On Modeling, Analysis, and Optimization of Packet Aggregation Systems

Jung Ha Hong and Khosrow Sohraby, *Senior Member, IEEE*

***Abstract*—In packet communication systems, a header is attached to the transmitted packet at each layer. The overhead due to the transmission of the individual header can have a significant impact on the performance of the communication system especially when the system operates in heavy load. In order to increase data throughput, a number of packets sharing a single header can be aggregated into a frame.**

In this paper, we present a mathematical model for a packet aggregation system assuming a general distribution for the packet length. For a given header size, we obtain the minimum system utilization where packet aggregation improves the system performance. We also analyze the asymptotic behavior of such systems leading to a simple heuristic policy on the optimum aggregation level. It is shown that the impact of the variability of the packet length distribution on different system performance measures is rather insignificant when the system load is low or moderate.

***Index Terms*—Asymptotic analysis, batch service queues, delay optimization, framing, packet aggregation, performance bounds, queueing delay analysis.**

I. Introduction

A packet transmission via Open System Interconnection (OSI) layers [3, p. 41] assumes that a header is attached to the transmitted packet at each layer. The overhead due to the transmission of the individual header may have a significant impact on the performance of a packet communication system especially when the system operates in heavy load. In order to reduce the overhead and increase data throughput, a number of packets can be aggregated into a frame at the time of encapsulation when the bit error rate is not very high. Otherwise, in a very noisy environment, the cost of frame retransmission due to transmission error may offset any performance gain from packet aggregation, since a frame consisting of a number of packets needs to be retransmitted instead of a single packet.

Packet aggregation methods have been proposed for about a decade now [5], [17]. More recently, there has been renewed interest of their applications in wireless networks [10], [16] where fairness and inefficiency issues due to a small payload in 802.11-based wireless systems are examined.

Another example of packet aggregation in real systems is Frame Relay (FR) systems [3, p. 266], where every packet which is transmitted over the system is aggregated into the FR frame. FR also allows packing of small packets into a single FR frame as in the case of voice packet transmission over FR defined in FRF.11 [4]. In this case, every aggregated packet is called a sub-frame which consists of original packet and a sub-header. IEEE 802.16 standard for broadband wireless access systems [8] is also another example of packet aggregation. Here, the media access control layer operates with protocol data units (PDU's), each of which consists of a header and one or more service data units (SDU's) which represent PDU's payload. In other words, SDU's are aggregated in a single PDU.

In our model, we limit the maximum number of packets inserted into a frame and transmitted as a single entity (i.e., batch) by J. A header is appended to the frame before its transmission. Thus, a completed frame consists of j ($j = 1, \cdots, J$) packets and a header. The service time of a frame is considered as the sum of the service times of all individual packets and the header. As described in more details in the next section, the number of packets in a frame (batch) is dictated by the number of the packets arriving during the service time of the previous frame. Therefore, in our model, the service times of successive frames (batches) are dependent, which indicates a major difference from the existing queueing models.

In this paper, we present a mathematical model and its analysis of packet aggregation systems. Our model consists of a queueing system with batch service, where a complete performance analysis is provided. We note that the analytical model of such systems is new and its solution has not been reported in the open literature. A simple variation of the model was first reported by Bailey [1]. In his model, if the server finds less than J waiting on completion of a batch service, then it takes all of them in a batch for service. If it finds more than J waiting, then it takes a batch of size J for service, while others, in excess of J units, wait for service in the queue. He assumed that the intervals of time between successive occasions of service are independent and identically distributed. Neuts [14] studied the distribution of the busy period for the same system assuming that the service times of successive batches are conditionally independent given the batch sizes, but may depend on their batch sizes. Neuts [15] proposed the "general bulk service rule" in which service initiates only when at least a certain number, L, of packets are waiting in the queue and a batch has a maximum size J, ($L \leq J$). Jaiswal [9] considered a batch service queue in which the batch size is random. In all these batch service models, the statistical information of the batch size is a priori specified unlike our model which is a by-product of the aggregation system under the consideration.

One may think of the header in our system as a vacation in

Paper approved by T.-S. P. Yum, the Editor for Packet Access and Switching of the IEEE Communications Society. Manuscript received December 11, 2008; revised April 23, 2009.

The authors are with the Department of Computer Science Electrical Engineering, University of Missouri - Kansas City, 5100 Rockhill Road, Kansas City, MO 64110 USA (e-mail: {jhhk86, sohrabyk}@umkc.edu).

Digital Object Identifier 10.1109/TCOMM.2010.080667

gated/limited service systems with a single vacation reported by Takagi [18, p.202, p.227] which are not batch service models. However, in our model, all packets in a frame depart from the system together with a header. Therefore, such models do not apply to our packet aggregation system.

A brief and preliminary model of the aggregation systems was presented by the first author [6], [7], where a Phase-Type service time distribution for individual packets and an Erlang distribution for the header were considered. Furthermore, the analysis approach was based on the Markov chain methodology.

This paper gives a detailed analysis of the packet aggregation system assuming that the packet arrivals follow a Poisson process. Although the generalization to other (Markovian) arrival processes are possible in principle, such analysis are quite space-intensive and are not presented here. We assume a general packet (service) and header length distribution, where the Supplementary Variable Technique [2, p. 57] is used for the analysis. Therefore, constant header size is readily analyzed as a special case, which is the most common in packet communication systems. Using the proposed model, we provide the analysis of the end-to-end delay of a packet and the distribution of the frame size. It is numerically demonstrated that the system performance measures approach their asymptotic values quickly as the maximum frame size J increases. Thus, we consider the case of infinite J, which leads to a functional equation [13] satisfying the system generating function. The exact solution of this equation is simpler and provides a significant reduction in the numerical complexity of the solution compared to the finite system. Moreover, the asymptotic model yields simple and nevertheless accurate upper and lower bounds for the first order statistics of important system performance measures. In addition, a simple heuristic supported by numerical results is provided to determine the optimum level of aggregation in such systems.

The remainder of this paper is organized as follows. We provide the queueing model and the analysis for finite maximum frame size J in section II and the asymptotic analysis of the system for large J in section III. Section IV covers the heuristics for the selection of optimum level of aggregation to minimize the end-to-end packet delay. We present numerical results in section V and conclude this paper in section VI.

II. System Model and Analysis

We assume that packets arrive according to a Poisson process with rate λ to a single server queue. Arriving packets are stored in an infinite buffer until they are transmitted to the destination peer. The transmission over the data link layer is done frame by frame. Data packets received from the network layer can be aggregated in a single frame. An overhead packet (header) is then appended in front of the frame. Therefore, each frame consists of a header and a number of data packets. The maximum number of data packets which can be aggregated in a single frame is J. Whenever the system is ready for transmission and the buffer is not empty, a frame is created of packets currently residing in the buffer starting from the head of line packet. If the number of packets in the buffer is less than J, then all packets in the buffer form a frame. If the number of packets in the buffer is greater than J, then only the first J packets are inserted into a frame and other packets wait for the next available frame in the buffer. Packets arriving during the transmission of a frame cannot be added to the current transmitting frame but wait for the available transmission in the buffer. Upon arrival at a buffer having i, $(i \geq 0)$ packets waiting for transmission, the arriving packet waits in the queue until the current frame and $\lfloor i/J \rfloor$ frames are transmitted, where $\lfloor \cdot \rfloor$ denotes the largest integer not exceeding the argument. Upon arrival at an empty buffer with no current transmission, the single arriving packet itself forms a frame and its transmission starts immediately. This will ensure a work conserving system.

