Effect of payload size on mean response time when message segmentations occur: Case of burst packet arrival

Takashi Ikegawa*
Waseda University
Tokyo, Japan
ikegawa@aoni.waseda.jp

ABSTRACT

This paper investigates the effect of payload size on mean response time for communication networks where message segmentations occur and segmented packets within a message arrival over a physical link in bursts. To achieve this, we develop an $M^X/G/1$ queue that represents multiple segmented packets per arrival and a correlated packet size sequence within a message, given a payload size and a message-size distribution. From numerical results, we show that the curves of the mean response time for a packet in the case of low offered load conditions are concave in payload sizes for an actual scenario where Web objects are delivered over a physical link. In addition, we derive the analytical form of a payload size yielding the minimum response time in the case of no offered load conditions when message sizes are exponentially distributed, which is approximately given by $\sqrt{2 \ell^{(m)} \ell^{(h)}/3}$ where $\ell^{(m)}$ is the mean message size and $\ell^{(h)}$ is the header/trailer size.

CCS CONCEPTS

• Computing methodologies → Modeling methodologies.

KEYWORDS

Message segmentation, payload size, mean response time, burst packet arrival, $M^X/G/1$ queue, correlated service time.

ACM Reference Format:

Takashi Ikegawa. 2019. Effect of payload size on mean response time when message segmentations occur: Case of burst packet arrival. In *12th EAI International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2019)*, March 12–15, 2019, Palma, Spain. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3306309.3306313>

1 INTRODUCTION

Multiple controllable (manageable) parameters including the maximum permitted packet size, namely payload size, and window size affect many quality of service (QoS) characteristics for users

of packet-based transfer networks. Optimization of these parameters against QoS characteristics such as mean response time and goodput [7, Section 1.4], [6, Section 1.2] is the most important issue for network design and management (e.g., [10] for traditional packet-based transfer networks, [3] for wireless networks, and [4] for industrial wireless sensor networks).

Among the control parameters, the payload size, which restricts the maximum size of packets, i.e., data units transferred over physical links, is one of the essential ones. The reasons for this include that the link-level transmission delay is simply given by the packet size divided by link capacity [7, pp. 64–65] and packet-corruption probability is approximately proportional to the packet size [10, p. 132].

Sizes of messages, that is data units generated by applications such as when transferring content of Web pages, are frequently larger than the payload size. To convey such messages over a network, a sender implements a message-segmentation function, which is specified, for example, in TCP/IP and the IEEE 802.11 media access control protocol.

The message-segmentation function divides a single message into multiple segments of data when it is larger than the payload size. Furthermore, the sender creates a packet by adding an appropriate header and/or trailer to each data segment.

To optimize payload size when corrupted packets are retransmitted by error-recovery functions, the following tradeoff conditions when goodput is employed as a QoS parameter should be taken into consideration [2, 8]. Here, goodput is defined as the number of bits successfully received by a receiver per unit time.

- If the payload is oversized, packet corruption can occur, which necessitates considerable packet retransmissions. As a result, goodput becomes very low.
- If the payload size is smaller than the previous one, there is a decrease in packet-corruption probability, equivalent to the mean number of retransmissions required for successful packet delivery. Therefore, goodput becomes higher in comparison.
- If the payload is undersized, the network would be operating inefficiently because the overhead, that is the header/trailer, for each packet is redundant, even though packet corruption is less likely to occur. Thus, goodput becomes low.

Hence, there is a need to determine the optimal payload size to maximize goodput.

In situations where message segmentations occur, the following tradeoff conditions when the mean response time for a packet is employed as a QoS parameter should be taken into consideration. The response time for a packet is defined as the time from arrival

*He is also with Graduate School of Mathematical Sciences, the University of Tokyo, Japan.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

VALUETOOLS 2019, March 12–15, 2019, Palma, Spain

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6596-3/19/03.

<https://doi.org/10.1145/3306309.3306313>

at a buffer to completion of the transmission. It consists of waiting time in the buffer and transmission time over a physical link.

- If the payload is oversized, message segmentations cannot occur, and the mean waiting time due to the burstiness of packet arrivals is small. However, mean transmission time is larger than that with message segmentations.
- If the payload size is smaller than the previous one, the mean number of segmented packets per message increases. As a result, mean waiting time due to the burstiness of packet arrivals is larger. However, mean transmission time is smaller because mean packet size is smaller due to message segmentations. Consequently, if the payload size is not undersized, mean response time decreases because the rate of increase for mean waiting time is smaller compared to that of decrease for mean transmission time.
- If the payload is undersized, the mean number of segmented packets per message increases significantly, resulting in too large waiting time compared to mean transmission time because the burstiness of packet arrivals is too large and the utilization of a physical link increases due to the packet overhead increment within a message. Therefore, mean response time is too large.

From the above discussion, a tradeoff between mean response time and payload size when message segmentations occur will exist.

In previous studies, the latter tradeoff has not been discussed in depth. Especially, to our knowledge, the analysis for this tradeoff has not been performed where a payload size and a message-size distribution are explicitly given and message segmentations occur. The purpose of this paper is to investigate this tradeoff by using a batch (bulk) arrival queueing system.

