J. Operations Research Soc. of Japan
Vol. 18, No. 3 & No. 4, September 1975

SOME BOUNDS FOR QUEUES

MASAO MORI

Tokyo Institute of Technology

(Received April 30; Revised August 5, 1974)

Abstract. We will give some bounds for the variance of waiting time of the system *GI/G/1*, which will give fairly good evaluations and are handy to calculate. And some bounds on the mean waiting time and on the probability of no wait will be improved for the systems *GI/G/1* and *GI/G/k*. A notion of *"the virtual waiting time vector"* is usefull to derive some of these bounds.

1. Introduction

In these years several authors have investigated inequalities for some queueing systems. For the usual single server queue *GI/G/1*. Kingman [5] obtained some upper bounds on the mean waiting time at first, and Marshall [8] gave the lower bounds. The results in [8] are very good for the class of systems of which inter-arrival distribution functions have increasing failure rate *(IFR)*. Further Kingman [6] showed the upper and lower bounds for the tail of waiting time distribution function. In the bulk queue with a single server, similar results were shown by Marshall [9] and Suzuki and Yoshida [12].

For the variance of the waiting time of *GI/G/1* queue, an upper bound in [5] is available, but the result is not handy to calculate. In this paper we will give some fairly good bounds on the variance,

which are easy to calculate. And further we will improve bounds on the mean waiting time and an upper bound on the probability of no wait.

In the section 5, we will deal with many server systems. For the queue $GI/G/k$, Kingman [6] derived lower and upper bounds on the mean waiting time, but this upper bound is not so good. Suzuki and Yoshida [12] also obtained both bounds for the special cases. Recently, Brumelle [1] has obtained both bounds for the system with stationary inputs. In the case of renewal inputs, this upper bound is the same as Kingman's result. But the lower bound is rather sharp. We will try here to improve the upper bound, which is rather appraisable especially in $GI/E_p/k$, and to obtain some other inequalities.

At first, in the section 3, we will introduce the notion of virtual waiting time vector and the total residual service time for servers. By using these quantities we will give a few fundamental inequalities which are useful to derive some bounds in the mean waiting time in the later sections.

2. Notations

In order to describe the system $GI/G/k$, let us introduce the following notations;

t_n : the inter-arrival time between the $(n-1)$-th and n-th arrivals $(n = 1, 2, 3, \ldots)$,

s_n : the service time of the n-th arriving customer $(n = 0, 1, 2, \ldots)$,

$T_n = \sum_{j=1}^{n} t_j$: the time point of the n-th arrival $(T_0 \equiv 0)$,

$u_n = \frac{s_n}{k} - t_{n+1}$,

$A(x)$: the distribution function of t_n,

$B(x)$: the distribution function of s_n,,

$K(x)$: the distribution function of u_n,

λ : the mean arrival rate (or $1/\lambda = E(t)$),

$c_a = \sqrt{\mathrm{var}(t)}/E(t)$: the coefficient of variation of $A(x)$,

$\rho = \lambda E(s)/k$: the traffic intensity for the system,

$J_k = \mathrm{var}(u)/2E(-u)$,

$W_n = (w_{n1}, w_{n2}, \ldots, w_{nk})$: the waiting time vector where

$$0 \leqq w_{n1} \leqq w_{n2} \leqq \cdots \leqq w_{nk} ,$$

$w_n = w_{n1}$: the waiting time in the queue of the n-th arriving customer,

W : the limiting random vector of the sequence $\{W_n\}$,

w : the limiting random variable of the sequence $\{w_n\}$,

$W(x)$: the distribution function of w,

$a_0 = W(0+) = P\{w=0\}$: the stationary probability of no wait,

$V(t) = (v_1(t), v_2(t), \ldots, v_k(t))$: the virtual waiting time vector

where $0 \leqq v_1(t) \leqq v_2(t) \leqq \cdots \leqq v_k(t)$,

$v(t) = v_1(t)$: the virtual waiting time of the customer who were assumed to arrive at the time t,

$Y(t)$: the total residual service time for all k servers at the time t,

$Y_n = Y(T_n - 0)$,

R_{n+1} : the total idle time for k servers in the interval (T_n, T_{n+1}),

$y_n = Y_n/k$,

$r_{n+1} = R_{n+1}/k$,

$z_n = y_n - w_n$,

$$\bar{Y} = \lim_{t\to\infty} \frac{1}{t} \int_0^t Y(s)ds,$$

$$\hat{Y} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} w_i$$

$w_n^{(i)}$: the waiting time of the n-th arrival for the modified i-server system $GI/G/i$, to which the input process $\{(\frac{i}{k}s_n, t_{n+1})\}$ is brought in (i.e. this $GI/G/i$ system has the same traffic intensity as the original system $GI/G/k$),

$N(t)$: the renewal number of the renewal process generated by the arrival process $\{t_n\}$,

$\theta(t) = t - T_{N(t)}$: the age of the above renewal process at time t.

In this paper, we will consider the equilibrium queueing process in almost cases, so we drop the suffix n when it is not necessary to specify the order of the arrivals. Thus we will often write as w, W, y and so on in the stationary state.

3. Fundamental Lemmas

Now we will describe the system $GI/G/k$. Let it be assumed that the input processes $\{t_n\}$ and $\{s_n\}$ are independent, and they are mutually independent and identically distributed respectively. And it is assumed that customers are served in the order of their arrivals (*FIFS* discipline).