We consider the transmission time of a frame as the sum of the transmission times of all individual packets and the header. We assume that the service (transmission) times of the header and the packet have general distributions and are independent. Let Y_0 and Y_1 be the service times of the header and the packet, respectively. The distribution function $B_k(x)$ of the service time Y_k is given by

$$B_k(x) = 1 - \exp\left[-\int_0^x \eta_k(t)dt\right], \quad k = 0, 1,$$

where $\eta_k(t)$ is the intensity function. Let $\mathrm{h_n}(\mathrm{m_n})$ be the nth moment of the service time $Y_0(Y_1)$. Conditioning that the server is busy at time t, we define the state of the server by

$$\xi(t) = \begin{cases} 0 & \text{if server is transmitting header,} \\ 1 & \text{if server is transmitting packet.} \end{cases}$$

We also define

$$\begin{aligned} P_0(t) = &\ \text{Prob}\,[N_q(t) = 0, N_r(t) = 0], \\ P_{i,j,k}(x,t)dx = &\ \text{Prob}\,[N_q(t) = i, N_r(t) = j, \xi(t) = k, \\ &\qquad x < X_k(t) \leq x + dx], \\ &\qquad i \geq 0,\ 1 \leq j \leq J,\ k = 0, 1, \end{aligned}$$

where $N_q(t)$ is the number of packets in the queue at time t, $N_r(t)$ is the number of remaining packets to be served in the frame under service at time t, which includes the packet being served at time t, and $X_0(t)\,(X_1(t))$ is the elapsed service time of the header(packet) under service at time t. We then can write the following equations of the process by considering the transitions occurring in small time interval $\triangle t$:

$$\begin{aligned} P_0(t+\triangle t) &= P_0(t)(1-\lambda\triangle t) \\ &\quad +(1-\lambda\triangle t)\int_0^\infty P_{0,1,1}(x,t)\eta_1(x)dx\triangle t + o(\triangle t), \\ P_{i,j,k}(x+\triangle t, t+\triangle t) &= (1-\lambda\triangle t)(1-\eta_k(x)\triangle t)P_{i,j,k}(x,t) \\ &\quad +\lambda\triangle t P_{i-1,j,k}(x,t)(1-\eta_k(x)\triangle t) + o(\triangle t), \\ &\quad i \geq 0,\ 1 \leq j \leq J,\ k = 0, 1, \end{aligned}$$

where $P_{-1,j,k}(x,t) = 0$. When $\triangle t \to 0$, we get the following differential equations:

$$\begin{aligned} \frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \int_0^\infty P_{0,1,1}(x,t)\eta_1(x)dx \\ \frac{\partial P_{i,j,k}(x,t)}{\partial x} &+ \frac{\partial P_{i,j,k}(x,t)}{\partial t} = -(\lambda + \eta_k(x))P_{i,j,k}(x,t) \\ &\quad + \lambda P_{i-1,j,k}(x,t), \quad i \geq 0, 1 \leq j \leq J, k = 0, 1. \end{aligned}$$

We note that $\rho + \lambda \mathrm{h}_1/J$ must be strictly less than one for the existence of an equilibrium solution, where $\rho = \lambda \mathrm{m}_1$. Let

$$P_0 \triangleq \lim_{t\to\infty} P_0(t),$$
$$P_{i,j,k}(x) \triangleq \lim_{t\to\infty} P_{i,j,k}(x,t), \quad i \geq 0,\ 1 \leq j \leq J,\ k = 0,1.$$

Then the differential equations become the following:

$$\lambda P_0 = \int_0^\infty P_{0,1,1}(x)\eta_1(x)dx, \tag{1}$$

$$\frac{\partial P_{i,j,k}(x)}{\partial x} = -(\lambda + \eta_k(x))P_{i,j,k}(x) + \lambda P_{i-1,j,k}(x), \quad i \geq 0,\ 1 \leq j \leq J,\ k = 0,1. \tag{2}$$

These equations are to be solved under the boundary conditions

$$P_{0,1,0}(0) = \lambda P_0 + \int_0^\infty P_{1,1,1}(x)\eta_1(x)dx, \tag{3}$$

$$P_{0,j,0}(0) = \int_0^\infty P_{j,1,1}(x)\eta_1(x)dx, \quad 2 \leq j \leq J-1, \tag{4}$$

$$P_{i,J,0}(0) = \int_0^\infty P_{i+J,1,1}(x)\eta_1(x)dx, \quad i \geq 0, \tag{5}$$

$$P_{i,j,0}(0) = 0, \quad i \geq 1,\ 1 \leq j \leq J-1, \tag{6}$$

$$P_{i,j,1}(0) = \int_0^\infty P_{i,j+1,1}(x)\eta_1(x)dx + \int_0^\infty P_{i,j,0}(x)\eta_0(x)dx, \ i \geq 0, 1 \leq j \leq J-1, \tag{7}$$

$$P_{i,J,1}(0) = \int_0^\infty P_{i,J,0}(x)\eta_0(x)dx, \quad i \geq 0, \tag{8}$$

and the normalization condition

$$P_0 + \sum_{i,j,k} \int_0^\infty P_{i,j,k}(x)dx = 1. \tag{9}$$

Now we define

$$G_{j,k}(z;x) \triangleq \sum_{i=0}^\infty P_{i,j,k}(x)z^i, \quad 1 \leq j \leq J, \quad k = 0,1.$$

From (2), we then get

$$\frac{\partial G_{j,k}(z;x)}{\partial x} = [\lambda z - \lambda - \eta_k(x)]\, G_{j,k}(z;x), \quad 1 \leq j \leq J,\ k = 0,1, \tag{10}$$

and the boundary conditions give the recursive relations

$$G_{1,0}(z;0) = \lambda P_0 + \int_0^\infty P_{1,1,1}(x)\eta_1(x)dx,$$
$$G_{j,0}(z;0) = \int_0^\infty P_{j,1,1}(x)\eta_1(x)dx, \quad 2 \leq j \leq J-1,$$
$$G_{J,0}(z;0) = \sum_{i=0}^\infty z^i \int_0^\infty P_{i+J,1,1}(x)\eta_1(x)dx,$$
$$G_{j,1}(z;0) = \int_0^\infty G_{j,0}(z;x)\eta_0(x)dx + \int_0^\infty G_{j+1,1}(z;x)\eta_1(x)dx, \ 1 \leq j \leq J-1,$$
$$G_{J,1}(z;0) = \int_0^\infty G_{J,0}(z;x)\eta_0(x)dx.$$

(10) has the solution

$$G_{j,k}(z;x) = G_{j,k}(z;0)[1 - B_k(x)]e^{-\lambda(1-z)x}, \tag{11}$$

which yields

$$\int_0^\infty G_{j,k}(z;x)\eta_k(x)dx = G_{j,k}(z;0)b_k^*(\lambda - \lambda z), \tag{12}$$

where $b_k^*(\cdot)$ is the Laplace Stieltjes Transform of the service time distribution $B_k(x)$, $k = 0,1$. Let

$$x_i \triangleq \int_0^\infty P_{i,1,1}(x)\eta_1(x)dx, \quad 0 \leq i \leq J-1.$$

From (1) and (12), the recursive relations are represented by

$$G_{1,0}(z;0) = x_0 + x_1,$$
$$G_{j,0}(z;0) = x_j, \quad 2 \leq j \leq J-1,$$
$$G_{J,0}(z;0) = \frac{1}{z^J}\left[G_{1,1}(z;0)b_1^*(\lambda - \lambda z) - \sum_{i=0}^{J-1} x_i z^i\right], \tag{13}$$
$$G_{j,1}(z;0) = G_{j,0}(z;0)b_0^*(\lambda - \lambda z) + G_{j+1,1}(z;0)b_1^*(\lambda - \lambda z), \ 1 \leq j \leq J-1, \tag{14}$$
$$G_{J,1}(z;0) = G_{J,0}(z;0)b_0^*(\lambda - \lambda z). \tag{15}$$