The batch arrival queueing system enables us to capture the behavior of burstiness of arriving segmented packets. However, traditional batch arrival queueing systems, e.g., [1], [11, Section 9.3], made unrealistic assumptions such as independent service times (i.e., packet-transmission times) although service times are correlated when message segmentations occur. To solve this problem, we propose a batch arrival queueing system that represents realistic behaviors using the segmented-packet-size sequence property.

The main contributions of this paper are as follows.

- (1) We propose an $M^X/G/1$ queue to represent message-segmentation behavior.
- (2) Using the $M^X/G/1$ queue to represent an actual scenario where Web objects are transferred over a physical link, we show that the curves of mean response time for a packet are concave under low offered load conditions.
- (3) We indicate that the payload size to minimize the mean response time in the case of no offered load conditions is approximately given by $\sqrt{2\ell^{(m)}\ell^{(h)}/3}$ where $\ell^{(m)}$ is the mean message size and $\ell^{(h)}$ is the header/trailer size when message sizes are exponentially distributed with mean $\ell^{(m)}$.

The rest of the paper is organized as follows. In Section 2, we describe the communication network model underlying our study. In Section 3, we analyze the segmented-packet-size sequence. In Section 4, we derive the forms of stationary distributions of packet sizes and the number of segmented packets per message using the segmented-packet-size sequence model described in Section 3. In

Figure 1: Communication network model.

Section 5, we derive the closed form of the mean response time using an $M^X/G/1$ queue. In Section 6, we investigate the effect of payload size on the mean response time when message sizes are distributed lognormally and exponentially. Finally, in Section 7, we summarize this paper and mention future work.

2 COMMUNICATION NETWORK MODEL

In this paper, we consider a simple two-layered communication network model to investigate the effect of payload size on packet-level performance when message segmentations occur. First, in this section, we describe our two-layered model. Next, we explain the assumptions made for analytical tractability.

2.1 Layer model

The communication network under consideration is shown in Fig. 1. Each station (a sender and a receiver) has two layers: application and segmentation-reassembly (SGM-RAS).

Application layer: it contains a traffic source and sink. The traffic source generates data units, which will be referred to as *messages*.

SGM-RAS layer: it implements a message SGM-RAS function. This function enables a single message to be divided into multiple segments if the message size is larger than the payload size, denoted by $\ell^{(d)}(> 0)$. In addition, this layer creates data units, which are referred to as *packets*, by encapsulating the (segmented) data into an information field and adding the appropriate protocol control information (PCI), such as a header and/or trailer. Furthermore, it transmits the packets over physical links at a sender.

2.2 Assumptions

For analytical tractability, we make the following assumptions related to message and PCI sizes.

- A1:** Let $L_i^{(m)} > 0$ denote a size of the i ($= 1, 2, \dots$)th message generated in the sender's application layer. A sequence of random variables $\{L_i^{(m)} : i = 1, 2, \dots\}$ is mutually independent and identically distributed (i.i.d.) in accordance with a common message-size distribution function, denoted by $F^{(m)}(\cdot)$.

Figure 2: Creation of packets through message segmentation.

A2: The message-size distribution $F^{(m)}(\cdot)$ has a finite mean $\ell^{(m)}$ and a finite variance $\{\sigma^{(m)}\}^2$.

A3: The size of the PCI is constant and equal to $\ell^{(h)}$.

In actual communication networks, the corrupted/lost packets are retransmitted by an error-recovery function as mentioned in the previous section. In this paper, packets are assumed to be not lost.

3 PACKET-SIZE SEQUENCE MODEL

The creation of packets from a message through message segmentation is shown in Fig. 2. If $L_i^{(m)} > \ell^{(d)}$, then the i th message is divided into multiple packets. As the number of packets created from the i th message, X_i is given by $\lceil L_i^{(m)} / \ell^{(d)} \rceil$.¹

Denoting the size of an information field of the j th packet for the i th message by $L_{ij}^{(p)} (> 0)$, where $i = 1, 2, \dots$ and $j = 1, 2, \dots, X_i$, we have

$$L_{ij}^{(p)} = \begin{cases} \ell^{(d)}, & j = 1, 2, \dots, X_i - 1, \\ L_i^{(m)} - (X_i - 1)\ell^{(d)}, & j = X_i. \end{cases} \quad (1)$$

We categorize the packets within the i th message into two kinds: body and edge packets. A segmented packet appearing between the head and the penultimate packets, i.e., the packet from first to $(X_i - 1)$ st is a *body* packet, and the final packet, i.e., the X_i th packet, is an *edge* packet.

If $L_i^{(m)} \leq \ell^{(d)}$, then the i th message is not segmented (i.e., $X_i = 1$). In this case, a single packet, whose information field is identical to the original message, is created. This is also an edge packet because the size of the information field for such a packet, that is $L_{i1}^{(p)}$, satisfies (1).

We note that the size of body packets is always equal to $\ell^{(d)} + \ell^{(h)}$, whereas that of edge packets is variable but does not exceed $\ell^{(d)} + \ell^{(h)}$ (see Fig. 2).

4 STATIONARY DISTRIBUTIONS OF PACKET SIZES AND NUMBER OF SEGMENTED PACKETS PER MESSAGE

In this section, we derive the forms of a stationary packet-size distribution and a stationary distribution of the number of segmented packets per message, using the packet-size sequence model mentioned in the previous section. Statistical parameters obtained from

¹The operator $\lceil a \rceil$ represents the smallest integer that is greater than or equal to a .

