Effect of payload size on mean response time when message segmentations occur using $M^X/G/1$ queueing model

Takashi Ikegawa

Abstract

This paper proposes the $M^X/G/1$ queueing model to represent arrivals of segmented packets when message segmentations occur. This queueing model enables us to derive the closed form of mean response time, given payload size, message size distribution and message arrival rate. From a numerical result, we show that the mean response time is more convex in payload sizes if message arrival rate is larger in a scenario where Web objects are delivered over a physical link.

Index Terms

Message segmentation, payload size, response time, $M^X/G/1$ queueing model.

I. INTRODUCTION

The size of packets, i.e., data units transferred over physical links, affects several quality of service (QoS) parameters for users of packet-based transfer networks such as the Internet. For example, the response time between hosts is highly dependent on packet size because it contains link-level transmission time, which is simply given by the packet size divided by capacity of the physical link [1, p. 9]. In addition, the packet-loss rate due to bit errors depends on the packet size because it is approximately proportional to the packet size [2, p. 132].

The packet size is limited for various reasons [3, pp. 406–409]. They include a) the data-link structure (e.g., the width of a transmission slot), b) compliance with standard protocol specifications, and c) satisfaction of the QoS parameters by applications such that the round-trip time of interactive applications is less than the time that a user is willing to wait [4], [5].

Messages, i.e., data units generated by applications, are frequently larger than the maximum permitted packet size. To convey such messages over the network, some communication protocols, such as transmission control protocol (TCP), Internet protocol (IP) for the Internet [6], IEEE 802.11 media access control (MAC) protocol [7] for wireless local-area networks (LANs) and radio link control (RLC) protocol [8] for mobile wide-area networks, specify a message segmentation/reassembly function. The message segmentation function enables a sender to divide a single message larger than the payload size $\ell_d$ into multiple packets. Furthermore, the sender adds an appropriate header, i.e., overhead, to a packet.

There have been several studies on optimization of packet size (or payload size) to satisfy the user’s QoS such as [9]–[11]. However, the purpose of these studies was to solve the tradeoff issue between the desire to reduce the header overhead by making packet large, and the need to reduce packet-loss rate due to bit errors in noisy links by using small packet size.

Another tradeoff exists when message segmentations occur. Consider the message whose variance is large enough in size, such as Web files (objects). While too large payload size is employed, message segmentations hardly occur. In this case, the waiting times of packets for transmission using a physical link is large because the packet size distribution can be identified with the message size distribution. On the other hand, when payload size is small, waiting times of packets may be small due to decreasing the variance of packet sizes because the constant

T. Ikegawa is with Waseda Research Institute for Science and Engineering, Waseda University and Graduate School of Mathematical Sciences, the University of Tokyo, Japan (e-mail: ikegawa@aoni.waseda.jp or tikegawa@ms.u-tokyo.ac.jp).
packets in size, which is payload size, are dominant in the all created packets. When payload size is too small, the number of segmented packets per message increases significantly, resulting in very large waiting time due to the burstiness of packet arrivals.

In previous work, this tradeoff issue has not been discussed. The purpose of this paper is to discuss the effect of payload size on mean response time when message segmentations occur using an $M^X/G/1$ queueing model [12, Chapter 4], which can capture the behavior of the burstiness of segmented packet arrivals.

The rest of the paper is organized as follows. In the next section, we describe the communication network model underlying our study. Section III analyzes the segmented packet sequence. Section IV derives the closed form of mean response time using an $M^X/G/1$ queueing model. Section V investigates the effect of message segmentation on the mean response time for actual message size distributions. Finally, Section VI summarizes this paper and mentions future work.

**II. COMMUNICATION NETWORK MODEL**

In this section, we first explain the two-layered communication network model under consideration. Next, the model of data units introduced in this paper at the respective layer is described. In final, we explain some assumptions for analytical tractability.