From (14), we have

$$G_{j,1}(z;0) = [G_{1,1}(z;0) - x_0\alpha(z)][\beta(z)]^{j-1} - \alpha(z)\sum_{i=1}^{j-1} x_i[\beta(z)]^{j-i}, \quad 2 \leq j \leq J, \tag{16}$$

where $\alpha(z) = b_0^*(\lambda - \lambda z)$ and $\beta(z) = 1/b_1^*(\lambda - \lambda z)$. From (13), (15), and (16), we therefore get

$$G_{1,1}(z;0) = \frac{\beta(z)f_1(z) - z^J\left[f_2(z) + x_0\{\beta(z)\}^J\right]}{1 - \{z\beta(z)\}^J/\alpha(z)}, \tag{17}$$

where $f_1(z) = \sum_{i=1}^{J-1} x_i z^i$ and $f_2(z) = \sum_{i=1}^{J-1} x_i[\beta(z)]^{J-i+1}$. Applying the L'Hospital's rule into (17), we get

$$G_{1,1}(1;0) = \frac{(1-\rho)\left[J\sum_{i=0}^{J-1} x_i - \sum_{i=1}^{J-1} i x_i\right] + \rho x_0}{(1-\rho)J - \lambda \mathrm{h}_1}, \tag{18}$$

which is always non-negative for a stable system. Let

$$G_{j,k}(z) \triangleq \int_0^\infty G_{j,k}(z;x)dx, \quad 1 \leq j \leq J,\ k = 0,1.$$

From (11), we then have

$$G_{j,k}(z) = G_{j,k}(z;0)\frac{1 - b_k^*(\lambda - \lambda z)}{\lambda - \lambda z}, \ 1 \leq j \leq J, k = 0,1. \tag{19}$$

Applying the L'Hospital's rule into (19), we get

$$G_{j,0}(1) = \mathrm{h}_1 G_{j,0}(1;0), \quad 1 \leq j \leq J,$$
$$G_{j,1}(1) = \mathrm{m}_1 G_{j,1}(1;0), \quad 1 \leq j \leq J.$$

Let $H(z) \triangleq \sum_{j=1}^J G_{j,0}(z)$ and $F(z) \triangleq \sum_{j=1}^J G_{j,1}(z)$. Then from the recursive relations, we obtain

$$H(1) = \mathrm{h}_1 G_{1,1}(1;0),$$
$$F(1) = \mathrm{m}_1\left[x_0 + JG_{1,1}(1;0) + \sum_{i=1}^{J-1} i x_i - J\sum_{i=0}^{J-1} x_i\right], \tag{20}$$

where $G_{1,1}(1;0)$ is given in (18). The normalization condition (9) is represented by

$$P_0 + H(1) + F(1) = 1. \tag{21}$$

Let $\Xi(z)$ and $\Delta(z)$ be the numerator and denominator of $G_{1,1}(z;0)$ in (17), respectively. Then we have

$$\Delta(z) = 0 \implies z^J = b_0^*(\lambda - \lambda z)\{b_1^*(\lambda - \lambda z)\}^J. \tag{22}$$

The sequence x_i, $0 \le i \le J-1$ are obtained using the analyticity of $G_{1,1}(z;0)$ inside and on the unit circle. At each root of $\Delta(z)$ inside and on the unit circle, $\Xi(z)$ should vanish. Now, we consider the following iteration with $z_0 = 0$, for $0 \le k \le J-1$,

$$z_{n+1} = b_1^*(\lambda - \lambda z_n) b_0^*(\lambda - \lambda z_n)^{1/J} e^{i2\pi k/J}, \quad n \ge 0.$$

It can be easily shown that for each k, if $|z_n| \le 1$, then $|z_{n+1}| \le 1$. Therefore, the iteration provides J roots of (22) on $|z| \le 1$, which includes a root $z = 1$ when $k = 0$. Substituting $J-1$ roots when $1 \le k \le J-1$ (except $z = 1$) into $\Xi(z)$, we get $J-1$ linear equations in terms of x_i, $(0 \le i \le J-1)$. Solving these equations with the normalization condition (21), we obtain all J unknowns, x_i, $(0 \le i \le J-1)$, resulting in the complete set of x_i's and hence, all the state probabilities.

Now, we provide the frame size distribution in terms of the sequence x_i found above. Since the header is transmitted first within a frame, the stationary frame size is identical to the number of the remaining packets to be transmitted in the frame while the header is transmitted. Denote $\tilde{\pi}_n^J (1 \le n \le J)$ as the stationary probability that a frame consists of n packets. Then we can easily find that

$$\tilde{\pi}_n^J = \frac{G_{n,0}(1)}{H(1)} = \frac{G_{n,0}(1;0)}{G_{1,1}(1;0)},$$

which yields

$$\begin{aligned}
\tilde{\pi}_1^J &= \frac{x_0 + x_1}{G_{1,1}(1;0)}, \\
\tilde{\pi}_n^J &= \frac{x_n}{G_{1,1}(1;0)}, \quad 2 \le n \le J-1, \\
\tilde{\pi}_J^J &= 1 - \frac{\sum_{n=0}^{J-1} x_n}{G_{1,1}(1;0)}.
\end{aligned}$$

The mean frame size $\bar{J}$ can be also obtained as

$$\bar{J} = \sum_{n=1}^{J} n\tilde{\pi}_n^J = \frac{F(1)}{\mathrm{m}_1 G_{1,1}(1;0)} = \frac{\lambda}{G_{1,1}(1;0)}, \tag{23}$$

where $G_{1,1}(1;0)$ is given in (18).

Here, we obtain the Laplace Stieltjes Transform(LST) $W^*(s)$ of the queue waiting time distribution and the LST $S^*(s)$ of the distribution of the total time in the system of a packet by conditioning on the state that the packet sees at its arrival epoch. Assuming that there are i packets in the queue at the arrival epoch of a packet, the queue waiting time of a packet is the sum of the remaining service time of a frame in transmission and the service time of $\lfloor \frac{i}{J} \rfloor$ frames. Let $r_0^*(s) = \frac{1-b_0^*(s)}{s\mathrm{h}_1}$ $\left(r_1^*(s) = \frac{1-b_1^*(s)}{s\mathrm{m}_1}\right)$ be the LST of the remaining service time distribution of the header (packet) when sampled at a random point and

$$P_{i,j,k} \triangleq \int_0^\infty P_{i,j,k}(x)dx, \quad i \ge 0,\ 1 \le j \le J,\ k = 0, 1.$$

By the PASTA (Poisson Arrivals See Time Averages)[12, p. 71] property, $W^*(s)$ is obtained as

$$\begin{aligned}
W^*(s) &\triangleq E[e^{-sW_q}] \\
&= P_0 + \sum_{i\ge 0} \sum_{1\le j\le J} r_0^*(s)\,[b_1^*(s)]^j \left[b_0^*(s)\{b_1^*(s)\}^J\right]^{\lfloor \frac{i}{J} \rfloor} P_{i,j,0} \\
&\quad + \sum_{i\ge 0} \sum_{1\le j\le J} r_1^*(s)[b_1^*(s)]^{j-1} \left[b_0^*(s)\{b_1^*(s)\}^J\right]^{\lfloor \frac{i}{J} \rfloor} P_{i,j,1}.
\end{aligned}$$

The total time in the system of a tagged packet is the sum of its queue waiting time and the service time of its frame. The frame size is determined by the number of packets arriving during the queue waiting time. Assume that there are i packets in the queue, j remaining packets in a frame and the header is being transmitted at the arrival epoch of a tagged packet. Let $i = nJ + r$, for $n \ge 0$ and $r = 0, 1, \cdots, J-1$. Then the number of packets arriving during the queue waiting time of the packet has the probability generating function (PGF)

$$r_0^*(\lambda - \lambda z)[b_1^*(\lambda - \lambda z)]^{j+nJ}[b_0^*(\lambda - \lambda z)]^n.$$