Figure 3: Sketch of message-size distribution, packet-size distribution, and edge-packet occurrence probability.

these distributions, such as mean and variance, are available for the derivation of mean response time described in Section 5.

4.1 Stationary packet-size distribution

We construct a random variable sequence $\{L_{\kappa}^{(p)} : \kappa = 1, 2, \dots\}$, replacing the pair of epoch labels i, j with a packet in-sequence number $\kappa = 1, 2, \dots$ for $\{L_{ij}^{(p)}\}$ in (1).

Denoting the stationary packet-size distribution, i.e., stationary distribution of $\{L_{\kappa}^{(p)} + \ell^{(h)} : \kappa = 1, 2, \dots\}$ by $F^{(p)}(\cdot)$, we have

$$\begin{aligned} F^{(p)}(x) &\triangleq \Pr(L_{\kappa}^{(p)} \leq x - \ell^{(h)}) \\ &= (1 - \pi^{(E)})\mathbf{1}(x - (\ell^{(d)} + \ell^{(h)})) + \pi^{(E)}F^{(E)}(x), \end{aligned} \quad (2)$$

where $\pi^{(E)}$ is an edge-packet occurrence probability and $F^{(E)}(\cdot)$ is a stationary edge-packet size distribution (see Fig. 3).²

We define u_n for $n = 0, 1, 2, \dots$ as:

$$u_n \triangleq \int_{n\ell^{(d)}}^{\infty} dF^{(m)}(x) = 1 - F^{(m)}(n\ell^{(d)}), \quad n = 1, 2, \dots, \quad (3)$$

with $u_0 = 1$. We regard that $u_{n+1}/u_n = 0$ when $u_n = 0$.

Then, the forms of $\pi^{(E)}$ and $F^{(E)}(\cdot)$ are given by

$$\pi^{(E)} = \frac{1}{\sum_{n=0}^{\infty} \int_{n\ell^{(d)}}^{\infty} dF^{(m)}(x)} = \frac{1}{\sum_{n=0}^{\infty} (1 - F^{(m)}(n\ell^{(d)}))} = \frac{1}{\sum_{n=0}^{\infty} u_n}, \quad (4)$$

$$F^{(E)}(x) = \begin{cases} 0, & 0 \leq x < \ell^{(h)}, \\ \sum_{n=0}^{\infty} (F^{(m)}(x + n\ell^{(d)} - \ell^{(h)})) \\ \quad - F^{(m)}(n\ell^{(d)} - \ell^{(h)}), & \ell^{(h)} \leq x \leq \ell^{(d)} + \ell^{(h)}, \\ 1, & x > \ell^{(d)} + \ell^{(h)}. \end{cases} \quad (5)$$

²An indicator function $\mathbf{1}(x)$ in (2) is defined as $\mathbf{1}(x) \triangleq 1$ if $x \geq 0$ and 0 if $x < 0$.

Let $\ell^{(p_i)}$ be the mean of stationary distribution of $\{L_\kappa^{(p_i)}\}$, that is the mean information field size. From the assumptions of A1 and A2, $\ell^{(p_i)}$ is finite. From (2)–(5), we have

$$\ell^{(p_i)} = \int_0^\infty x dF^{(p)} - \ell^{(h)} = \pi^{(E)} \ell^{(m)}. \quad (6)$$

The detailed derivations of (2)–(6) can be found in [5].

4.2 Stationary distribution of number of created packets per message

Denote the stationary distribution of $\{X_i : i = 1, 2, \dots\}$ by $\beta_n (\triangleq \Pr(X_i = n))$ where $n = 1, 2, \dots$. From (3), the form of β_n is given by

$$\begin{aligned} \beta_n &= \Pr((n-1)\ell^{(d)} < L_i^{(m)} \leq n\ell^{(d)}) \\ &= F^{(m)}(n\ell^{(d)}) - F^{(m)}((n-1)\ell^{(d)}) = u_{n-1} - u_n. \end{aligned} \quad (7)$$

Let $\bar{\beta}$ and $\bar{\beta}^{(2)}$ be the mean and the second moment of the random variable X_i with the stationary distribution $\{\beta_n\}$, respectively. From (4) and (7), we have

$$\bar{\beta} = \sum_{n=1}^{\infty} n \beta_n = \sum_{n=0}^{\infty} u_n = \frac{1}{\pi^{(E)}}, \quad (8)$$

$$\bar{\beta}^{(2)} = \sum_{n=1}^{\infty} n^2 \beta_n = \sum_{n=0}^{\infty} (2n+1) u_n. \quad (9)$$

5 MEAN RESPONSE TIME

Let a random variable R_κ be the response time for the κ th packet, which is one of the important QoS characteristics. To derive the form of the expectation $E[R_\kappa]$, we represent a communication network as a single-server queueing system as illustrated in Fig. 4.

The server in the queueing system corresponds to a physical link. A service time of the κ th customer (i.e., a packet), denoted by S_κ , is equal to the transmission time of a packet. Thus, we have

$$S_\kappa = \frac{L_\kappa^{(p_i)} + \ell^{(h)}}{\mu_c}, \quad (10)$$

where μ_c is the capacity of the physical link.

