end delay.

The results reported in this paper can be used by network engineers to dimension a network in terms of bandwidth requirement and to develop scheme distributing the load among a number of web servers, in order to improve the waiting delay perceived by end users. The objective of this study is to find the theoretical upper bound of the actual waiting and turnaround times of users in a real environment when they download web objects using HTTP over SCTP in the narrowband network, which does not allow fast retransmission.

We achieve our objectives by developing an analytical model to compute the mean waiting and turnaround time of an end user when multiple users simultaneously access the web server. In contrast to previous work [12,13,14], which only considered the response time of an object for single user, we first consider the response time for single user and then find waiting delay for multiple users. The results of this paper will allow us to compute more realistic end-to-end delay experienced by a user in the real environment.

Since the estimated mean waiting time in this paper can be considered as QoS of end-users, it can be used as a benchmark to pre-estimate waiting time by considering size of objects, bandwidth, and round trip time. To validate the proposed mean waiting time model, we experimented in a simple test-bed and compared the results with estimated value. In addition, we compared the values with the mean waiting time of HTTP over TCP.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-263-9

Sections 2 and 3 describe the estimation model and algorithm of mean response and waiting time for HTTP over SCTP, respectively. Section 4 discusses performance evaluation and analysis. We conclude this paper in section 5.

II. MEAN RESPONSE TIME MODEL FOR SINGLE USER

In this section, we first describe the mean response time model, when single user retrieves a web object in the narrowband network [14].

Fig. 1 shows the congestion control mechanism of SCTP in the narrowband network. In Fig.1, $th(1)$, $th(2)$, and $th(3)$ are the slow start thresholds and initially $th(1)=\infty$. y coordinate is the congestion window($cwnd$) and its initial value is $2\times mtu$. Here, mtu represents the maximum transfer unit of the link. Thus SCTP executes the slow-start period by increasing $cwnd$ exponentially such as 2, 4, 8, ... and detects the packet loss when timeout occurs at ①. SCTP responds to this as following.

$$th(2) = max(cwnd/2,\ 2 \times mtu\ )$$
$$cwnd = 1 \times mtu \qquad (1)$$

That is, the threshold of next stage is reduced to half size of the window in which packet loss occurred and slow-start period is repeated with congestion windows exponentially increased from 1 to 2, 4, 8, etc. When the congestion window exceeds threshold $th(2)$, congestion avoidance period is started. Since this period needs an acknowledgement every packet, it is called linearly increasing period. If a packet loss occurs as Fig. 1, ② in this period, there are two choices according to timeout. First of all, using (1) new threshold ($th(3)$) is obtained. If three duplicate acknowledgements are obtained before timeout, then fast retransmission (Fig. 1, ③) is started. Otherwise slow-start (Fig. 1, ④) is executed. In this paper, we assume the narrowband network which is not able to receive three duplicate acknowledgements during timeout. Thus the slow-start is executed.

In order to simplify the model we assume that sizes of web objects are identical and received packets are transmitted in an upper layer in terms of window unit. Let the size of an object to transfer be θ bits and maximum transfer unit mtu bits, then the number of packets to transfer for an object is $n = \lceil \theta/mtu \rceil$.

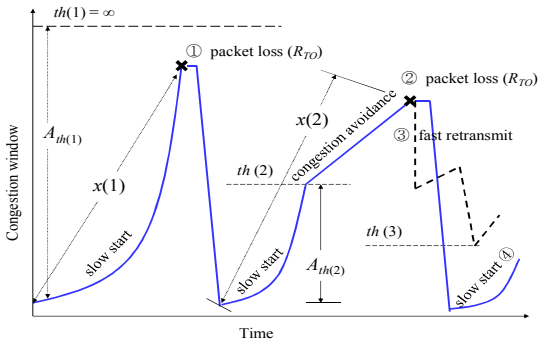

Figure 1. Congestion control of SCTP in the narrowband network

When the probability of a packet loss is p, the expected number of total packet loss is $\alpha = \lceil np \rceil$ in terms of binomial distribution. At this moment, a certain packet loss occurs during either slow-start phase or congestion avoidance phase.

From the above, we can identify the packet loss phase by comparing, for k^{th} packet loss, the possible number of packets ($A_{th(k)}$) to transmit until the threshold ($th(k)$, k=1,2,..,a) at which congestion avoidance starts, with the expected number of packets ($x(k)$: k=1,2,..,a) transmitted before the packet loss. At this time, $x(k)$ is calculated as a function of remained packets $N(k)$ and packet loss rate p.