Let u_x be the probability that the frame size is equal to x given this scenario. Then it is clear that u_x is equal to the coefficient of z^{x-r-1} in above PGF for $r+1 \le x \le J-1$ and $u_J = 1 - \sum_{x=r+1}^{J-1} u_x$. Let v_x be the probability that the frame size is equal to x, given that there are i packets in the queue, a packet is being transmitted, and there are j remaining packets in the frame at the arrival epoch of the tagged packet. Similarly, v_x is obtained as the coefficient of z^{x-r-1} in PGF

$$r_1^*(\lambda - \lambda z)[b_1^*(\lambda - \lambda z)]^{j-1+nJ}[b_0^*(\lambda - \lambda z)]^n$$

for $r+1 \le x \le J-1$ and $v_J = 1 - \sum_{x=r+1}^{J-1} v_x$. Therefore, $S^*(s)$ is obtained as

$$\begin{aligned}
S^*(s) &\triangleq E[e^{-sT}] \\
&= P_0 b_0^*(s) b_1^*(s) \\
&\quad + \sum_{n,r,j} \sum_{r+1\le x\le J} r_0^*(s)[b_1^*(s)]^{j+nJ+x}[b_0^*(s)]^{n+1} u_x P_{i,j,0} \\
&\quad + \sum_{n,r,j} \sum_{r+1\le x\le J} r_1^*(s)[b_1^*(s)]^{j-1+nJ+x}[b_0^*(s)]^{n+1} v_x P_{i,j,1}.
\end{aligned}$$

Thus, the mean queue waiting time $E[W_q]$ and the mean total time $E[T]$ in the system can be found as $-\frac{d}{ds}W^*(s)\big|_{s=0}$ and $-\frac{d}{ds}S^*(s)\big|_{s=0}$, respectively.

We remark that significant numerical challenges have been observed even in the case of exponential packet service time unless the maximum frame size J is small. The problem appears to be in finding x_i's, $(i = 0, \cdots, J-1)$ requiring the solution of a system of linear equations. This system becomes ill-conditioned for even moderate J. We have checked various performance measures such as mean frame size, mean queue length, and mean total delay of a packet and observed the quick convergence to their asymptotic values as J increases. Thus, in the next section, we consider the case of infinite J resulting in a functional equation [13] satisfied by the generating function of the sequence x_i. The solutions turn out to be significantly easier to represent and compute.

III. Asymptotic Analysis and Bounds

In this section, we consider the case of infinite J. Whenever a transmission is ready to take place in this system, all packets in the queue are aggregated in a single frame, which is transmitted with a header. Packets arriving during the transmission of a frame cannot be added to the current transmitting frame and will be inserted into the next frame. Upon arrival at an empty buffer with no current transmission, the single arriving packet itself forms a frame and its transmission starts immediately. We assume that ρ $(= \lambda \mathrm{m}_1)$ is strictly less than one for the stability of the system. This system is also formulated by the equations (1) and (2), the boundary conditions (3) to (8) except (5) and (8), and the normalization condition (9) with $J = \infty$. We define

$$G_k(z,w;x) = \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} P_{i,j,k}(x) z^i w^j, \quad k = 0, 1.$$

From (2), we then get the differential equation

$$\frac{\partial G_k(z,w;x)}{\partial x} = [\lambda z - \lambda - \eta_k(x)]\, G_k(z,w;x), \quad k = 0, 1,$$

which has the solution

$$G_k(z,w;x) = G_k(z,w;0)[1 - B_k(x)]e^{-\lambda(1-z)x}. \tag{24}$$

From the boundary conditions, we get

$$G_0(z,w;0) = \lambda P_0 w + \sum_{j=1}^{\infty} w^j \int_0^{\infty} P_{j,1,1}(x)\eta_1(x)dx, \tag{25}$$

$$G_1(z,w;0) = \frac{1}{w}\int_0^{\infty} G_1(z,w;x)\eta_1(x)dx \tag{26}$$
$$+\int_0^{\infty} G_0(z,w;x)\eta_0(x)dx - \sum_{i=0}^{\infty} z^i \int_0^{\infty} P_{i,1,1}(x)\eta_1(x)dx.$$

Let $F(z) \triangleq \sum_{i=0}^{\infty} x_i z^i$, where

$$x_i = \int_0^{\infty} P_{i,1,1}(x)\eta_1(x)dx, \quad i \geq 0.$$

From (24), we have

$$\int_0^{\infty} G_k(z,w;x)\eta_k(x)dx = G_k(z,w;0) b_k^*(\lambda - \lambda z), \ k = 0, 1,$$

and therefore (25) and (26) are represented by

$$G_0(z,w;0) = \lambda P_0(w-1) + F(w), \tag{27}$$

$$G_1(z,w;0) = \frac{1}{w} G_1(z,w;0) b_1^*(\lambda - \lambda z) + G_0(z,w;0) b_0^*(\lambda - \lambda z) - F(z). \tag{28}$$

Substituting (27) into (28), we get

$$G_1(z,w;0) = \frac{[\lambda P_0(w-1) + F(w)]b_0^*(\lambda - \lambda z) - F(z)}{1 - \frac{1}{w} b_1^*(\lambda - \lambda z)}. \tag{29}$$

Applying the L'Hospital's rule into (29), we get

$$G_1(1,1;0) = \frac{F'(1) - \lambda \mathrm{h}_1 F(1)}{\rho}, \tag{30}$$

where $F'(1) = \left.\frac{dF(z)}{dz}\right|_{z=1}$. Let $\Xi(z,w)$ and $\Delta(z,w)$ be the numerator and denominator of $G_1(z,w;0)$ in (29), respectively. Then we have

$$\Delta(z,w) = 0 \implies w = b_1^*(\lambda - \lambda z).$$

Now, we define $\sigma(z)$ on $|z| \leq 1$ as

$$\sigma(z) \triangleq b_1^*(\lambda - \lambda z).$$

It can be easily shown that $|\sigma(z)| \leq 1$ for $|z| \leq 1$. Since $G_1(z,w;0)$ is analytic on $\{(z,w) : |z| \leq 1, |w| \leq 1\}$, $\Xi(z,w)$ should vanish at $(z, \sigma(z))$, which provides the functional equation

$$F(z) = a(z) + b(z)F(\sigma(z)), \tag{31}$$

where $a(z) = \lambda P_0\{\sigma(z) - 1\} b_0^*(\lambda - \lambda z)$ and $b(z) = b_0^*(\lambda - \lambda z)$. Let $\sigma_k(z)$ be the kth iteration of $\sigma(z)$, i.e.,

$$\sigma_0(z) = z, \quad \sigma_{k+1}(z) = \sigma(\sigma_k(z)), \ k \geq 0.$$

It can be shown that $\sigma_k(z)$ converges to 1 as k goes to infinity for $|z| \leq 1$. Iterating (31) formally, we therefore obtain its solution as

$$F(z) = \sum_{k=0}^{\infty} a(\sigma_k(z)) \prod_{j=0}^{k-1} b(\sigma_j(z)) + F(1) \prod_{j=0}^{\infty} b(\sigma_j(z)). \tag{32}$$