The response time R_κ is defined as the total time from the κ th packet arrival at the queue until its service completion. It is interpreted as the waiting time, denoted by W_κ , i.e., time spent in the queue alone, plus the service time S_κ .

Figure 4: Queueing system.

To represent arrivals of segmented packets, we assume that the segmented packets within a message arrive at a physical link simultaneously, implying the *batch* arrival.

For analytical tractability, we introduce the following assumptions:

- B1:** the batches arrive at the queueing system in accordance with a Poisson process with arrival rate λ ,
- B2:** the maximum number of packets that can be accommodated in the queueing system is infinite,
- B3:** packets are transmitted, i.e., served, in FIFO (First-In, First-Out) scheduling discipline, and
- B4:** offered load (or traffic intensity) ρ satisfies $\rho < 1$.³

From the assumptions of B1–B3, we can treat the queueing system as an $M^X/G/1$ queue, where X implies the batch size.

From the assumption of B4, the $M^X/G/1$ queue becomes stable.

Batch arrival queues, i.e., $M^X/G/1$ queues, have been analyzed by many researchers including Burke [1] and Tijms [11, Section 9.3]. However, most of those queues, which are referred to as “ordinary $M^X/GI/1$ queues”,⁴ assume that

- the service times are independent and identically distributed random variables, referred to as an i.i.d. service time assumption, and
- the customers within the same batch are served in a random scheduling discipline, i.e., random scheduling assumption.

In this paper, we analyze the proposed batch arrival queue in a different way using the basic results on $M/G/1$ queues.

In actual systems, the assumptions of the ordinary $M^X/GI/1$ queues will be violated. For example, the sizes of packets within a message that cause the segmentation are correlated (see (1)). Thus, the actual service times of customers for the same batch will not be an i.i.d. In addition, the actual service order within a batch is in accordance with the FIFO scheduling discipline rather than a random one (see assumption B3).

To overcome these problems, we propose an $M^X/G/1$ queue that represents the correlated packet sizes and the packet transmissions in accordance with the FIFO scheduling discipline.

We divide the random variable W_κ into two independent random variables $W_\kappa^{(m)}$ and $W_\kappa^{(p)}$ for the κ th packet (see Fig. 5):

³In a single-server queue, ρ becomes the utilization of a server.

⁴We use “GI” (general independent distribution) rather than “G” (general distribution) to note a service time distribution, because service times are assumed to be independently distributed.

Figure 5: Relation between $W_\kappa^{(m)}$ and $W_\kappa^{(p)}$.

$W_\kappa^{(m)}$: waiting time for a batch, which is measured from the arrival time of the batch until the transmission start time of the *first* packet for the batch,

$W_\kappa^{(p)}$: waiting time for the κ th packet, namely the test packet, which is measured from the start time of the first packet of the batch having the test packet until that of the transmission for the test packet.

By solving the proposed $M^X/G/1$ queue, we have

$$E[R_\kappa] = E[S_\kappa] + E[W_\kappa], \quad (11)$$

$$E[S_\kappa] = \frac{\ell^{(p)} + \ell^{(h)}}{\mu_c}, \quad (12)$$

$$E[W_\kappa] = E[W_\kappa^{(m)}] + E[W_\kappa^{(p)}], \quad (13)$$

$$E[W_\kappa^{(m)}] = \frac{\lambda}{2(1-\rho)} \times \frac{\left[\{\ell^{(m)}\}^2 + \{\sigma^{(m)}\}^2 + 2\bar{\beta} \ell^{(m)} \ell^{(h)} + \bar{\beta}^{(2)} \{\ell^{(h)}\}^2 \right]}{\mu_c^2}, \quad (14)$$

$$\rho = \frac{\lambda(\ell^{(m)} + \bar{\beta} \ell^{(h)})}{\mu_c}, \quad (15)$$

$$E[W_\kappa^{(p)}] = \frac{(\ell^{(d)} + \ell^{(h)}) (\bar{\beta} - 1)}{2\mu_c}. \quad (16)$$

The derivations of (11)–(16) can be found in Appendix A.

Example 5.1. Case of exponential message-size distribution.

We assume that message sizes are independently and exponentially distributed with mean $\ell^{(m)}$:

$$F^{(m)}(x) = 1 - e^{-\frac{x}{\ell^{(m)}}}, \quad x > 0. \quad (17)$$

Substituting (17) into (3), we have

$$u_n = e^{-\frac{n \ell^{(d)}}{\ell^{(m)}}} = u^n, \quad n = 0, 1, 2, \dots, \quad (18)$$

with $u \triangleq e^{-\frac{\ell^{(d)}}{\ell^{(m)}}}$.

In this case, the terms appearing in (12), (14), (15), and (16), that is $\ell^{(p)}$, $\sigma^{(m)}$, $\bar{\beta}$, and $\bar{\beta}^{(2)}$, can be simplified as:

$$\ell^{(p)} = (1-u) \ell^{(m)}, \quad \sigma^{(m)} = \ell^{(m)}, \quad (19)$$

$$\bar{\beta} = \frac{1}{1-u}, \quad \bar{\beta}^{(2)} = \frac{1+u}{(1-u)^2}. \quad (20)$$

6 NUMERICAL RESULTS AND DISCUSSIONS

In this section, we examine the effect of payload size $\ell^{(d)}$ on mean response time $E[R_\kappa]$ by utilizing the results described in Section 5. First, we explain the assumptions for the numerical results. Next, the effect of $\ell^{(d)}$ on $E[R_\kappa]$ is investigated when Web objects (i.e., messages) whose sizes are represented by a lognormal distribution are transferred. Finally, we consider the scenario where message sizes are exponentially distributed, which enables us to derive the closed form of $E[R_\kappa]$.