3.1 On the process $\{W_n\}$ and $\{Y_n\}$

At first we will describe the process $\{W_n\}$ and $\{Y_n\}$, both of which are observed at the time just before arrivals. $W_n = (w_{n1}, w_{n2}, \ldots, w_{nk})$ is the waiting time vector introduced by Kiefer and Wolfowitz [4] where $0 \leqq w_{n1} \leqq \ldots \leqq w_{nk}$. The meaning of w_{ni} is the time length from the time point of the n-th arrival to the i-th smallest time point of the work completion times of k servers on which each servers would become idle but for the arrivals after the time T_{n+0}. Thus w_{n1} means the waiting time spent in the queue of the n-th arriving customer, denoted by w_n. Then $\{W_n\}$ satisfies the following recurrence relation:

$$(3.1) \quad W_{n+1} = \mathcal{R}((w_{n1} + s_n - t_{n+1})^+, (w_{n2} - t_{n+1})^+, \ldots, (w_{nk} - t_{n+1})^+)$$

where $\mathcal{R}$ is the rotational operator which rearranges the components of a vector in ascending order and $a^+ = max(0, a)$. And throughout this paper let it be assumed that the condition of ergodicity is satisfied, i.e. $\rho = E(s)/kE(t) < 1$. So there exists the limiting random vector W of the sequence $\{W_n\}$.

Next we will derive the relation for $\{Y_n\}$ or $\{y_n\}$. From the definitions of Y_n and R_{n+1} we get easily

$$(3.2) \quad Y_{n+1} = \sum_{i=1}^{k} w_{n+1,i} = \sum_{i=1}^{k} (w_{ni} + \delta_{i1} s_n - t_n)^+$$

and

$$(3.3) \quad R_{n+1} = \sum_{i=1}^{k} (w_{ni} + \delta_{i1} s_n - t_{n+1})^-$$

where δ_{ij} denotes Kronecker's delta and $a^- = max(0, -a)$. From the above equation we easily obtain

$$(3.4) \quad Y_{n+1} - R_{n+1} = Y_n + ku_n \quad \text{or} \quad y_{n+1} - r_{n+1} = y_n + u_n \ .$$

In the case of $k = 1$, this equation represents the recurrence equation for the waiting time process in which Y_n means the waiting time itself and R_{n+1} is the idle time between the n-th and the $(n+1)$-st arrivals. And in this case $(k = 1)$, $R_{n+1} \cdot Y_{n+1} = 0$ always holds, but it is not always true for $k \geqq 2$. Hence generally we have

$$(3.5) \quad Y_{n+1} \geqq (Y_n + ku_n)^+ \quad \text{or} \quad y_{n+1} \geqq (y_n + u_n)^+$$

where equality always holds in the case of $k = 1$. Equality in (3.5) also holds even for $k \geqq 2$ during the time interval in which all servers are busy. This fact will be suggestive in considering the bounds when ρ is near 1.

For convienience we rewrite y_n as

$$(3.6) \quad y_n = w_n + z_n \quad \text{where} \quad z_n + \frac{1}{k} \sum_{i=1}^{k} (w_{ni} - w_{n1}).$$

If $w_n > 0$, the quantities $\{w_{ni} - w_n : i = 2, 3, \ldots, k\}$ are the residual service times of the customers who are being served just at the beginning of the n-th arriving customer's service, i.e. at the time $T_n + w_n$, thus they are expected to be as large as the magnitude of length of a service time or so. However, if $w_n = 0$, some of $\{w_{ni} - w_n\}$ may be 0.

3.2 On the processes $\{V(t)\}$ and $\{Y(t)\}$

Now we will describe the processes $\{V(t)\}$ and $\{Y(t)\}$, which are observed at arbitrary time points. We introduce the virtual waiting time vector $V(t) = (v_1(t), \ldots, v_k(t))$, where $0 \leqq v_1(t) \leqq \ldots \leqq v_k(t)$,

which is interpreted as follows: the meaning of $v_i(t)$ is the time length from the time t to the i-th time point of the times rearranged in ascending order on which each servers would become idle but for arrivals after the time t. Thus $v_1(t)$ is the so called virtual waiting time at t, which we will denote by $v(t)$.

Now we are trying to represent $V(t)$ by using $\{W_n\}$. Let us consider a renewal process generated by inter-arrival times $\{t_n\}$. $N(t)$ is defined as the renewal number by the time t, i.e. $N(t)$ is a number such that $T_{N(t)} \leqq t < T_{N(t)+1}$ holds. Hereafter we will often write $N(t)$ as N for abreviation. And $\theta(t)$ is defined as 'the age' at the time t, i.e. $\theta(t) = t - T_N$. Then from the definitions of $\{W_n\}$ and $\{V(t)\}$ we have

$$(3.7) \quad V(t) = \mathcal{R}((w_{N1} + s_N - \theta(t))^+, (w_{N2} - \theta(t))^+, \ldots, (w_{NK} - \theta(t))^+).$$

And it has been shown by the author [10] that if $A(x)$ is non-arithmetic distribution function, then for an arbitrary subset A of R^k

$$(3.8) \quad \lim_{t\to\infty} P\{W_N \in A, \theta(t) \leqq \theta\} = P\{W \in A\} \cdot \lambda \int_0^\theta \{1 - A(y)\}dy.$$

It has been proved by using fact that the two dimensional stochastic process $\{(W_n, t_{n+1})\}$ forms the so-called (J, X)-process[1] with continuous state space R^k. The equation (3.8) enables us to consider that W_N and s_N are independent of $\theta(t)$ in the equilibrium state. And the restriction that $A(x)$ is non-arithmetic is removed in considering the equilibrium state.

---

(1) As for (J, X)-process, see the paper: R. Pyke, Markov renewal process: definitions and preliminary properties, Ann. Math. Statist., 32 (1961), 1231-1242.