From (31), we get $F'(1) = \frac{\lambda\{\rho P_0 + \mathrm{h}_1 F(1)\}}{1-\rho}$, which represents (30) as

$$G_1(1,1;0) = \frac{\lambda\{P_0 + \mathrm{h}_1 F(1)\}}{1 - \rho}. \tag{33}$$

Let $G_k(z,w) = \int_0^{\infty} G_k(z,w;x)dx, \ k = 0, 1$, then from (24), we have

$$G_k(z,w) = G_k(z,w;0)\frac{1 - b_k^*(\lambda - \lambda z)}{\lambda - \lambda z}, \quad k = 0, 1. \tag{34}$$

Applying the L'Hospital's rule into (34), we then get

$$G_0(1,1) = \mathrm{h}_1 G_0(1,1;0) \quad \text{and} \quad G_1(1,1) = \mathrm{m}_1 G_1(1,1;0).$$

From (27) and (33), we thus have

$$\begin{aligned} G_0(1,1) &= \mathrm{h}_1 F(1), \\ G_1(1,1) &= \frac{\rho\{P_0 + \mathrm{h}_1 F(1)\}}{1 - \rho}. \end{aligned} \tag{35}$$

Therefore, the normalization condition (9) is represented by $P_0 + G_0(1,1) + G_1(1,1) = 1$, which gives

$$F(1) = \frac{1 - \rho - P_0}{\mathrm{h}_1}. \tag{36}$$

Substituting $F(1)$ in (36) into (35), we obtain $G_1(1,1) = \rho$. Since $F(0) = \lambda P_0$ by definition of $F(z)$, we get P_0, which results in the complete stationary probabilities, by substituting $F(1)$ in (36) into (32) and putting $z = 0$ on both sides of (32):

$$P_0 = \frac{(1-\rho)\mathrm{C}_{\infty}}{\lambda \mathrm{h}_1 \left[1 + \sum_{k=0}^{\infty}\{1 - \sigma_{k+1}(0)\} b(\sigma_k(0))\mathrm{C}_{k-1}\right] + \mathrm{C}_{\infty}},$$

where $\mathrm{C}_{-1} = 1$ and $\mathrm{C}_k = \prod_{j=0}^{k} b(\sigma_j(0)), \ k \geq 0$.

The frame size distribution can be found by the similar argument used for finite J. Denote $\tilde{\pi}_n$ as the stationary probability that a frame consists of n packets, then

$$\tilde{\pi}_1 = \frac{x_0 + x_1}{F(1)}, \quad \tilde{\pi}_n = \frac{x_n}{F(1)}, \ n \geq 2,$$

and the mean frame size $\bar{J}$ is obtained as

$$\bar{J} = \sum_{n=1}^{\infty} n\tilde{\pi}_n = \frac{x_0 + F'(1)}{F(1)} = \frac{\lambda}{F(1)},$$

where $F(1)$ is given in (36).

Let $Q(z)$ be the probability generating function of the number of packets in the queue. Then $Q(z)$ is obtained as

$$Q(z) \triangleq \sum_{n=0}^{\infty} P[L_q = n] z^n = P_0 + G_0(z,1) + G_1(z,1)$$

and the mean queue length is obtained as $E[L_q] = \frac{dQ(z)}{dz}\Big|_{z=1}$. It can be shown that $E[L_q]$ is represented by

$$E[L_q] = \frac{\lambda^2 \mathrm{h}_1(2\mathrm{h}_1\mathrm{m}_1 + \mathrm{m}_2) + \lambda \mathrm{h}_2(1-\rho-P_0)}{2\mathrm{h}_1(1-\rho^2)}. \quad (37)$$

By Little's law [12, p. 47], we also get the mean queue waiting time $E[W_q]$ as $E[L_q]/\lambda$. We note that $P_0 = 0\,(P_0 = 1)$ gives the upper (lower) bound for $E[L_q]$.

For infinite J, the queue waiting time of a packet is the same as the remaining service time of a frame in transmission and the total time in the system of a packet is the sum of its queue waiting time and the service time of its frame, which can be derived by conditioning on the state of the system at its arrival epoch. Let

$$P_{i,j,k} \triangleq \int_0^{\infty} P_{i,j,k}(x)dx, \quad i \geq 0,\ j \geq 1,\ k = 0, 1.$$

By the PASTA property [12, p. 71], the LST $\mathcal{W}^*(s)$ of the queue waiting time distribution and the LST $\mathcal{S}^*(s)$ of the total time distribution in the system of a packet are obtained as

$$\begin{aligned}
\mathcal{W}^*(s) &\triangleq E[e^{-s\mathcal{W}_q}] \\
&= P_0 + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} r_0^*(s)[b_1^*(s)]^j P_{i,j,0} + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} r_1^*(s)[b_1^*(s)]^{j-1} P_{i,j,1} \\
&= P_0 + r_0^*(s)G_0(1, b_1^*(s)) + \frac{r_1^*(s)}{b_1^*(s)} G_1(1, b_1^*(s)), \\
\mathcal{S}^*(s) &\triangleq E[e^{-s\mathcal{T}}] \\
&= P_0 b_0^*(s) b_1^*(s) + \sum_{k=0}^{1} c_k G_k(b_1^*(s), b_1^*(s) b_1^*(\lambda - \lambda b_1^*(s))),
\end{aligned}$$

where $c_0 = b_0^*(s)b_1^*(s)r_0^*(s)r_0^*(\lambda - \lambda b_1^*(s))$ and $c_1 = \frac{b_0^*(s)r_1^*(s)r_1^*(\lambda - \lambda b_1^*(s))}{b_1^*(\lambda - \lambda b_1^*(s))}$. Thus, we have the mean queue waiting time $E[\mathcal{W}_q]$ and the mean total time $E[\mathcal{T}]$ in the system as $-\frac{d}{ds}\mathcal{W}^*(s)\big|_{s=0}$ and $-\frac{d}{ds}\mathcal{S}^*(s)\big|_{s=0}$, respectively. It can be shown that these means are represented by

$$\begin{aligned}
E[W_q] &= \frac{\lambda \mathrm{h}_1 \mathrm{m}_2 + 2{\mathrm{h}_1}^2\rho + (1-\rho-P_0)\mathrm{h}_2}{2(1-\rho^2)\mathrm{h}_1}, \\
E[T] &= \mathrm{m}_1 + \frac{\mathrm{d} + (1+2\rho)(1-\rho-P_0)\mathrm{h}_2}{2(1-\rho^2)\mathrm{h}_1},
\end{aligned}$$

where $\mathrm{d} = 2(1+\rho+\rho^2){\mathrm{h}_1}^2 + (1+2\rho)\lambda \mathrm{h}_1 \mathrm{m}_2$. Here, we remark that as $\rho \to 0$, we have $E[\mathcal{W}_q] \to 0$ and $E[\mathcal{T}] \to \mathrm{m}_1 + \mathrm{h}_1$, as expected. Since $0 \leq P_0 \leq 1$, we obtain the upper and lower bounds for the mean total time $E[\mathcal{T}]$ as

$$U = \mathrm{m}_1 + \frac{\mathrm{d} + (1+2\rho)(1-\rho)\mathrm{h}_2}{2(1-\rho^2)\mathrm{h}_1}, \quad (38)$$

$$L = \mathrm{m}_1 + \frac{\mathrm{d} - \rho(1+2\rho)\mathrm{h}_2}{2(1-\rho^2)\mathrm{h}_1}. \quad (39)$$

We note that the exact P_0 approach to zero (one) as ρ goes to one (zero). Therefore, we expect that the upper (lower) bound should be close to the exact value in heavy (light) traffic.

In the next section, we address the question of "aggregate or not to aggregate?" and the optimum level of aggregation.