We can determine that an arbitrary k^{th} packet loss occurs either during slow-start phase or congestion avoidance phase, when either $x(k)<A_{th(k)}$ or $x(k)\geq A_{th(k)}$, respectively. For example, in Fig. 1, the total number of packets transmitted is $x(1)$ until the first loss ① and the possible number of packets to transmit is $A_{th(1)}$ until $th(1)$. And since $x(1)<A_{th(1)}$, it is considered that the packet loss occurs during slow-start phase. Similarly, since the number of packets sent before the loss ② is $x(2)>A_{th(2)}$, it is determined that the packet loss occurs during congestion avoidance.

Mean response time for HTTP over SCTP is given as (2).

$$E(T_{sctp}) = \sum_{k=1}^{\alpha} \left[\beta E(T_{slow}^{k}) + (1-\beta)E(T_{cong}^{k})\right] + R \qquad (2)$$

Since the first packet loss (k=1) of SCTP in (2) occurs always during slow-start phase as shown in Fig. 1, $E(T_{slow}^{1})$ needs to be added. Packet losses after second one occur during either slow-start phase or congestion avoidance phase. $E(T_{slow}^{k})$ and $E(T_{cong}^{k})$ represent mean response time, when the k^{th} packet loss (k=2,3,...a) occurs during slow-start phase and congestion avoidance phase, respectively. Detailed computation procedures of $E(T_{slow}^{k})$ and $E(T_{cong}^{k})$ are presented in [14]. Since an arbitrary packet loss cannot occurs simultaneously during slow-start phase and congestion avoidance phase, β is either 0 or 1 for the given k^{th} packet loss. That is, if k^{th} packet loss occurs during slow-start phase and β=1, then $E(T_{sctp})$ is accumulated by adding $E(T_{slow}^{k})$. Similarly, if k^{th} packet loss occurs during congestion avoidance phase and β=0, then $E(T_{sctp})$ is accumulated by adding $E(T_{cong}^{k})$. Therefore the total mean response time of an object needs to add either $E(T_{slow}^{k})$ or $E(T_{cong}^{k})$ (k=1,2,...,a) as the expected value of lost packet number (a). R, which is the time to transfer the remained data, $N(a+1)$ after the last packet loss occurred, can be calculated without considering additional packet losses since the expected value of packet losses is already equal to a. That is, if $N(a+1)$ is less than the possible amount of data to transfer until the last threshold $th(a+1)$, the transmission is completed during slow-start phase. Therefore R is sum of slow-start time ($ST(N(a+1))$) and transmission time ($N(a+1)\times mtu/\mu$) until then. μ represents the bandwidth of the link. Otherwise the transmission is completed during congestion avoidance phase. Thus R is sum of slow-start time ($ST(A_{th(a+1)})$) and transmission

time ($N(a+1)\times mtu/\mu$) until the threshold adding the extra time ($(N(a+1)-A_{th}(a+1))\times rtt$) in congestion avoidance phase.

III. MEAN WAITING TIME MODEL FOR MULTIPLE USERS

The mean response time of HTTP over SCTP ($E(T_{sctp})$) found in the previous section is total time for a user to connect to a web server and download an object. Mean waiting and turnaround time are defined as the performance measure when multiple users access the web server simultaneously.

We assume the asynchronous TDM (time division multiplexing) based on packet for web service. A web object consists of n packets, thus, packet response time (τ) is equal to $E(T_{sctp})/n$ when every τ is the same. Also, n is given by $\lceil \theta/mtu \rceil$. Now, if we assume that four clients (m=4) request the same file, each user's expected response time ($E(T_{sctp})$) will be the same. For example, we consider the case where n=3 with the asynchronous TDM. When a client requests an object from the server, three packets are included in the object. $E(T_{sctp})$ means total response time that each client expects.

Now, we develop analytical models for the mean waiting and turnaround times for two cases depending on whether the packet response times are same or not.

When the web servers are connected to the external users through only one link, the total waiting time, the mean waiting time(W_{sctp}^{same}), total turnaround time, and mean turnaround time(T_{sctp}^{same}) are given by the following equations:

$$total\ \ waiting\ \ time = \sum_{i=1}^{m}(m-i)\tau + m(n-1)(m-1)\tau \tag{3}$$

$$W_{sctp}^{\ same} = \frac{\sum_{i=1}^{m}(m-i)\tau + m(n-1)(m-1)\tau}{m} \tag{4}$$

$$total\ \ turnaround\ \ \ time = m\tau[m(n-1) + \frac{m+1}{2}] \tag{5}$$

$$T_{sctp}^{same} = \frac{1}{m}[m(n\text{-}1) + \frac{m+1}{2}]\tau = [\frac{2mn\text{-}m+1}{2}]\tau \tag{6}$$