Analogously to (3.2), $Y(t)$can be represented as

$$(3.9) \quad Y(t) = \sum_{i=1}^{k} v_i(t) = \sum_{i=1}^{k} (w_{Ni} + \delta_{i1} s_N - \theta(\tau))^+.$$

Now we will give the fundamental equality and inequalities on the process $Y(t)$, from which some upper and lower bounds on $E(w)$ will be derived. It is worthy to notice that $E(\hat{Y}) = E(Y)$, where Y represents the limiting random variable of the sequence $\{Y_n\}$. Now we have the following two lemmas.

Lemma 3.1 (*Brumelle* [1])

In a system GI/G/k, if $\rho < 1$, $E(s^2) < \infty$ *and* $E(w) < \infty$, *we have*

$$(3.10) \quad E(\overline{Y}) = \lambda E(s) \cdot E(w) + \frac{\lambda E(s^2)}{2}$$

$$= k\rho E(w) + \frac{\lambda E(s^2)}{2} \quad .$$

Lemma 3.2

With the same conditions stated in Lemma 3.1, we have

$$(3.11) \quad E(\overline{Y}) \geqq \rho E(Y) + \frac{\lambda E(s^2)}{2k} \quad .$$

and

$$(3.12) \quad E(\overline{Y}) \leqq E(Y) + \frac{1}{2} E(s).$$

Proof) We shall derive the above relations by observing a sample path of $Y(t)$. Fig.1, 2 illustrate a sample path of $Y(t)$ for the case of $k = 3$.

At first we will show (3.11). Let us draw straight lines with slope $-k$ from each points (T_n, Y_n) and $(T_n, Y_n + s_n)$ (Fig.1). In evaluating

$E(\bar{Y})$ we defy the dotted area in Fig.1, then we have

$$(3.13) \quad \frac{1}{t}\int_0^{t^*} Y(s)ds \geqq \frac{1}{t}\sum_{n=0}^{N(t)} \frac{1}{2}(Y_n + s_n)\cdot\frac{s_n}{k}$$

where $t^* = max(t, T_N + \frac{1}{k}(Y_N + s_N))$. If we take t at any time point on which all servers are idle, then we have $t = t^*$. But in general case, $t^* - t \leqq |\frac{1}{k}(Y_N + s_N) - \theta(t)|$. If we take $W_0 = (0, \ldots, 0)$, from [4] it holds that $W_0 \overset{st.}{\leqq} W_1 \overset{st.}{\leqq} W_2 \overset{st.}{\leqq} \cdots \overset{st.}{\leqq} W$, where $W_n \overset{st.}{\leqq} W_{n+1}$ means that W_n is stochastically smaller than W_{n+1}, i.e. $P\{W_n > a\} \leqq P\{W_{n+1} > a\}$ for any $a \in R^k$. So we have $Y_0 \overset{st.}{\leqq} Y_1 \overset{st.}{\leqq} \cdots \overset{st.}{\leqq} Y$. And if $\rho < 1$, W has a proper distribution. Thus, recalling (3.8), we can consider the quantity $\{\frac{1}{k}(Y_N + s_N) - \theta(t)\}$ as a properly distributed random variable, and we can obtain

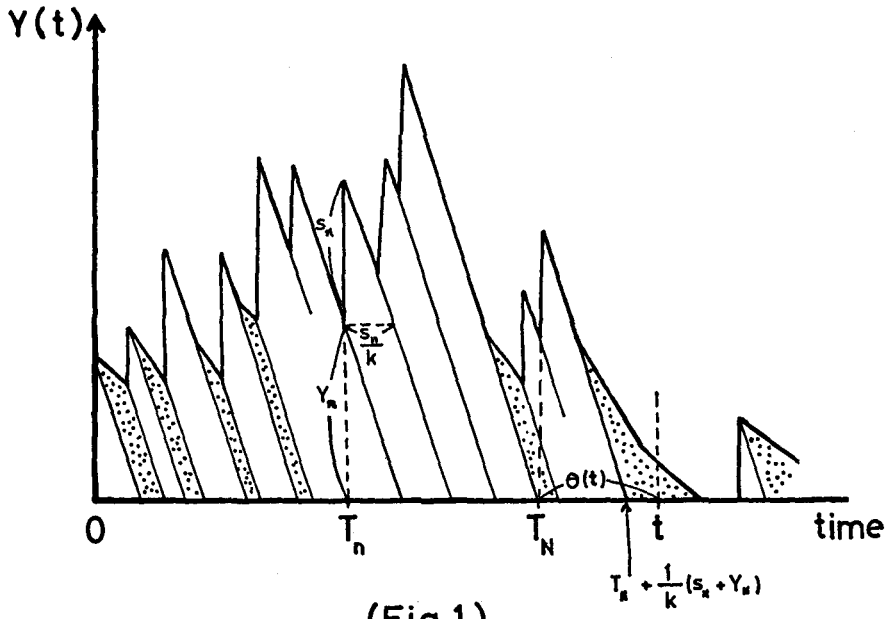
$$E\{\lim_{t\to\infty}\frac{1}{t}\int_0^{t^*} Y(s)\} = E\{\lim_{t\to\infty}\frac{1}{t}\int_0^{t} Y(s)ds\}\ .$$

Therefore, by evaluating the right side of (3.13) in the usual manner, (3.11) is derived directly.