IV. Simple Heuristics on Optimum Level of Aggregation

Recall that the stability condition for a finite system is $\rho + \lambda \mathrm{h}_1/J < 1$, where $\rho = \lambda \mathrm{m}_1$, $\mathrm{m}_1(\mathrm{h}_1)$ denotes the mean packet (header) size, J is the maximum number of packets in a frame, i.e., referred to the level of aggregation, and λ denotes the arrival rate of packets to the system. It is clear that for sufficient system utilization ρ, a system with no aggregation ($J = 1$) may be unstable depending on the header size. But the stability (finite delay) is achieved as the level of aggregation J increases. In fact, we must have $J > \frac{\lambda \mathrm{h}_1}{1-\rho}$. Therefore, the aggregation process acts as a system "stabilizer" in the heavy load. It is also possible to reduce the packet end-to-end delay significantly for marginally stable systems (ρ close to unity) when we aggregate.

We now address two fundamental questions. The first one is what the minimum system utilization is for a given (fixed) header size where aggregation improves the system performance. The second one is what the optimum level of aggregation should be (i.e., finding the optimum J to minimize the average packet end-to-end delay).

We formulate the first question by considering the same average packet delay for the two extreme cases of $J = \infty$ (maximum aggregation) and $J = 1$ (no aggregation). Here, we use the simple upper bound for $E[\mathcal{T}]$ in (38) as an approximation for the average packet delay for $J = \infty$. We note that the case of $J = 1$ is a classical $M/G/1$ queue, where the service time S is the sum of the packet and the header transmission times. The overall system load (including the header) is $a = \lambda(\mathrm{m}_1 + \mathrm{h}_1)$ and $E[\mathcal{T}]$ is given as [11, p. 190]

$$E[\mathcal{T}] = E[S] + \frac{E[S^2]}{2E[S]} \frac{a}{1-a}. \quad (40)$$

We assume a deterministic header (i.e., $\mathrm{h}_2 = {\mathrm{h}_1}^2$) and express the second moment of the packet service time in terms of its squared coefficient of variation, i.e., $\mathrm{m}_2 = (1+c^2){\mathrm{m}_1}^2$. Equating (40) to the upper bound for $E[\mathcal{T}]$ in (38), we then get the following quadratic equation in h;

$$\begin{aligned}
\rho(1+\rho)(2+\rho)h^2 + \left\{-1 + \rho^2 + (1+2\rho)\rho^2(1+c^2)\right\} h \\
- (1-\rho)\rho^2(1+c^2) = 0, \quad (41)
\end{aligned}$$

where $h = \frac{\mathrm{h}_1}{\mathrm{m}_1}$ is the normalized header size. We note that h approaches to zero as the utilization ρ approaches to one. We solve (41) for h and present its non-negative solution as the following heavy traffic expansion:

$$h = \frac{1-\rho}{3} + \frac{2(1-\rho)^2}{9} + \frac{4(2+c^2)(1-\rho)^3}{27(1+c^2)} + O\left((1-\rho)^4\right). \quad (42)$$

We note that the first two terms in (42) are independent of c^2. As it will be demonstrated numerically in section V, the first two terms provide an excellent approximation for the minimum header size in the moderate to heavy traffic. The first three terms in (42) can be an excellent approximation for all range of utilizations and c^2. However, we may have to decide if aggregation should take place depending on the level of utilization ρ, since the header size is fixed in practice. It

TABLE I
AVERAGE FRAME SIZE $\bar{J}$ FOR ρ AND MAXIMUM FRAME SIZE J.

	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9	ρ=0.95	ρ=0.98
J=2	1.0742	1.2021	1.4252	1.8350	1.9917	unstable
J=5	1.0888	1.2718	1.6900	3.0524	4.0040	4.9837
J=7	1.0890	1.2739	1.7124	3.3548	4.7531	6.4830
J=9	1.0890	1.2742	1.7186	3.5088	5.2334	7.6844
J=12	1.0890	1.2743	1.7210	3.6207	5.6690	9.0512
J=15	1.0890	1.2743	1.7215	3.6716	5.9187	10.0388
J=17	1.0890	1.2743	1.7216	3.6901	6.0274	10.5462
J=18	1.0890	1.2743	1.7216	3.6967	6.0701	10.7645
J=∞	1.0890	1.2743	1.7216	3.7277	6.3825	13.5412

would be best to consider just the first two terms in (42) and solve for ρ as function of h. We then get

$$\rho^\star = \frac{1}{4}\left(7 - 3\sqrt{1+8h}\right), \tag{43}$$

which provides a simple heuristic to determine the level of utilization for which aggregation should take place for a given header size to improve the system performance, i.e., we aggregate if the system utilization $\rho > \rho^\star$. For example, if the normalized header size h is 0.1, we need a minimum utilization $\rho^\star = 0.744$ to warrant any level of aggregation. If $h = 0.05$, then $\rho^\star$ is increased to 0.863. We note that for $h > \frac{5}{9}$ (about 56%), $\rho^\star$ becomes negative and we always need to aggregate. This size of header is certainly unrealistic in practice.

The question of optimal J is much harder to be addressed analytically. We can either resort to (pure) numerical computations or use a heuristic approach supported by numerical experiments. We discuss the latter in this section.

Let's concentrate on the typical behavior of average packet delay vs. J. If the optimum level of aggregation $J^\star$ is 1 (i.e., no aggregation), then the average delay is monotonically increasing and reaches its asymptotic value for $J = \infty$. However, if $J^\star > 1$, then the average packet delay is a decreasing (increasing) function of J for $J \le J^\star$ ($J > J^\star$). We observe numerically that around $J^\star$, the the average packet delay $E[\mathcal{T}]$ is rather insensitive to J. This observation is useful since the "penalty" for not choosing the exact optimum aggregation level is minimal, as long as this level is in the neighborhood of the optimal value.

The optimum level of aggregation depends on two parameters, namely the system utilization ρ (excluding the header) and the packet service time variability c^2. Here, we propose the following heuristics. For a given (normalized) header size h, we do not aggregate (i.e., $J^\star = 1$) if the system utilization (excluding the header) is less than $\rho^\star$. Otherwise, the optimum aggregation level is given by the following heuristic:

$$J^\star = \begin{cases} \max\left(1, \left\lfloor \dfrac{1}{3(1-\rho)} \right\rfloor\right), & \rho \ge \rho^\star, 0 \le c^2 \le 1, \\ \max\left(1, \left\lfloor \dfrac{1+c^2}{6(1-\rho)} \right\rfloor\right), & \rho \ge \rho^\star, c^2 \ge 1, \end{cases} \tag{44}$$

where $\lfloor \cdot \rfloor$ denotes the largest integer not exceeding the argument.

In the next section, we discuss numerical results obtained from the analytical model.

TABLE II
HEADER SIZE FOR THE SAME AVERAGE TOTAL PACKET DELAY IN CASES OF MAXIMUM AGGREGATION ($J = \infty$) AND NO AGGREGATION.

	c^2=0.5	c^2=1	c^2=2	c^2=5	c^2=10
ρ=0.3	0.4241	0.4250	0.4267	0.4299	0.4325
ρ=0.5	0.2451	0.2457	0.2465	0.2478	0.2486
ρ=0.7	0.1240	0.1241	0.1243	0.1246	0.1248
ρ=0.9	0.0357	0.0357	0.0357	0.0357	0.0357
ρ=0.99	0.0034	0.0034	0.0034	0.0034	0.0034

V. NUMERICAL RESULTS

In this section, we use the constant header size for numerical results, which is most common in packet communications systems. We normalize the average packet size m_1 to 1. For a different distribution of the packet size, we use different squared coefficient of variation, c^2 (variance divided by mean squared). For $c^2 = 0.5$, Erlang-2 distribution is used. For $c^2 \ge 1$, we use hyperexponential distribution with parameters $p_1, p_2(= 1-p_1), \mu_1$, and μ_2 and assume $\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2}$ (i.e., the hyperexponential distribution has "balanced mean"), leaving us with two degrees of freedom for the determination of the parameters. As long as $c^2 \ge 1$, we can use arbitrary mean and variance (satisfying the obvious moment inequalities). In the case of $J = 1$, i.e., no packet aggregation, we have a classical $M/G/1$ queue, where the service time is the sum of the packet and the header transmission times.