When the web servers are connected to the external users through several links of different bandwidths, the mean waiting and turnaround time are given by (7) and (8) respectively. First, we consider the mean waiting time. To find the waiting time of i^{th} user, we divide the total time into two intervals: the first interval represents the time when all the packets except the last packet of each user has been received; the second interval represents the time when the last packet of each user has been received. Total waiting time of i^{th} user until the first interval is (the number of packets–1)×[(the number of users for group including i^{th} user–1)×τ_i+(total packet response time excluding i^{th} group)]. The waiting time of i^{th} user is the sum of response times of other users prior to him. By generalizing and adding this all, we obtain the following equation for the mean waiting time. Both m_0 and τ_0 are zeros in the equation.

$$W_{sctp}^{\ diff} = \frac{(n-1)\sum_{i=1}^{P} m_i[(m_i-1)\tau_i + \sum_{i=1, j\neq i}^{P} m_j\tau_j]] + \sum_{i=1}^{P}[\sum_{j=1}^{i} m_i(m_{j-1}\tau_{j-1}) + \sum_{j=1}^{m_i}(j-1)\tau_i]}{m} \tag{7}$$

Now, we consider the mean turnaround time. If we use the same procedure as the waiting time, total turnaround time of i^{th} user until the second interval is (the number of packets–1)× [the number of users (m_i)×the sum of packet response time (τ_i)]. The turnaround time of any user in the second interval is the sum of response times of other users prior to him and his own packet response time. Thus, by generalizing and adding this all, we obtain the following equation. Both m_0 and τ_0 are zeros in the equation.

$$T_{sctp}^{\ diff} = \frac{m(n-1)\sum_{i=1}^{P} m_i\tau_i + \sum_{i=1}^{P}[\sum_{j=1}^{i} m_i(m_{j-1}\tau_{j-1}) + \sum_{j=1}^{m_i} j\tau_i]}{m} \tag{8}$$

IV. PERFORMANCE EVALUATION

Based on the model discussed in section 2 and 3, we can construct an algorithm for the whole procedure as in Algorithm 1 (Fig. 2). When the number of packets for an object is n, the complexity of the algorithm is $O(n)$.

We consider a simulation of web server for TCP and SCTP, and an environment to emulate HTTP. Desktop computers are used as client-server to send data. In order to simulate real network, we use a laptop computer with NIST emulator [15] between a client and a server, and adjust various network conditions such as packet loss (p), bandwidth (μ), and RTT (rtt).

Algorithm 1. mean waiting and turnaround time for multiple users

```
Begin
Compute the total number of packets in object
    (n = ⌈θ/mtu⌉)
Compute the expected number of packet loss (α = ⌈np⌉)
Set N(1) = n and th(1) = ∞
Set E(T_sctp) = 0
for all k such that k=1,2…, α do
      Find E(T_slow^k) and E(T_cong^k)
end for
Find the mean response time, E(T_sctp)= E(T_sctp) + R
Find the packet response time, τ = E(T_sctp) / n
If τ is same for all bandwidth type i,
      Find mean waiting ( W_sctp^diff ) and turnaround time using
      (4) and (6) respectively.
else
      Find mean waiting and turnaround time using (7) and
      (8), respectively.
endif
End
```

Figure 2. mean waiting and turnaround time for multiple users

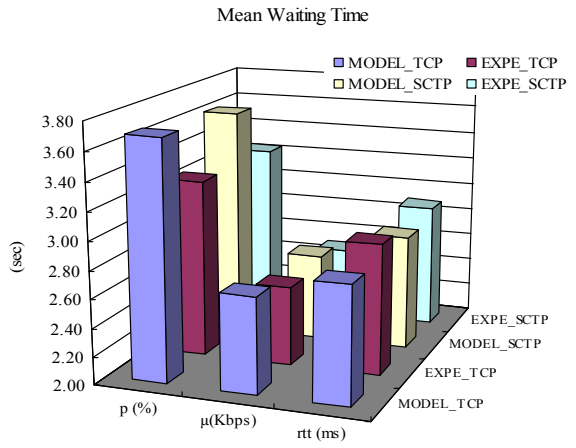

Figure 3. Mean waiting times for p, μ, rtt

Except the number of initial windows, HTTP over TCP model is basically same as HTTP over SCTP. That is, except that mean response time ($E(T^1_{slow})$) for the case of first packet loss occurred in slow-start phase of Algorithm 1 is computed differently, the procedures are same. Mean object size (θ) is 13.5KB and maximum transmission unit (mtu) is 536B. A HTML file contains five web objects.

Our experiments were performed as follows: Firstly, we changed p from 0.4 to 2% after fixing rtt=256ms and μ=40Kbps. Secondly, we changed μ from 400Kbps to 3Mbps after fixing p=1% and rtt=256ms. Finally, we changed rtt from 55ms to 256ms after fixing p=1% and μ=40Kbps.