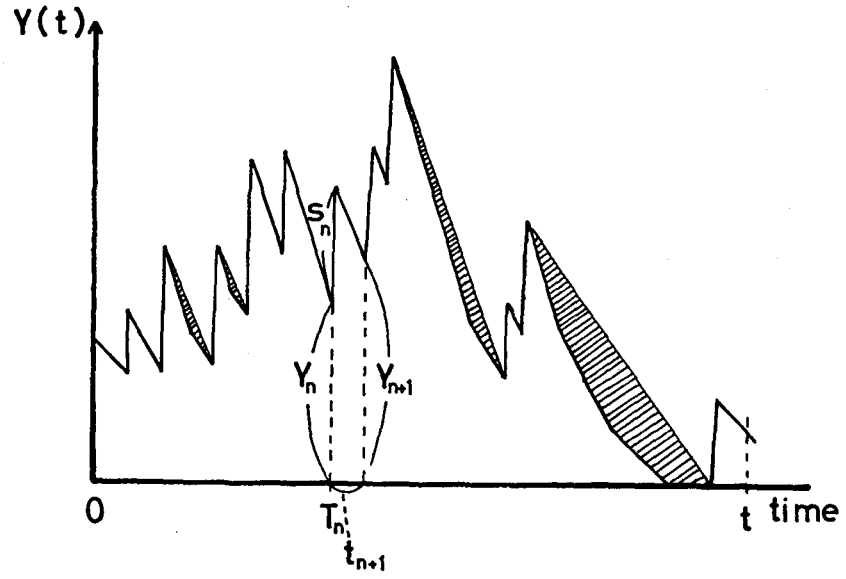
For the derivation of (3.12), we draw straight lines between $(T_n, Y_n + s_n)$ and (T_{n+1}, Y_{n+1}) for each n (Fig.2), and in estimating $E(\bar{Y})$, add the area increased by drawing these lines, which is the part with short oblique lines in Fig.1. Then we have

$$(3.14) \quad \frac{1}{t}\int_0^{t} Y(s)ds \leqq \frac{1}{t}\sum_{n=0}^{N(t)} t_{n+1}\cdot\frac{1}{2}(Y_n + s_n + Y_{n+1})\ .$$

Here t_{n+1} is independent of Y_n and s_n, so $cov(t_{n+1}, Y_{n+1}) \leqq 0$ from (3.2) for each n. Thus we can derive (3.12).



(Fig 1)



(Fig 2)

Next we will study the relation between the stationary distributions of $\{W_N\}$ and $\{V(t)\}$. Before deriving this relation, we will mention the notion of γ-*MRLA*, *IFR* and so on, which will take an important role in the later. Let $F(x)$ be a distribution function, defined on $[0, \infty)$.

Definition 1. γ-*MRLA (mean residual life bounded by* γ *from above).*

$$F(x) : \gamma\text{-MRLA} \longleftrightarrow \frac{\int_x^\infty [1 - F(y)]dy}{1 - F(x)} \leqq \gamma \quad \text{for all } x \geqq 0.$$

Definition 2. *IFR (increasing failure rate).*

$$F(x) : IFR \longleftrightarrow \frac{F(x + \Delta) - F(x)}{1 - F(x)} \quad \text{is increasing in } x \text{ for any } \Delta > 0.$$

Again, γ-*MRLB* (mean residual life bounded by γ from below) and (decreasing failure rate) are defined in the same way by reversing inequality and by replacing the word 'increasing' into the word 'decreasing' in the above definitions. We denote the queueing system for which the distribution function $A(x)$ is $\frac{1}{\lambda}$ - *MRLA* by $\frac{1}{\lambda}$ - *MRLA/G/k*. *IFR/G/1* and so on should be also interpreted similarly.

Now we assume that the queueing process is already in the stationary state. Thus the distribution function of $\theta(t)$ may be written as follows using the renewal theory;

$$(3.15) \quad P\{\theta(t) \leqq x\} = \lambda \int_0^x [1 - A(y)]dy \equiv A^*(x)$$

And we denote a typical random variable distributed with $A(x)$ by τ.

If $A(x)$ is $\frac{1}{\lambda}$ - *MRLA*, $\theta(t)$ is stochastically smaller than τ, i.e. $\theta(t) \overset{st.}{\leqq} \tau$, because $1 - A^*(x) \leqq 1 - A(x)$ is easily shown from the definition.

And further from (3.8), W_N is samely distributed as the stationary waiting time vector W, denoting this by $W_N \sim W$. Thus it holds

$$(3.16) \quad W \sim \mathcal{R}((w_{N1} + s_N - \tau)^+, \ (w_{N2} - \tau)^+, \ \ldots, \ (w_{Nk} - \tau)^+).$$

By comparing the above relation with (3.7), we get $W \overset{st.}{\leqq} V(t)$ in the system $\frac{1}{\lambda}$ - *MRLA/G/k*. Then we have the following lemma from the above discussions.

Lemma 3.3

In a system $\frac{1}{\lambda}$ - *MRLA/G/k, we have in equilibrium state*

$$(3.17) \quad W \overset{st.}{\leqq} V(t),$$

$$(3.18) \quad Y \overset{st.}{\leqq} Y(t)$$

and

$$(3.19) \quad E(Y) \leqq E(\overline{Y}).$$

In the case of $\frac{1}{\lambda}$ - *MRLB/G/k*, the aboves are true by reversing inequalities. And equalities hold in the system *M/G/k*.

4. The Single Server System

In this section we deal with the single server system, i.e. $k = 1$. In this case, it is neccessary to notice that $W_n = Y_n = w_n$ and $V(t) = Y(t) = v(t)$ always hold, and that $\{w_n\}$ satisfies the following recurrence equation.

$$(4.1) \quad w_{n+1} = (w_n + u_n)^+ \quad \text{or} \quad w_{n+1} - r_{n+1} = w_n + u_n \ .$$

Now in this section, let us consider that the queueing system is already being in equilibrium state.