**A. Layer model**

To investigate the effect of message segmentation on performance, we consider a communication network of which conceptual representation is shown in Fig. 1. Each station (a sender and a receiver) has two layers: application and segmentation-reassembly (SGM-RAS).

The application layer contains a traffic source and sink. The traffic source generates the data units, which will be referred to as messages. On the other hand, the traffic sink terminates the corresponding data units.

The SGM-RAS layer implements message SGM-RAS function. In addition, it has a function to transfer data units, which will be referred to as packets, over physical links at a sender.
B. Data-unit model

We define data units exchanged between peer entities at the respective layer: messages and packets.

Message: a data unit generated by a traffic source with a given size distribution,
Packet: a data unit created from a message through segmentation function by adding a header and/or trailer, i.e., control information, to the (divided) message. The message segmentation function implemented in the sender’s SGM-RAS layer enables a single message to be divided into several packets if the message size is larger than the payload size $\ell_d (> 0)$. The receiver’s SGM-RAS layer performs a message reassembly function, thus reassembling the segmented generated packets before delivering them to the application layer.

C. Assumptions

For analytical tractability, we make the following assumptions.

A1: message sizes are mutually independent and identically distributed according to a common message-size distribution $F(m)(\cdot)$. The distribution $F(m)(\cdot)$ has a finite mean value $\ell(m)$, which is referred to as the mean message size.

A2: the finite variance $\{\sigma^2(m)\}^2$ of the message-size distribution exists.

III. SEGMENTED PACKET SEQUENCE MODEL

The creation of packets from a message through message segmentation is shown in Fig. 2. If a message is larger than payload size $\ell_d$, the message is divided into multiple packets. Two kinds of packets exist:

body packet: a segmented packet appearing between the head and the penultimate packets in the original message, whose packet size is always equal to $\ell_d$, and
edge packet: the final segmented packet if a message is segmented, or the message itself if it is not segmented, whose packet size is variable but does not exceed $\ell_d$.

Let a random variable $L$ be a packet size excluding control information (header). Letting $F(p)(\cdot) \triangleq \Pr.(L \leq x)$ be the stationary distribution of packet sizes, we have

$$F(p)(x) = (1 - \pi(E)) \cdot 1(x - \ell_d) + \pi(E)F(E)(x).$$

(1)
where $F^{(E)}(\cdot)$ is the stationary edge-packet-size distribution and $\pi_E$ is the edge packet occurrence probability. The form of $F^{(E)}(x)$ is given by
\[
F^{(E)}(x) = \begin{cases} 
0, & \text{if } x \leq 0, \\
\sum_{n=0}^{\infty} \left\{ F^{(m)}(x + n \ell_d) - F^{(m)}(n \ell_d) \right\}, & \text{if } 0 < x \leq \ell_d, \\
1, & \text{if } x > \ell_d.
\end{cases}
\tag{2}
\]

The form of $\pi^{(E)}$ can be written as
\[
\pi^{(E)} = \frac{1}{\sum_{n=0}^{\infty} u_n}.
\tag{3}
\]

Here, the term $u_n$ is defined as
\[
u_n \triangleq \int_{n \ell_d}^{\infty} dF^{(m)}(x) = 1 - F^{(m)}(n \ell_d),
\quad \text{for } n = 1, 2, \cdots,
\tag{4}
\]
with $u_0 = 1$, and we regard that $u_{n+1}/u_n = 0$ when $u_n = 0$.