Figure 6: Mean response time versus payload sizes for different batch arrival rates in case of lognormal message sizes.

6.1 Assumptions

We consider a scenario where packets are transferred over an IEEE 802.11g LAN, i.e., physical link capacity $\mu_c = 54$ Mbps. In addition, we assume that the size of PCI $\ell^{(h)}$ is 38 bytes.

6.2 Case of lognormal message sizes

The sizes of the Web objects have a long-tailed nature. A lognormal distribution is well known for representing long-tailed characteristics [6, p. 34]. The message sizes are assumed to follow the lognormal distribution given by

$$F^{(m)}(x) = \int_{y=0}^{y=x} \frac{1}{\sqrt{2\pi}\sigma_w y} e^{-\frac{(\log y - \mu_w)^2}{2\sigma_w^2}} dy, \quad x > 0. \quad (21)$$

The distribution parameters μ_w and σ_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 [9].

Figure 6 shows mean response time $E[R_\kappa]$ versus payload sizes $\ell^{(d)}$ for different batch arrival rates λ . In this example, we used the values 0.1, 0.5, and 1.0 [messages/msec] for λ . In this figure, we also plot the mean response times obtained from the ordinary $M^X/GI/1$ queues, denoted by $E[\tilde{R}_\kappa]$, to evaluate the error caused by the i.i.d. service time assumption (see Remark 1). The brief explanation for the derivation of $E[\tilde{R}_\kappa]$ can be found in Appendix B.

From Fig. 6, we find the following features:

- C1:** as payload size $\ell^{(d)}$ tends to infinity, the value of $E[R_\kappa]$ approaches a finite value,
- C2:** the value of $E[R_\kappa]$ increases when payload size $\ell^{(d)}$ decreases in a small payload-size region, and
- C3:** the curves of $E[R_\kappa]$ exhibit a convex property against payload sizes $\ell^{(d)}$ under low offered load conditions.

To discuss the features of C1 and C2 in depth, we consider the cases of payloads that are oversized and undersized.

• Case of oversized payloads

In this case, since message segmentations cannot occur, we have:

- the edge-packet occurrence probability $\pi^{(E)}$ is approximately one because u_n for $n = 1, 2, \dots$ is almost zero. Hence, the sizes of an information field for each packet can be almost identified with that of an entire message, and
- the stationary distribution $\{\beta_1, \beta_2, \beta_3, \dots\}$ for a set of batch sizes $\{X_i\}$ can be approximated by $\{1, 0, 0, \dots\}$. As a result, the mean $\bar{\beta}$ and the second moment $\bar{\beta}^{(2)}$ for $\{X_i\}$ are almost one from (8) and (9).

Consequently, the queueing system for oversized payloads $\ell^{(d)}$ can be modeled as an M/G/1 queue where a service time is the transmission time of a data unit, which consists of an entire message and PCI.

From the analysis of M/G/1 queues and the above arguments, we have

$$\lim_{\ell^{(d)} \rightarrow \infty} E[R_K] = E[S_K] + \frac{\lambda E[S_K^2]}{2(1 - \lambda E[S_K])}, \quad (22)$$

where $E[S_K]$ and $E[S_K^2]$ are given by

$$E[S_K] = \frac{\ell^{(m)} + \ell^{(h)}}{\mu_c}, \quad (23)$$

$$E[S_K^2] = \frac{\{\sigma^{(m)}\}^2 + \{\ell^{(m)} + \ell^{(h)}\}^2}{\mu_c^2}. \quad (24)$$

The values of $E[R_K]$ for oversized payloads are finite because $E[S_K]$ and $E[S_K^2]$ are finite due to the assumptions of A2 and A3.

• Case of undersized payloads

From $\lim_{\ell^{(d)} \rightarrow 0} u_n = \lim_{\ell^{(d)} \rightarrow 0} \int_n^\infty dF^{(m)}(x) = 1$ for $n = 0, 1, 2, \dots$, using (8), we have

$$\lim_{\ell^{(d)} \rightarrow 0} \bar{\beta} = \lim_{\ell^{(d)} \rightarrow 0} \sum_{n=0}^{\infty} u_n = \sum_{n=0}^{\infty} \lim_{\ell^{(d)} \rightarrow 0} u_n \rightarrow \infty. \quad (25)$$

Applying (25) to (14) and (16), we obtain

$$\lim_{\ell^{(d)} \rightarrow 0} E[W_K^{(m)}] \rightarrow \infty \quad \text{and} \quad \lim_{\ell^{(d)} \rightarrow 0} E[W_K^{(p)}] \rightarrow \infty. \quad (26)$$

From (26), we find that mean waiting time $E[W_K] (= E[W_K^{(m)}] + E[W_K^{(p)}])$ is too large when payload size $\ell^{(d)}$ is too small. As a result, mean response time $E[R_K]$ is too large.