4.1 On the mean waiting time

At first we will derive a lower bound on the mean waiting time by using results in the section 3, which is true for all *GI/G/k* queues. Rewriting (3.10) and (3.12) in this case, we have

$$E(\bar{v}) = \rho E(w) + \frac{\lambda E(s^2)}{2} \tag{4.2}$$

and

$$E(\bar{v}) \leqq E(w) + \frac{1}{2} E(s) \tag{4.3}$$

where $\bar{v} = \lim_{t\to\infty} \frac{1}{t} \int_0^t v(s)ds.$

By eliminating $E(\bar{v})$ and inserting $\rho = \lambda E(s)$, the above relations imply

$$E(w) \geqq \frac{\lambda \mathrm{var}(s)}{2(1-\rho)} - \frac{1}{2} E(s). \tag{4.4}$$

This lower bound gives a good estimate for the system in which the variance of the inter-arrival time is small compared with that of the service time, such as in a *D/M/1* queue. But in the reverse case of the above such as in a *M/D/1* queue, this estimate is not so good. Kingman [6] gives a lower bound, i.e. $E(w) \geqq E(u^+)^2/2E(-u)$, but it is not so handy to calculate. Recently, a method of obtaining good lower bounds has been proposed by Cox and Bloomfield [3] for the case of small ρ, but it is also not so easy to calculate.

$\frac{1}{\lambda}$ *-MRLA/G/1 queue*

Now we consider a $\frac{1}{\lambda}$ - *MRLA/G/1* queue. Rewriting (3.19) for the single server system, we have

(4.5) $E(w) \leqq E(\bar{v})$.

By eliminating $E(\bar{v})$ from (4.2) and (4.5), we get easily

(4.6) $$E(w) \leqq \frac{\lambda E(s^2)}{2(1-\rho)} .$$

Summing up this results with Marshall's [8], we have

(4.7) $$J_1 - \frac{1+\rho}{2\lambda} \leqq E(w) \leqq min\{J_1, \frac{\lambda E(s^2)}{2(1-\rho)}\}$$

for a $\frac{1}{\lambda}$ - *MRLA/G/1* queue.

Comparing J_1 with $\lambda E(s^2)/2(1-\rho)$, we have

$$J_1 - \frac{\lambda E(s^2)}{2(1-\rho)} = \frac{\lambda var(t) - \lambda(E(s))^2}{2(1-\rho)} = \frac{\frac{1}{\lambda}(c_a^2 - \rho^2)}{2(1-\rho)}$$

where c_a is the coefficient of variation of $A(x)$. In the case that $var(t)$ is small such as in a *D/G/1* queue, J_1 is smaller. But on the contrary, in the case that $var(t)$ is large such as in a *M/G/1* queue, $\frac{\lambda E(s^2)}{2(1-\rho)}$ is smaller.

For a *IFR/G/1* queue, a better lower bound is given in [8], that is $J_1 - \frac{c_a^2 + \rho}{2\lambda} \leqq E(w)$.

$\frac{1}{\lambda}$ - *MRLB/G/1 queue*

In this case, instead of (4.5) we have

$$E(w) \geqq E(\bar{v}),$$

thus we can obtain

$$(4.8) \quad E(w) \geqq \frac{\lambda E(s^2)}{2(1-\rho)} .$$

This is a new result. Thus, combining this and the results in [8], we have

$$(4.9) \quad J_1 - \frac{1+\rho}{2\lambda} \geqq E(w) \geqq \frac{\lambda E(s^2)}{2(1-\rho)}$$

for a $\frac{1}{\lambda}$ - *MRLB/G/1* queue. The difference between the upper and the lower

$$\left(J_1 - \frac{1+\rho}{2\lambda}\right) - \frac{\lambda E(s^2)}{2(1-\rho)} = \frac{c_a^2 - 1}{2\lambda(1-\rho)}$$

is rather small also for this case in light traffic, but this difference is rather large in heavy traffic, since $(1-\rho)$ becomes small.

For a *DFR/G/1* queue, the upper bound is improved in [8], so we have

$$(4.10) \quad J_1 - \frac{c_a^2 + \rho}{2\lambda} \geqq E(w) \geqq \frac{\lambda E(s^2)}{2(1-\rho)} .$$

4.2 On the variance of the waiting time

Now let us take expectations of both sides of (4.1) or (3.4), we get at first

$$(4.11) \quad E(r) = E(-u).$$

Let us square (4.1) and take expectations of both sides, we have

$$(4.12) \quad E(w) = \frac{E(u^2)}{2E(-u)} - \frac{E(I^2)}{2E(I)} .$$

And further in the same manner as above, by multiplying (4.1) with its square and taking expectations of both sides, we can easily reduce

$$(4.13) \quad \mathrm{var}(w) = -\{ \frac{E(u^3)}{3E(u)} - [\frac{E(u^2)}{2E(u)}]^2 \} + \{ \frac{E(I^3)}{3E(I)} - [\frac{E(I^2)}{2E(I)}]^2 \} .$$

In the above equations I represents a typical idle time of the server, to be more precisely, we put $I = r_{n+1}$ when $r_{n+1} = (w_n + u_n)^- > 0$. Thus we have

$$(4.14) \quad E(r^m) = a_0 \, E(I^m)$$

for all $m > 0$ where $a_0 = P\{w = 0\}$, if either of both sides exist. And we denote the distribution function of I by $H(x)$.