Letting $\ell^{(p)}$ be the mean packet size, from (1) and assumption of A1, the form of $\ell^{(p)}$ is given by
\[
\ell^{(p)} \triangleq \mathbb{E}[L] = \pi^{(E)} \ell^{(m)}.
\tag{5}
\]

Let $\{\sigma^{(p)}\}^2$ be the variance of packet sizes. From (1) and assumption of A2, the variance of packet sizes $\{\sigma^{(p)}\}^2$ is given by
\[
\{\sigma^{(p)}\}^2 = \mathbb{E}[L^2] - \mathbb{E}[L]^2
= \pi^{(E)} \left( \{\ell^{(m)}\}^2 + \{\sigma^{(m)}\}^2 \right) + 2 \pi^{(E)} \ell^{(m)} \ell_d \\
- 2 \pi^{(E)} \ell_d \left( \sum_{n=0}^{\infty} v_n - \ell_d \sum_{n=1}^{\infty} n u_n \right) - \left\{ \ell^{(p)} \right\}^2,
\tag{6}
\]
where
\[
v_n \triangleq \int_{n \ell_d}^{\infty} x dF^{(m)}(x),
\quad \text{for } n = 1, 2, \cdots,
\tag{7}
\]
with $v_0 = \ell^{(m)}$.

The detailed derivations of (1) to (7) can be found in [13].

Let a random variable $X$ be the number of packets created from a corresponding message. Note that the random variable $X$ is identified with a batch size, which will be introduced in Section IV.

The forms of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ are given by
\[
\mathbb{E}[X] = \sum_{n=0}^{\infty} n \Pr(X = n) = \sum_{n=1}^{\infty} n (u_{n-1} - u_n) = \sum_{n=0}^{\infty} u_n \left( \frac{1}{\pi^{(E)}} \right),
\tag{8}
\]
\[
\mathbb{E}[X^2] = \sum_{n=0}^{\infty} n^2 \Pr(X = n) = \sum_{n=1}^{\infty} n^2 (u_{n-1} - u_n) = \sum_{n=0}^{\infty} (2n + 1) u_n.
\tag{9}
\]
IV. MEAN RESPONSE TIME

In this paper, we employ the response time $R$ as the performance. To derive the form of $E[R]$, we represent a communication network as a queueing system as shown in Fig. 3. The response time $R$ is defined as the total time from a packet arrival at the queue until its service completion. In the communication network, a service time $S$ in a queueing system where the server is a physical link is equal to transmission time, which is given by $\ell/\mu_c$ where a data unit is $\ell$ in size and $\mu_c$ is the capacity of the physical link. The response time $R$ is interpreted as the waiting time $W$, i.e., time spent in the queue alone, plus the service time $S$.

For analytical tractability, we introduce the following assumptions:

B1: packets created from a corresponding message, referred to as a batch, arrive at the queueing system simultaneously at a time.

B2: the batches arrive at the queueing system according to Poisson process with mean arrival rate $\lambda$.

B3: the maximum number of packets that can be accommodated in the queueing system is infinite.

B4: packets are served in FIFO scheduling dripline.

B5: offered load $a = \lambda E[X] E[S]$ satisfies $a < 1$.

B6: the size of SGM-RAS-layer’s control information is constant and equal to $\ell_h$.

From assumptions of B1 – B4, we can treat the queueing system as an $M^X/G/1$ queueing model in the Kendall notation, where $X$ is the batch size, which is defined as the number of packets simultaneously arriving at the queueing system.

From assumption of B5, the $M^X/G/1$ becomes stable.

By solving the $M^X/G/1$ queueing model, we have

\[
E[R] = E[S] + E[R], \quad (10)
\]
\[
E[S] = E[L] + \ell_h = \frac{\ell(p) + \ell_h}{\mu_c}, \quad (11)
\]
\[
E[S^2] = E[(L + \ell_h)^2] = \frac{\sigma^2(p)^2 + (\ell(p) + \ell_h)^2}{\mu_c^2}, \quad (12)
\]
\[
E[W] = \frac{\lambda \left( E[X] E[S^2] + (E[X^2] - E[X]) E[S]^2 \right)}{2(1 - a)} + \frac{\ell_d + \ell_h}{\mu_c} (E[X] - 1). \quad (13)
\]

The derivations of (10) – (13) can be found in Appendix.
V. NUMERICAL RESULTS AND DISCUSSIONS

In this section, we examine the effect of payload size on mean response time by utilizing the results in Section IV. We consider a scenario in which Web objects are transferred over the IEEE 802.11g, i.e., physical link capacity \( \mu_c = 54 \) Mbps and control information field size \( \ell_h = 38 \) bytes.