Fig. 3 depicts the summary of mean waiting times (sec) for each p, μ, rtt. In the figure, MODEL_SCTP and EXPE_SCTP represent W^{diff}_{sctp} and T_{sctp}, respectively. MODEL_TCP and EXPE_TCP also represent W^{diff}_{tcp} and T_{tcp}, respectively. Fig. 3 shows that both models for HTTP over SCTP and HTTP over TCP overestimate mean waiting times for p and μ, respectively, but, models underestimate them for rtt.

Now, we define the mean difference ratio between models and experiments by (9).

$$DIFF_{mean} = \sum_{i=1}^{n} \left[\frac{W^{diff}_{sctp} - T_{sctp}}{W^{diff}_{sctp}} + \frac{W^{diff}_{tcp} - T_{tcp}}{W^{diff}_{tcp}} \right] / n \times 100 \qquad (9)$$

The computed $DIFF_{mean}$ is 4.17%. This value is small; however, more experiments and model adjustments are necessary to describe the real environment exactly. Additionally, we find that the mean waiting time of HTTP over SCTP is less than HTTP over TCP on both the model and experiment.

V. CONCLUSIONS

This paper presents an analytical model to estimate mean waiting time of web service using HTTP over SCTP in the narrowband network when multiple users access web server simultaneously. We first describe the mean response time model for single user, which is one of QoS offered to web users and one of essential parameters to evaluate web performance. We then extend the mean response time model to the mean waiting and turnaround time models for multiple users. Simple test-bed simulation results show that the mean difference ratio, between the analytical model and experiment, is small. Further extension of this work includes model with higher accuracy for the real environment.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A4A01003651)

REFERENCES

[1] V. Paxon, M. Allman, and W. Stevens, "TCPs congestion control," RFC 2581, 1999. http://www.ietf.org/rfc/rfc2581.txt.

[2] R. Stewart, "Stream control transmission protocol (SCTP), RFC 4960, 2007. http://www.ietf.org/rfc/rfc4960.txt.

[3] L. Budzisz, J. Garcia, A. Brunstrom, and R. Ferrus, "A Taxonomy and Survey of SCTP research," ACM Computing Surveys, vol. 44, no. 4, 2012, pp. 18:1-18:36.

[4] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling TCP Reno performance: A simple model and its empirical validation," ACM Transactions on Networking, vol. 8, no. 2, 2000, pp. 133-145.

[5] N. Cardwell, S. Savage, and Y. Anderson, "Modeling TCP latency," Proceeding of the 2000 IEEE Infocom Conference, 2000, pp. 1742-1751.

[6] Z. Jiong, Z. Shu-Jing, and Qi-Gang, "An adapted full model for TCP latency," Proceedings of. the 2002 IEEE TENCON Conference, Vol. 2, 2002, pp.801-804.

[7] D. Oliveria and R. Braun, "A dynamic adative acknowledgement strategy for TCP over multihop wireless networks," Proceedings of the IEEE INFOCOM Conference, 2005, pp.1863-1874.

[8] Lin-Huang Chang, Ming-Yi Liao and De-Yu Wang, "Analysis of FTP over SCTP in Congested Network," 2007 International Conference on Advanced Information Technologies (AIT), 2007, pp. 82-89.

[9] Chia-Wen Lu and Quincy Wur, "Perfomance study on SNMP and SIP over SCTP in wireless sensor networks," 14th International conference on advanced communication technology (ICACT), 2012, pp. 844-847.

[10] Fei Ge, Liansheng Tan, Jinsheng Sun, and Moshe Zukerman, "Latency of fast TCP for HTTP transactions," IEEE Communications Letters, vol. 15, no. 11, 2011, pp. 1259-1261.

[11] J. Eklund, K. Grinnemo, A.Brunstrom, G. Cheimonidis, and Y. Ismailov, "Impact of Slow Start on SCTP Handover Performance," Proceedings of the 20th international conference on computer communications and networks, 2011, pp.1-7.

[12] Y. Lee, M. Atiquzzaman, and S. Sivagurunathan, "Mean response time estimation for HTTP over SCTP in wireless environment," Proceedimgs of the 2006 IEEE International Conference on Communications, 2006, pp.164-169.

[13] Y. Lee and M. Atiquzzaman, "Mean waiting delay for web object transfer in wireless environment," Proceedings of the 2009 IEEE International Conference on Communications, 2009, pp.1-5.

[14] Y. Lee, "Mean response delay estimation for HTTP over SCTP in wireless Internet," Journal of the Korea Contents Association, vol. 8, no. 6, 2008, pp. 43-53.

[15] Mark Carson and Darin Santay, "NIST Net – A Linux-based Network Emulation Tool," ACM SIGCOMM Computer Communication Review, vol. 33, no. 3, 2003, pp. 111-126. http://snad.ncsl.nist.gov/itg/nistnet

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-263-9