First we derive the lower bound in the following:

Proposition 4.1

For all GI/G/1 queues, if $\rho < 1$ *and* $E(s^3) < \infty$, *we have*

$$(4.15) \quad \mathrm{var}(w) \geqq - \frac{E(u^3)}{3E(u)} + [\frac{E(u^2)}{2E(u)}]^2 \equiv \underline{\sigma_w^2} .$$

Proof) Consider the renewal process generated by the sequence of independent random variables $\{I_n\}$, where each I_n is distributed with the distribution function $H(x)$. Denote the residual life of this renewal process at time t by $\eta(t)$, and the well-known results in the renewal theory imply

$$\lim_{t\to\infty} P\{\eta(t) \leq x\} = \frac{\int_0^x [1 - H(y)]dy}{E(I)} .$$

The variance of this limiting distribution is

$$\frac{E(I^3)}{3E(I)} - \left[\frac{E(I^2)}{2E(I)}\right]^2 \equiv \sigma^2_{r,I} .$$

Therefore this quantity is non-negative. Inserting this to (4.13), thus (4.15) is derived.

To our regret, however, we have failed to derive an upper bound for $\sigma^2_{R,I}$, so we cannot give an upper bound on $var(w)$ for the general cases. Kingman [5] gave the following upper bound under the conditions that $\rho < 1$, $E(u^2) < 0$ and $E(e^{\varepsilon s}) < \infty$ for some $\varepsilon > 0$;

$$var(w) \leq \min_{\theta}\{\frac{4}{e^2\theta^2} \log \frac{1}{1 - \phi(\theta)}\}$$

where $\phi(\theta) = E(e^{\theta u})$. But it has not been ready to enumerate. Now we are trying to investigate upper bounds for some cases.

$\frac{1}{\lambda}$ - *MRLA/G/1 queue*

In this case we will use the following lemma directly derived from the proof of Theorem 4 in [8].

Lemma 4.2 *(Marshall)*

If $A(x)$ is γ-MRLA(B), then $H(x)$ is also γ-MRLA(B).

Proposition 4.3

For a $\frac{1}{\lambda}$ *- MRLA/G/1 queue, we have*

$$(4.16) \quad \mathrm{var}(w) \leqq \sigma_w^2 + \frac{2}{\lambda^2} - \frac{(1-\rho)^2}{4\lambda^2}$$

Proof) In this case, $H(x)$ is also $\frac{1}{\lambda}$ - *MRLA*. We denote by η the limiting random variable of the residual life $\eta(t)$ mentioned in the proof of Proposition 4.1. The fact $H(x)$ is $\frac{1}{\lambda}$ - *MRLA* implies

$$(4.17) \quad \frac{\mathrm{E}(I^j)}{j\mathrm{E}(I)} = \mathrm{E}(\eta^{j-1}) \leqq (j-1)! \cdot \left(\frac{1}{\lambda}\right)^{j-1}$$

for all $j \geqq 2$. Inserting this relation of the case $j = 3$ into (4.13), we get

$$\mathrm{var}(w) \leqq \sigma_w^2 + \frac{2}{\lambda^2} - \left[\frac{\mathrm{E}(I^2)}{2\mathrm{E}(I)}\right]^2 .$$

Now we are trying to obtain a lower bound for the last term of the right-hand side. By considering $\mathrm{var}(I) \geqq 0$, we have $\mathrm{E}(I^2)/2\mathrm{E}(I) \geqq \mathrm{E}(I)/2$. Noticing $\mathrm{E}(r) = \mathrm{E}(-u)$, and taking (4.14) into consideration, we can easily obtain

$$(4.18) \quad \mathrm{var}(w) \leqq \sigma_w^2 + \frac{2}{\lambda^2} - \frac{(1-\rho)^2}{4\lambda^2} \cdot \frac{1}{a_0}$$

from which (4.16) is directly derived.

If we can get a good estimate of a_0 in hand, we are able to improve this upper bound.

In the case of *M/G/1*, I is also distributed with exponential distribution with mean $1/\lambda$, so the equality in (4.17) holds for all $j \geqq 2$. Thus we have

$$(4.19) \quad \mathrm{var}(w) = \sigma_w^2 + \frac{1}{\lambda^2}$$

for a *M/G/1* queue. Comparing (4.15) and (4.16) with the above equation, we notice that both bounds just obtained give very good estimates.

IFR/G/1 queue

In this case, we can improve the upper bound slightly.

Lemma 4.4 *(Theorem 5 (iii) of [8])*

For a IFR/G/1 queue,

$$(4.20) \quad \frac{\int_x^\infty [1 - H(y)]dy}{E(I)} \leqq \lambda \int_x^\infty [1 - A(y)]dy \ .$$

(In the case of *DFR/G/1* queues, it is enough to reverse the inequality).

From (4.20), we can easily obtain

$$(4.21) \quad \frac{E(I^j)}{E(I)} \leqq \frac{E(t^j)}{E(t)}$$

where t represents a typical random variable distributed with $A(x)$. Thus we have

$$(4.22) \quad \mathrm{var}(w) \leqq \sigma_w^2 + \frac{\lambda E(t^3)}{3} - \frac{(1-\rho)^2}{4\lambda^2}$$

for a *IFR/F/1* queue.

In the case of *D/G/1* queue, (4.22) is followed to

$$\mathrm{var}(w) \leqq \sigma_w^2 + \frac{1}{3\lambda^2} - \frac{(1-\rho)^2}{4\lambda^2} \ .$$

This upper bound (4.22) is expected to give a good estimate specially for the case that $\mathrm{var}(t)$ is small.

To our regret, we cannot give upper bounds in the case of $\frac{1}{\lambda}$ - *MRLB/G/1* queues and *DFR/G/1* queues.