The sizes of the Web objects are assumed to follow a lognormal distribution given by

\[
F^{(m)}(x) = \begin{cases} 
\int_{y=0}^{y=x} \frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{-(\log y - \mu_w)^2}{2\sigma_w^2}} dy, & x > 0, \\
0, & x \leq 0.
\end{cases}
\]  

(14)

The distribution parameters \( \mu_w \) and \( \sigma_w \) are assumed to be 6.34 and 2.07, respectively, on the basis of the measured mean message size \( \ell^{(m)} = 4827 \) bytes and the measured standard deviation \( \sigma^{(m)} = 41,008 \) bytes [14]. Note that this lognormal distribution can represent a long-tailed property.

Figure 4 shows mean response time \( R \) versus payload sizes \( \ell_d \) for different batch arrival rates \( \lambda \). This figure demonstrates the mean response time \( R \) is convex in payload sizes \( \ell_d \).

The reason for this is as follows.

- **When payload size \( \ell_d \) is large enough**

From (4), \( u_n \) for \( n = 1, 2, \ldots \) is approximately zero, resulting in \( \pi^{(E)} \approx 0 \). Hence, message segmentations hardly occur.

In addition, it yields \( E[X] \approx 1 \) and \( E[X^2] \approx 1 \). Hence, since batch size is approximately one, the queuing system can be approximated as an M/G/1 queueing model.

The mean waiting time \( E[W] \) for the M/G/1 queueing model is given by

\[
E[W] \approx \frac{\lambda E[S]^2}{2(1 - \lambda E[S])} = \frac{\lambda \{E[S]\}^2}{2} \left( \frac{\sqrt{E[S^2] - \{E[S]\}^2}}{E[S]} \right)^2 + 1.
\]  

(15)
From the above equation, we find that mean response time \( R \) depends on the coefficient of variation of \( S \), \( c_S \), that is, \( \sqrt{E[S^2] - \{E[S]\}^2}/E[S] \). If \( c_S \), equivalently the variance of message sizes in proportional, increases, \( R \) increases. In this example, the value of \( c_S \) is very high because the message size distribution exhibits long-tailed property.

* When payload size \( \ell_d \) is small
  When payload size decreases, the number of segmented packets per message increases, resulting in the smaller value of \( c_S \). Hence, the mean response time \( R \) may be smaller if payload size is smaller even though the mean batch size increases.

* When payload size \( \ell_d \) is small enough
  In this case, the segmented packets, of which size is \( \ell_d \) is dominant in all created packets. Therefore, the queueing system can be approximated as an \( M^X/D/1 \) queueing model. Although the value of \( c_S \) is almost zero, the mean batch size is too large, resulting in large mean response time.

From Fig. 4, we find that the mean response time is more convex in payload sizes if batch arrival rate is larger.

VI. CONCLUSION

This paper proposed the \( M^X/G/1 \) queueing model to discuss the effect of payload size on mean response time when message segmentations occur. We derived the closed form of mean response time, given payload size, message size distribution and message arrival rate. From a numerical result, we have demonstrated that the mean response time is more convex in payload sizes if message arrival rate is larger, in a scenario where Web objects are delivered over a physical link.

The remaining issues include the clarification of the relationship among several parameters to express the optimized payload size, the extension of our model to noisy links, and development of payload adaptation algorithm to minimize the mean response time.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP15K00139.

APPENDIX

This appendix provides derivations of (10) – (13).

Equations (10) – (12) are clear from the definitions of \( R \) and \( S \).