Table I provides the average number of packets aggregated in a frame, $\bar{J}$, for various ρ ($= \lambda \mathrm{m}_1$) and the maximum frame size J using $\mathrm{h}_1 = 0.1$ and exponential packet size. Table I shows that $\bar{J}$ approaches quickly to the limiting case as J increases even if the system utilization ρ is high. It also shows that $\bar{J}$ is rather insensitive to the maximum frame size J for moderate utilization ($\rho = 0.5$ or 0.7). On the other hand, for high utilization ($\rho \ge 0.9$), $\bar{J}$ is rather large, implying that a considerable level of packet aggregation takes place. For light load ($\rho \le 0.5$), the reverse is true.

Table II provides the constant header size yielding the same average packet end-to-end delay for $J = \infty$ (maximum aggregation) and $J = 1$ (no aggregation). For $\rho = 0.9$ and $c^2 = 1$, packet aggregation will reduce the average total packet delay when the constant header size is larger than 0.0357, i.e., about 3.5% of the (average) packet length. In the very heavy load region of $\rho = 0.99$, a header size of 0.34% of a packet size is sufficient to justify packet aggregation. It is remarkable that the header size giving the same average packet delay for the two extreme cases is not sensitive to the variability of a packet transmission time. We note that the simple three-term heavy traffic approximation expressing h as function of ρ from (42) is accurate for all entries in Table II. It is observed that the simple approximation $\rho^\star$ in (43) obtained from the first two terms in (42) is also accurate for moderate to high utilization ($\rho \ge 0.5$) but the accuracy diminishes for lighter load. We have checked the variance of the total delay for the given header size in Table II. It is interesting to note that the variances are almost the same.

The goal of the next experiment is to investigate the degree of aggregation for $J = \infty$. Here, $\mathrm{h}_1 = 0.1$ is assumed. Table III shows the dependency of $\bar{J}$, the average number of packets in a frame, for different ρ and c^2. When $\rho = 0.9$, $\bar{J}$ is 46% higher for $c^2 = 10$ comparing to $c^2 = 1$, i.e., $\bar{J}$ increases as

TABLE III
AVERAGE FRAME SIZE FOR DIFFERENT ρ AND c^2 ($J = \infty$).

	c^2=0.5	c^2=1	c^2=2	c^2=3	c^2=5	c^2=10
ρ=0.3	1.073	1.089	1.110	1.126	1.147	1.176
ρ=0.5	1.232	1.274	1.326	1.361	1.408	1.467
ρ=0.7	1.620	1.722	1.839	1.919	2.024	2.156
ρ=0.9	3.368	3.728	4.166	4.476	4.896	5.451
ρ=0.99	21.708	24.324	30.284	33.345	38.333	45.462

c^2 does for a higher system load. In other words, for a high system load, packet aggregation takes place more often as packet size variability increases (unless limited by maximum frame size J). However, for light traffic, $\bar{J}$ is very close to unity for any c^2, which indicates packet aggregation takes place rarely.

Table IV, V, and VI present the average packet end-to-end delay for different aggregation levels (finite and infinite J), system utilizations (ρ), and squared coefficients of variation for packet service time (c^2). Here, a (normalized) header size h of 0.05 is assumed. The numbers in bold represent the minimum average packet delay achievable for each case, which provides the exact optimum level of aggregation. These tables also show the accuracy for the upper and lower bounds in (38) and (39).

Table IV shows that $J = 1$ (i.e., no aggregation) is optimum for light and moderate utilizations. However, the penalty for aggregation is rather minimal even though it may not be needed. The reverse may not be true. Table V shows that packet aggregation improves the system performance in the heavy load. For $\rho = 0.97$, the system is even unstable with no aggregation ($J = 1$), resulting in an infinite delay. But it becomes stable as soon as packet aggregation takes place. These results are consistent with $\rho^\star = 0.863$ in (43), i.e., we aggregate if $\rho > 0.863$ for the header size of $h = 0.05$.

It is interesting to note that the optimum aggregation levels obtained from the exact numerical results in Table V closely match to the heuristic result, $J^\star$, in (44). For $c^2 = 2$, $\rho = 0.9$ and 0.95 even gives the same results with the heuristic. For $c^2 = 0.5,\ \rho = 0.9$, the exact optimum J is 4 and $J^\star = 3$. However, the penalty for choosing a slightly less or greater aggregation level is minimal even if the difference between the heuristic and the optimum results tends to increase as the system utilization and c^2 become higher. For $c^2 = 4,\ \rho = 0.95$, the exact optimum J is 12 resulting in average packet delay of 57.56. If we opt for $J = 16$ as given by the heuristic result $J^\star$, the average delay increases very slightly to 57.93 (not shown in the table). For $c^2 = 0.5,\ \rho = 0.97$, the optimum J is 12 and $J^\star = 11$. For very high load, we have experienced severe numerical stability issues in getting exact result when J is high, especially as c^2 increases.

Table VI shows that the highest difference between the heuristic $\rho^\star$ and the exact numerical results occurs when system utilizations are close to but less than $\rho^\star$. The heuristic result, $\rho^\star = 0.863$ in (43) suggests no aggregation since $\rho < \rho^\star$ for all ρ in Table VI. However, the penalty for no aggregation is not significant as observed in Table VI even if the penalty increases as the squared coefficient of packet service time increases.

TABLE IV
AVERAGE TOTAL DELAY FOR DIFFERENT VALUES OF ρ AND c^2.

	ρ=0.3			ρ=0.5			ρ=0.7		
c^2	0.5	1	2	0.5	1	2	0.5	1	2
J=1	**1.40**	**1.51**	**1.73**	**1.89**	**2.16**	**2.68**	**3.17**	**3.83**	**5.15**
2	1.45	1.57	1.80	2.00	2.28	2.81	3.26	3.90	5.15
3	1.47	1.59	1.83	2.07	2.36	2.91	3.40	4.05	5.31
4	1.47	1.60	1.85	2.10	2.40	2.98	3.50	4.17	5.47
5	1.47	1.60	1.86	2.11	2.42	3.03	3.56	4.27	5.60
6	1.47	1.60	1.86	2.11	2.44	3.06	3.61	4.34	5.71
7	1.47	1.60	1.87	2.12	2.45	3.08	3.64	4.39	5.80
∞	1.47	1.60	1.87	2.12	2.45	3.12	3.69	4.51	6.16
U	1.50	1.63	1.90	2.15	2.48	3.15	3.72	4.54	6.19
L	1.46	1.59	1.85	2.08	2.42	3.08	3.60	4.43	6.07

* $h_1 = 0.05$ assumed. $U(L)$ represents the upper (lower) bound for infinite J.

VI. CONCLUSION

In this paper, we presented a mathematical model for packet aggregation systems. We showed that in the heavy load, system performance can be significantly improved if packet aggregation takes place. For a given header size, we found the minimum system utilization where aggregation improves the system performance. We also provided a simple heuristic result for the optimum level of aggregation. Our results were in close agreement with the exact numerical results derived from our mathematical model for such systems.

REFERENCES

[1] N.T.J. Bailey, "On queueing Processes with Bulk Service," *J. Royal Statistical Association*, 16 Series B, pp. 80-87, 1954.

[2] M.L. Chaudhry and J.G.C. Templeton, *A First Course in Bulk Queues*, Wiley, NY, 1983.

[3] B. Dunsmore and T. Skandier, *Telecommunications Technologies Reference*, Cisco Press, USA, 2003.