4.3 The probability of no wait

Now we shall derive an estimate for $a_0 = \mathrm{P}\{w = 0\}$. From the relation (4.1) and by using $w_n \geqq 0$, we have

$$w_{n+1} \geqq u_n^+ \geqq u_n \tag{4.23}$$

for all n. Then we have

$$\mathrm{P}\{w_n \leqq x\} \leqq K(x) = \int_0^\infty \{1 - A(y - x)\}dB(y) \tag{4.24}$$

for all $n \geqq 1$, which implies

$$a_0 \leqq K(0+) \tag{4.25}$$

for all *GI/G/1* queues.

Thus the result in [8] is improved as

$$(1 - \rho) \leqq a_0 \leqq K(0+) = \int_0^\infty \{1 - A(y)\}dB(y) \tag{4.26}$$

in the case of $\frac{1}{\lambda}$ - *MRLA/G/1* queue, for in [8] the upper bound of a_0 is given as 1 in this case. For example, in the system *M/D/1*, we have $a_0 = 1 - \rho$ and $K(0+) = e^{-\rho}$. The difference between the upper bound and the exact value of a_0 is $K(0+) - a_0 = \frac{1}{2!}\rho^2 - \frac{1}{3!}\rho^3 + \ldots$, which is rather small. But the bounds on the probability of no wait must be improved much more.

5. The many Server System

Now we shall deal with the many server system $GI/G/k$ defined in the section 3. At first let us introduce a modified system $GI/G/i$ in order to compare the original one $GI/G/k$. We only change the number of servers k and the service times $\{s_n\}$ of the original system into i and $\{\frac{i}{k} s_n\}$ respectively, but we let the arrival points $\{T_n\}$ unchanged. That is, let the waiting time vectors of this modified system $\{W_n^{(i)}\}$ be generated by the same samples $\{(s_n, t_{n+1})\}$ as those used in the original system in the following manner:

$$(5.1) \quad W_{n+1}^{(i)} = \mathcal{R}\left((w_{n1}^{(i)} + \frac{i}{k} s_n - t_{n+1})^+, (w_{n2}^{(i)} - t_{n+1})^+, \ldots, (w_{ni}^{(i)} - t_{n+1})^+\right).$$

Especially for $i = 1$, we have the following recurrence relation:

$$(5.2) \quad w_{n+1}^{(1)} = (w_n^{(1)} + \frac{1}{k} s_n - t_{n+1})^+.$$

By using $\{W_n^{(i)}\}$ we can obtain some bounds for $E(w)$.

5.1 The lower bound on the mean waiting time

Now from the relations (3.6) and (5.2), if we assume $w_0^{(1)} \leq y_0$, then we derive

$$(5.3) \quad w_n^{(1)} \leq y_n = w_n + z_n$$

recursively for all n. If we estimate $E(z)$ in the equilibrium state, we are able to give lower bounds for $E(w)$ by using the results for the single server system stated in the section 4. Thus we are trying to estimate $E(z)$.

At first from the relations (3.6), (3.10) and (3.11), the inequality

$$(5.4) \quad E(z) \leq \frac{k-1}{k} \cdot \frac{E(s^2)}{2E(s)}$$

is easily obtained. Thus we can get

$$(5.5) \quad E(w) \geq E(w^{(1)}) - \frac{k-1}{k} \cdot \frac{E(s^2)}{2E(s)}$$

from (5.3) and (5.4). This lower bound is quite the same as in [1]. From the way of derivating (3.11), this evaluation is expected to be good for small k and large ρ. But for the case of large k or small ρ, this bound is not so good, because the estimation of the upper bound of $E(z)$ is too coarse. But in the following special cases, the results are slightly improved.

$\frac{1}{\lambda}$ - MRLA/G/k queue

From the relations (3.10) and (3.19), we have

$$(5.6) \quad (1-\rho)\cdot E(w) \leq \frac{\lambda E(s^2)}{2k} - E(z).$$

The left hand side of the above inequality is non-negative, so we have $E(z) \leq \frac{\lambda E(s^2)}{2k} = \rho \cdot \frac{E(s^2)}{2E(s)}$, from which the lower bound is slightly improved as

$$(5.7) \quad E(w) \geq \begin{cases} E(w^{(1)}) - \rho \cdot \frac{E(s^2)}{2E(s)}, & \text{if } \rho \leq \frac{k-1}{k}, \\ E(w^{(1)}) - \frac{k-1}{k} \cdot \frac{E(s^2)}{2E(s)}, & \text{if } 1 > \rho \geq \frac{k-1}{k}. \end{cases}$$

$\frac{1}{\lambda}$ -MRLB/G/k queue

From the relation (3.10) and (3.19) with reverse inequality,

calculation using (5.4) implies that

$$(5.8) \quad E(w) \geq \frac{\lambda E(s^2)}{2(1-\rho)\rho k^2} - \frac{E(s^2)}{2E(s)} .$$

5.2 The upper bounds on the mean waiting time

At first we will obtain a rather coarse upper bound for the general system *GI/G/k*. By substituting (3.6) into (3.4), we have

$$(5.9) \quad w_{n+1} + z_{n+1} - r_{n+1} = w_n + z_n + u_n .$$

And from the definitions of z_n and r_n, they are represented as follows:

$$(5.10) \quad z_n = \frac{1}{k}\{(w_{n-1,1} + s_{n-1} - t_n)^+ + (w_{n-1,2} - t_n)^+ + \cdots + (w_{n-1,k} - t_n)^+\} - w_n$$

and

$$(5.11) \quad r_n = \frac{1}{k}\{(w_{n-1,1} + s_{n-1} - t_n)^- + (w_{n-1,2} - t_n)^- + \cdots + (w_{n-1,k} - t_n)^-\} .$$