To derive (13), we observe an arbitrary packet in a batch (i.e., message), referred to as the tagged packet, and divide the waiting time \( W \) into two parts (see Fig. 5):
waiting time $W_1$: the time from when the first packet in a batch have the tagged packet arrives at the queueing system to when it enters service.

waiting time $W_2$: the time from when the first packet enters service to when the tagged packet enters service.

From the argument of [12, Chapter 4], the form of $E[W_1]$ is given by

$$E[W_1] = \frac{\lambda (E[X] E[S^2] + (E[X^2] - E[X]) \{E[S]\})}{2(1 - a)}.$$  \hfill (16)

Noting that one message consists of consecutive $(n - 1)$ body packets and an edge packet when it is divided to $n$ packets (see Fig. 2), we have

$$E[W_2] = \frac{\ell_d + \ell_h}{\mu_c} \sum_{n=1}^{\infty} (n - 1) Pr.(X = n) = \frac{\ell_d + \ell_h}{\mu_c} \left( \sum_{n=1}^{\infty} n Pr.(X = n)n - \sum_{n=1}^{\infty} Pr.(X = n) \right)$$

$$= \frac{\ell_d + \ell_h}{\mu_c} (E[X] - 1),$$  \hfill (17)

because of $Pr.(X = 0) = 0$. From (16) and (17), we obtain (13).

REFERENCES

[1] H. Kobayashi and B. Mark, System Modeling and Analysis: Foundations of System Performance Evaluation. Prentice Hall, 2008.
[2] M. Schwartz, Telecommunication Networks: Protocols, Modeling and Analysis. Addison-Wesley Publishing Company, 1987.
[3] A. S. Tanenbaum, Computer Networks, 3rd ed. Prentice Hall PTR, 1996.
[4] G. Montenegro, S. Davkins, M. Kojo, V. Magret, and N. Vaidya, Long Thin Networks. Network Working Group Request for Comments, Jan. 2000, no. RFC 2757.
[5] S. Davkins, G. Montenegro, M. Kojo, and V. Magret, End-to-end Performance Implications of Slow Links. Network Working Group Request for Comments, July 2001, no. RFC 3150.
[6] W. R. Stevens, TCP/IP Illustrated, Volume 1: The Protocols. Addison-Wesley Publishing Company, 1994.
[7] B. P. Crow, I. Widjaja, J. G. Kim, and P. T. Sakai, “IEEE 802.11 wireless local area networks,” IEEE Communications Magazine, vol. 35, no. 9, pp. 116–126, Sept. 1997.
[8] 3rd Generation Partnership Project, “Technical Specification Group Radio Access Network; Radio Link Control (RLC) protocol specification,” [Online]. Available: http://www.3gpp.org/, 2005.
[9] P. Lettieri and M. Srivastava, “Adaptive frame length control for improving wireless link throughput, range, and energy efficiency,” in Proc. IEEE INFOCOM’98, 1998, pp. 564–571.
[10] E. Modiano, “An adaptive algorithm for optimizing the packet size used in wireless ARQ protocols,” Journal Wireless Networks, vol. 5, no. 4, pp. 279–286, July 1999.
[11] P. R. Jelenkovic and J. Tan, “Dynamic packet fragmentation for wireless channels with failures,” in Proc. ACM MobiHoc’08, 2008, pp. 73–82.
[12] H. Akiyama and K. Kawashima, Teletraffic: Theory and Applications. Springer-Verlag, 1993.
[13] T. Itohada, Y. Kishi, and Y. Takahashi, “Data-unit-size distribution model when message segmentations occur,” Performance Evaluation, vol. 69, no. 1, pp. 1–16, Jan. 2012.
[14] M. Molina, P. Castelli, and G. Fedd. “Web traffic modeling exploiting TCP connections’ temporal clustering through HTML-REDUCE,” IEEE Network, vol. 14, no. 3, pp. 46–55, May/June 2000.