To clarify the feature of C3, we investigate the effect of payload size $\ell^{(d)}$ on mean response time $E[R_K]$ under low offered load conditions. Figure 7 shows mean response time $E[R_K]$ versus payload sizes $\ell^{(d)}$ for different batch arrival rates $\lambda = 0.001, 0.01$, and 0.1 [messages/sec] in addition to no offered load. This figure demonstrates that the mean response time $E[R_K]$ has the minimum value.

To explain this, we derive the form of $E[R_K]_{\lambda=0}$, which is defined as (virtual) mean response time $E[R_K]$ of no offered load, where $\lambda = 0$. The form of $E[R_K]_{\lambda=0}$ is given by $E[S_K] + E[W_K^{(p)}]$ because of $E[W_K^{(m)}] = 0$. Using (6) and (8), the form that is a function of

Figure 7: Mean response time versus payload sizes for different batch arrival rates in case of lognormal message sizes.

edge-packet occurrence probability $\pi^{(E)}$ can be rewritten as:

$$\begin{aligned} E[R_K]_{\lambda=0} &= E[S_K] + E[W_K^{(p)}] \\ &= \frac{\pi^{(E)} \ell^{(m)} + \ell^{(h)}}{\mu_c} + \left(\frac{\ell^{(d)} + \ell^{(h)}}{2\mu_c} \right) \left(\frac{1 - \pi^{(E)}}{\pi^{(E)}} \right). \end{aligned} \quad (27)$$

From (27), we observe that the curves of $E[R_K]_{\lambda=0}$ exhibits a convex property because

- $\pi^{(E)}$ increases monotonically in payload size $\ell^{(d)}$,⁵
- $E[S_K]$, the first term of the right side of (27), increases linearly in edge-packet occurrence probability $\pi^{(E)}$, and
- $E[W_K^{(p)}]$, the second term of the right side of (27), exhibits a convex property in $\pi^{(E)}$.

Note that (27) implies that a tradeoff between $E[S_K]$ and $E[W_K^{(p)}]$ against $\pi^{(E)}$ exists.

REMARK 1. Approximation error using ordinary M^X/GI/1 queues. From Fig. 6, the analysis using the ordinary M^X/GI/1 queues yields considerable errors for large payload sizes, because it cannot represent situations where service times are variable and correlated. However, the difference of mean response times between i.i.d. and actual (correlated) service times is small for small payload sizes because packet sizes are almost constant, i.e., body packets are dominant in transmitted packets.

6.3 Case of exponential message sizes

In this subsection, we consider the case of exponential message sizes. We denote mean response time in this case by $E[\hat{R}_K]$. Figure 8 shows the mean response time in the case of exponential message sizes $E[\hat{R}_K]$ versus payload sizes $\ell^{(d)}$ for different batch arrival rates

⁵This is verified in the case of exponential messages because $\frac{\partial \pi^{(E)}}{\partial \ell^{(d)}} = (e^{-\ell^{(d)}/\ell^{(m)}})/\ell^{(m)} > 0$ for any $\ell^{(d)}(> 0)$ (see Example 5.1). The proof for the general message-size distributions is an open issue.

Figure 8: Mean response time versus payload sizes for different batch arrival rates in case of exponential message sizes.

λ . We used the value of 4137 bytes for $\ell^{(m)}$, which is the same as in the case of lognormal message sizes, and the values of 0, 0.001, 0.01, and 0.1 [messages/msec] for λ . From this figure, we find that the tendency of mean response time in the case of exponential message sizes is similar to that of log-normal message sizes.

Let $\hat{\ell}^{(d)}$ be the payload size that yields the minimum mean response time, that is $\arg \min_{\ell^{(d)}} E[\hat{R}_K]$. As shown in Fig. 8, the values of $\hat{\ell}^{(d)}$ depends on the batch rates λ . Even though the closed form of $E[\hat{R}_K]$, which has a λ parameter, is available from the argument of Example 5.1, the closed form is too complex. We then attempt to derive the closed form of $\hat{\ell}^{(d)}$ in the case of no offered load (i.e., $\lambda = 0$), denoted by $\hat{\ell}_{\lambda=0}^{(d)}$.

Denote the mean response time $E[\hat{R}_K]$ in the case of $\lambda = 0$ by $E[\hat{R}_K]_{\lambda=0}$. Substituting $\pi^{(E)} = 1 - u$ into (27), we have

$$E[\hat{R}_K]_{\lambda=0} = \frac{(1-u)\ell^{(m)} + \ell^{(h)}}{\mu_c} + \frac{u(\ell^{(d)} + \ell^{(h)})}{2(1-u)\mu_c}. \quad (28)$$

To find the value of $\hat{\ell}_{\lambda=0}^{(d)}$, by differentiating (28) with respect to $\ell^{(d)}$ (noting that $u = e^{-\frac{\ell^{(d)}}{\ell^{(m)}}}$), we obtain

$$\frac{\partial E[\hat{R}_K]_{\lambda=0}}{\partial \ell^{(d)}} = \frac{e^{-\frac{\ell^{(d)}}{\ell^{(m)}}}}{\mu_c} \left[1 + \frac{\ell^{(m)} \left(1 - e^{-\frac{\ell^{(d)}}{\ell^{(m)}}} \right) - \ell^{(d)} - \ell^{(h)}}{2\ell^{(m)} \left(1 - e^{-\frac{\ell^{(d)}}{\ell^{(m)}}} \right)^2} \right]. \quad (29)$$

A necessary condition for $\hat{\ell}_{\lambda=0}^{(d)} = \arg \min_{\ell^{(d)}} E[\hat{R}_K]$ to satisfy is $\frac{\partial E[\hat{R}_K]_{\lambda=0}}{\partial \ell^{(d)}} = 0$. The right side of (29) represents a nonlinear equation. However, the value of $\hat{\ell}_{\lambda=0}^{(d)}$ can be solved using numerical techniques. In this example, $\hat{\ell}_{\lambda=0}^{(d)}$ is 366 bytes.