[4] W. A. Flanagan, *Voice over Frame Relay*, Telecom Books, NY, 1997.

[5] R.A. Gopalakrishna, "Network Packet Aggregation," U. S. Patent US 6614808 B1, Filed Sep. 2, 1999, Date of patent, Sept. 2, 2003.

[6] J.H. Hong, O. Gusak, N. Oliver, and K. Sohraby, "Performance analysis of packet encapsulation and aggregation," in *Proc. IEEE International Symp. Modeling, Analysis Simulation Computer Telecommun. Syst. (MASCOTS'06)*, Monterey, CA, pp. 137-146, Sept. 2006.

[7] J.H. Hong and K. Sohraby, "On the asymptotic analysis of packet aggregation systems," in *Proc. IEEE International Symp. Modeling, Analysis Simulation Computer Telecommun. Syst. (MASCOTS'07)*, Istanbul, Turkey, pp. 353-359, Oct. 2007.

[8] *Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Standard for Local and Metropolian Area Networks, Part 16, June 24, 2004.

[9] N.K. Jaiswal, "A bulk service queueing problem with variable capacity," *J. Roy. Statist. Soc.*, B26, pp. 143-148, 1964.

[10] K. Kim, S. Ganguly, R. Izmailov, and S. Hong, " On packet aggregation mechanisms for improving VoIP quality in mesh networks," *Veh. Technol. Conf. 2006 (VTC 2006-Spring)*, pp. 891-895, May 2006.

[11] L. Kleinrock, *Queueing Systems: Volume I: Theory*, John Wiley & Sons, Inc., Canada, 1975.

[12] H. Kobayashi and B. L. Mark, *System Modeling and Analysis: Foundations of System Performance Evaluation*, Upper Saddle River, NJ, 2009.

[13] M. Kuczma, B. Choczewski, and R. Ger, *Iterative Functional Equations*, Cambridge Univ., 1990.

[14] M.F. Neuts, "The busy period of a queue with batch service," *Operations Research*, vol. 13, no. 5, pp. 815-819, 1965.

[15] M.F. Neuts, "A general class of bulk queues with Poisson input," *Ann. Math. Statist.* 38, pp. 759-770, 1967.

[16] T. Razafindralambo, I.G. Lassous, L. Iannone, and S. Fdida, "Dynamic packet aggregation to solve performance anomaly in 802.11 wireless networks,", in *Proc. 9th ACM International Symp. Modeling Analysis Simulation Wireless Mobile Syst.*, pp. 247-254, 2006.

[17] S. Shaffer, D. Weiss, and J. Casuba, "Method for constructing adaptive packet lengths in a congested network," U. S. Patent US 6003089, Filed Mar. 31, 1997, Date of patent, Dec. 14, 1999.

[18] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation*, vol. 1, Vacation Priority Syst., Part 1, North-Holland, 1991.

TABLE V
AVERAGE TOTAL DELAY FOR DIFFERENT VALUES OF ρ AND c^2.

	ρ=0.9				ρ=0.95				ρ=0.97			
c^2	0.5	1	2	4	0.5	1	2	4	0.5	1	2	4
J=1	14.16	18.25	26.43	42.80	305.53	400.53	590.53	970.53	∞	∞	∞	∞
2	10.51	13.42	19.23	30.85	29.62	38.66	56.76	92.95	132.37	174.54	258.89	427.56
3	10.03	12.70	18.01	28.61	23.35	30.30	44.21	72.02	55.85	73.37	108.43	178.55
4	**10.00**	**12.57**	17.68	27.86	21.47	27.71	40.19	65.12	43.94	57.51	84.64	138.91
5	10.09	12.63	**17.65**	27.61	20.71	26.60	38.36	61.84	39.34	51.29	75.19	122.98
6	10.22	12.76	17.74	**27.57**	20.40	26.08	37.42	60.04	37.04	48.11	70.26	114.53
7	10.37	12.92	17.88	27.65	**20.28**	25.85	36.91	58.96	35.74	46.27	67.32	109.39
8	10.51	13.08	18.05	27.78	20.29	**25.78**	36.66	58.31	34.97	45.14	65.44	106.00
9	10.64	13.24	18.23	27.95	20.36	25.80	36.55	57.92	34.50	44.41	64.18	103.65
10	10.76	13.39	18.41	28.14	20.46	25.87	**36.54**	57.69	34.22	43.95	63.32	101.96
11	10.87	13.53	18.59	28.35	20.58	25.98	36.59	57.58	34.07	43.66	62.72	100.72
12	10.96	13.66	18.77	28.56	20.72	26.12	36.69	**57.56**	**34.01**	43.49	62.31	99.80
$J=\infty$	11.67	14.98	21.61	34.87	23.66	30.73	44.85	73.11	39.66	51.72	75.85	124.10
U	11.70	15.01	21.64	34.91	23.69	30.76	44.88	73.14	39.69	51.75	75.88	124.14
L	11.33	14.64	21.28	34.54	22.95	30.01	44.14	72.40	38.45	50.51	74.64	122.89

* $h_1 = 0.05$ assumed. $U(L)$ represents the upper (lower) bound for infinite J.

TABLE VI
AVERAGE TOTAL DELAY FOR DIFFERENT VALUES OF ρ AND c^2.

	ρ=0.74			ρ=0.78			ρ=0.82			ρ=0.86		
c^2	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
J=1	**3.71**	4.54	6.20	4.50	5.58	7.73	5.78	7.25	10.20	8.15	10.37	14.80
2	3.75	**4.53**	**6.07**	**4.42**	**5.41**	**7.36**	**5.42**	**6.72**	9.30	7.10	8.92	12.56
3	3.89	4.68	6.22	4.56	5.54	7.46	5.52	6.79	**9.29**	**7.07**	**8.81**	**12.26**
4	4.01	4.82	6.39	4.69	5.70	7.63	5.67	6.94	9.43	7.19	8.91	12.30
5	4.10	4.94	6.54	4.81	5.84	7.80	5.80	7.10	9.60	7.34	9.06	12.43
6	4.16	5.03	6.67	4.89	5.95	7.96	5.91	7.24	9.78	7.47	9.22	12.60
7	4.20	5.10	6.79	4.96	6.05	8.10	6.01	7.37	9.94	7.60	9.37	12.78
$J=\infty$	4.30	5.31	7.34	5.13	6.41	8.96	6.34	7.99	11.30	8.24	10.49	14.98
U	4.33	5.35	7.37	5.17	6.44	8.99	6.37	8.02	11.33	8.27	10.52	15.01
L	4.19	5.21	7.24	5.00	6.28	8.83	6.17	7.82	11.13	8.01	10.26	14.75

* $h_1 = 0.05$ assumed. $U(L)$ represents the upper (lower) bound for infinite J.

Jung Ha Hong received the B.S. and M.S degrees in mathematics from Korea University in 1999 and 2001, respectively, and M.S. in computer science from the University of Missouri - Kansas City in 2006. She is currently working toward the Ph.D. degree in computer science from the University of Missouri - Kansas City.

Her current research interests include queueing theory, design and analysis of telecommunications and computer networks.

Khosrow Sohraby (S'82–M'84–SM'89) received the B.Eng. and M.Eng. degrees from McGill University, Montreal, Canada, in 1979 and 1981, respectively, and the Ph.D. degree from the University of Toronto, Toronto, Canada, in 1985, all in electrical engineering.

His current research interests include design, analysis and control of high-speed computer and communications networks, traffic management and analysis, multimedia networks, networking aspects of wireless and mobile communications, analysis of algorithms, parallel processing and large-scale computations. Refer to http://www.sce.umkc.edu/ sohrabyk/ for a detailed biography.