By substituting $e^{-\frac{\ell^{(d)}}{\ell^{(m)}}} \approx 1 - \frac{\ell^{(d)}}{\ell^{(m)}} + \frac{1}{2} \left(\frac{\ell^{(d)}}{\ell^{(m)}} \right)^2$ into (29), the approximate $\hat{\ell}_{\lambda=0}^{(d)}$ is given by

$$\hat{\ell}_{\lambda=0}^{(d)} \approx \sqrt{\frac{2\ell^{(m)}\ell^{(h)}}{3}}, \quad (30)$$

which is 350 bytes in the example.

7 CONCLUSION

This paper investigated the effect of payload size on mean response time for a packet from the viewpoint of message segmentation. To achieve this, we proposed an $M^X/G/1$ queue that represents the burst arrival behavior of segmented packets and correlated packet-transmission times within a message. It enables us to derive an analytical form of the mean response time, given payload size, message-size distribution, and message arrival rate. From numerical results, we found that the curves of the mean response time are convex in payload sizes under low offered load conditions. Furthermore, we indicated that the payload size to minimize the mean response time under no offered load conditions is approximately given by $\sqrt{2\ell^{(m)}\ell^{(h)}/3}$ where $\ell^{(m)}$ is the mean message size and $\ell^{(h)}$ is the header/trailer size when message sizes are exponentially distributed with mean $\ell^{(m)}$.

The remaining issues include the extension of our model to multiple traffic sources and noisy physical links, and development of a simple and cost-effective payload adaptation algorithm in a dynamic network environment.

ACKNOWLEDGMENTS

This work was in part supported by JSPS KAKENHI Grant Number JP15K00139. The author would like to thank Dr. Atsushi Nakayasu for his valuable discussions in deriving (30).

REFERENCES

- [1] P. J. Burke. 1975. Delays in Single-Server Queues with Batch Input. *Operations Research* 23, 4 (Aug. 1975), 830–833. <https://doi.org/10.1287/opre.23.4.830>
- [2] G. Ciccarese, M. D. Blasi, P. Marra, C. Palazzo, and L. Patrono. 2011. An algorithm for controlling packet size in IEEE 802.16e networks. *Computer Networks* 55, 13 (2011), 2873 – 2885. <https://doi.org/doi.org/10.1016/j.comnet.2011.06.010>
- [3] B. Fu, Y. Xiao, H. Deng, and H. Zeng. 2014. A Survey of Cross-Layer Designs in Wireless Networks. *IEEE Communications Surveys & Tutorials* 16, 1 (2014), 110 – 126. <https://doi.org/10.1109/SURV.2013.081313.00231>
- [4] V. C. Gungor and G. P. Hancke. 2009. Industrial Wireless Sensor Networks: Challenges, Design Principles, and Technical Approaches. *IEEE Transactions on Industrial Electronics* 56, 10 (Oct 2009), 4258–4265. <https://doi.org/10.1109/TIE.2009.2015754>
- [5] T. Ikegawa, Y. Kishi, and Y. Takahashi. 2012. Data-Unit-Size Distribution Model when Message Segmentations Occur. *Performance Evaluation* 69, 1 (Jan. 2012), 1–16. <https://doi.org/10.1016/j.peva.2011.05.004>
- [6] H. Kobayashi and B.L. Mark. 2008. *System Modeling and Analysis: Foundations of System Performance Evaluation*. Prentice Hall.
- [7] J. Kurose and K. Ross. 2016. *Computer Networking: A Top-Down Approach* (seventh ed.). Pearson Education Limited.
- [8] E. Modiano. 1999. An Adaptive Algorithm for Optimizing the Packet Size Used in Wireless ARQ Protocols. *Journal Wireless Networks* 5, 4 (July 1999), 279–286. <https://doi.org/10.1023/A:1019111430288>
- [9] M. Molina, P. Castelli, and G. Foddis. 2000. Web Traffic Modeling Exploiting TCP connections' Temporal Clustering through HTML-REDUCE. *IEEE Network* 14, 3 (May/June 2000), 46–55. <https://doi.org/10.1109/65.844500>
- [10] M. Schwartz. 1987. *Telecommunication Networks: Protocols, Modeling and Analysis*. Addison-Wesley Publishing Company.
- [11] H. C. Tijms. 2003. *A First Course in Stochastic Models*. John Wiley & Sons, Ltd.

A DERIVATIONS OF (11)–(16)

This appendix provides the derivations of (11)–(16). As described in [1], [11, Section 9.3], $M^X/G/1$ queues can be easily solved using the arguments of analysis for $M/G/1$ queues.

- **Derivations of (11)–(13)**

Equations (11)–(13) are clear from the definitions of R_K , S_K , and W_K .

- **Derivations of (14) and (15)**

- **Preliminary**

To derive the forms of (14) and (15), we introduce a random variable $L_n^{(tp)}$, which is defined as the total size of all consecutive packets when n (segmented) packets are created from a message, i.e., $X_i = n$. The form of $L_n^{(tp)}$ can be expressed as:

$$L_n^{(tp)} = \sum_{j=1}^n L_{ij}^{(p)} + n \ell^{(h)} = L_i^{(m)} + n \ell^{(h)}. \quad (31)$$

because the total size of information fields for all packets within a message $\sum_{j=1}^n L_{ij}^{(p)}$ is equal to $L_i^{(m)}$. Then, $E[L_n^{(tp)}]$ and $E[\{L_n^{(tp)}\}^2]$ are given by

$$E[L_n^{(tp)}] = \ell^{(m)} + n \ell^{(h)}, \quad (32)$$

$$E[\{L_n^{(tp)}\}^2] = \{\ell^{(m)}\}^2 + \{\sigma^{(m)}\}^2 + 2n \ell^{(m)} \ell^{(h)} + n^2 \{\ell^{(h)}\}^2. \quad (33)$$

- **Equation (14)**

From the arguments of $M/G/1$ queues, e.g., [6, Section 4.3.3], $E[W_K^{(m)}]$ is given by

$$E[W_K^{(m)}] = \frac{\lambda E[\{S^{(b)}\}^2]}{2(1-\rho)}, \quad (34)$$

where $S^{(b)}$ is the total time needed to serve all customers (or packets) for a batch (or a message).

From the definition of $\{L_n^{(tp)}\}^2$, the form of $E[\{S^{(b)}\}^2]$ in (34) is given by

$$\begin{aligned} E[\{S^{(b)}\}^2] &= \sum_{n=1}^{\infty} \Pr(X_i = n) E[\{S^{(b)}\}^2 | X_i = n] \\ &= \sum_{n=1}^{\infty} \beta_n E[\{L_n^{(tp)}\}^2] \\ &= \frac{\sum_{n=1}^{\infty} \beta_n E[\{L_n^{(tp)}\}^2]}{\mu_c^2}. \end{aligned} \quad (35)$$

Substituting (35) into (34) and using (8), (9), and (33), we obtain (14).

- **Equation (15)**

From the definition of offered load ρ , we have

$$\rho = \lambda E[S^{(b)}]. \quad (36)$$

From the definition of $L^{(tp)}$, the form of $E[S^{(b)}]$ is given by

$$\begin{aligned} E[S^{(b)}] &= \sum_{n=1}^{\infty} \Pr(X_i = n) E[S^{(b)} | X_i = n] \\ &= \frac{\sum_{n=1}^{\infty} \beta_n E[L_n^{(tp)}]}{\mu_c}. \end{aligned} \quad (37)$$

Substitution of (37) into (36) and usage of (33) with (8) yields (15).

- **Derivation of (16)**

Noting that

- one message consists of consecutive $(n-1)$ body packets and an edge packet when it is divided into n packets (see Fig. 2),
- the size of a body packet is $\ell^{(d)} + \ell^{(h)}$, and
- the test packet that is in the r th position of a message (or a batch) has to wait for the transmission of the $(r-1)$ body packets,

we have

$$\begin{aligned} E[W_K^{(p)}] &= \sum_{n=1}^{\infty} \Pr(X_i = n) E[W_K^{(p)} | X_i = n] \\ &= \sum_{n=1}^{\infty} \beta_n \frac{\sum_{r=1}^n (r-1) (\ell^{(d)} + \ell^{(h)})}{n \mu_c} \\ &= \sum_{n=1}^{\infty} \beta_n \frac{(n-1) (\ell^{(d)} + \ell^{(h)})}{2 \mu_c}, \end{aligned} \quad (38)$$

which is (16).

B BRIEF DERIVATION OF MEAN RESPONSE TIME OBTAINED FROM ORDINARY $M^X/GI/1$ QUEUES

In this appendix, we provide the derivation of the mean response time obtained from ordinary $M^X/GI/1$ queues briefly due to the space limitation.

The terms that are affected by the i.i.d. service time assumption are the offered load and mean waiting time for a batch. We denote them by $\tilde{\rho}$ and $E[\tilde{W}_K^{(m)}]$, respectively.

Similar to Appendix A, to derive the form of $\tilde{\rho}$ and $E[\tilde{W}_K^{(m)}]$, we introduce a random variable $\tilde{L}_n^{(tp)}$, which is the total size of all consecutive packets within a message when $X_i = n$ under the i.i.d. service time assumption. Then, we have

$$E[\tilde{L}_n^{(tp)}] = n (\ell^{(p)} + \ell^{(h)}), \quad (39)$$

$$E[\{\tilde{L}_n^{(tp)}\}^2] = n^2 (\ell^{(p)} + \ell^{(h)})^2 + n \{\sigma^{(p)}\}^2, \quad (40)$$

where $\{\sigma^{(p)}\}^2$ is the variance of information field sizes for packets. The form of $\{\sigma^{(p)}\}^2$ can be found in [5, Proposition 8].