Recalling that u_n is independent of both w_n and z_n and that $w_{n+1} \cdot r_{n+1} = 0$, square (5.9) and take expectations, then in the equilibrium state we get

$$(5.12) \quad 2E(-u) \cdot E(w) = \mathrm{var}(u) + 2\mathrm{cov}(z_n, r_n) - \mathrm{var}(r)$$

by inserting $E(r) = E(-u)$, where the subscript n means the number of a typical customer in the stationary state. It is conjectured that $2\mathrm{cov}(z_n, r_n) - \mathrm{var}(r) \leq 0$ and that $E(w) \leq \mathrm{var}(u)/2E(-u)$, but we have not been successed to prove it. However, we can give a rough

bound for $\mathrm{cov}(z_n, r_n)$, from which we can estimate $E(w)$ as follows:

Proposition 5.1

For all GI/G/k queues, if $\rho < 1$ *and* $E(s^2) < \infty$, *we have*

$$E(w) \leqq J_k + \frac{\frac{k-1}{k^2}\left(1 - \frac{1}{k\rho}\right)^+ \cdot E(s^2)}{2E(-u)} \tag{5.13}$$

where $J_k = \mathrm{var}(u)/2E(-u)$.

Proof) In the case of $\rho \leqq \frac{1}{k}$, it was proved by Suzuki and Yoshida [12], in which they also showed that $E(r_n \cdot z_n) \leqq \frac{k-1}{k} E(t_n) \cdot E(z_n)$, for $z_n > 0$ implies $r_n \leqq \frac{k-1}{k} t_n$, and that $\mathrm{cov}(t_n, z_n) \leqq 0$.

The inequality $\mathrm{cov}(t_n, z_n) \leqq 0$ is shown by the fact that z_n is non-increasing function of t_n for fixed values of $\{w_{n-1,i}\}$ and s_{n-1} from (5.10) and that t_n is independent of $\{w_{n-1,i}\}$ and s_{n-1}. Thus we have $\mathrm{cov}(z_n, r_n) \leqq \frac{1}{k}\{E(s) - E(t)\} \cdot E(z) = \left(\rho - \frac{1}{k}\right)E(t) \cdot E(z)$. And by inserting (5.4) into this, we have

$$\mathrm{cov}(z_n, r_n) \leqq \left(\rho - \frac{1}{k}\right)\frac{1}{\lambda} \cdot \frac{k-1}{k} \cdot \frac{E(s^2)}{2E(s)}, \tag{5.14}$$

from which (5.13) is implied.

The upper bound just obtained seems to be not so good, but it is slightly better than the result obtained in [1] and [6], which is

$$E(w) \leqq J_k + \frac{\frac{k-1}{k^2} E(u^2)}{2E(-u)} .$$

Now we are trying to study special cases.

$\frac{1}{\lambda}$ - MRLA/G/k queue

In this case, if we could give a good lower bound on $E(z)$, we would be able to obtain a nice bound on $E(w)$ from (3.6), (3.10) and (3.19). Even if we imagine the worst case and put $E(z) \geqq 0$, we have

$$(5.15) \quad E(w) \leqq \frac{E(s^2)}{2k(1-\rho)} = \frac{\lambda E(\frac{s^2}{k^2})}{2(1-\rho)} + \frac{\frac{k-1}{k^2}\lambda E(s^2)}{2(1-\rho)}$$

where the first term in the right side is one of the upper bounds on $E(w^{(1)})$ in this case. Thus we may choose the smaller one by enumerating (5.13) and (5.15).

$GI/E_p/k$ queue

In this case we are also trying to evaluate $\mathrm{COV}(z_n, r_n)$ as in the case of *GI/G/k* and to improve the upper bound on $E(w)$. Recall that $(w_{ni} - w_n)$ $(i = 2, 3, \ldots, k)$ are the remaining service times of the customers who are being in service just at the time $T_n + w_n$. Let p_{ni} be the number of the phase of Erlang distribution at this time for the customer with the remaining service time
We put $p_{ni} = 0$, if $(w_{ni} - w_n) = 0$.

And if we know the number p_{ni}, $(w_{ni} - w_n)$ is conditionary independent of r_n, for the time length spent in each phase is exponentially distributed. So we have

$$\begin{aligned} & E([w_{ni} - w_n] \cdot r_n \mid p_{ni}) \\ = \; & E(w_{ni} - w_n \mid p_{ni}) \cdot E(r_n \mid p_{ni}) \\ \leqq \; & E(s) \cdot E(r_n \mid p_{ni}) \end{aligned}$$

where the last inequality is reduced from the property of Erlang distribution. And further by taking expectations of both sides of the above inequality as to p_{ni}, we have

$$E([w_{ni} - w_n] \cdot r_n) \leqq E(s) \cdot E(r) = E(s) \cdot E(-u).$$

Combining (3.6) with the above,

$$(5.16) \quad E(z_n, r_n) \leqq \frac{k-1}{k} E(s) \cdot E(-u),$$

is obtained. Substitute (5.16) into (5.12) and put $\mathrm{var}(r) \geqq 0$, then

$$(5.17) \quad E(w) \leqq J_k + \frac{k-1}{k} E(s) - E(z)$$

is obtained. In this case, however, we cannot give an appropriate lower bound on $E(z)$, so we put $E(z) \geqq 0$ and we have

$$(5.18) \quad E(w) \leqq J_k + \frac{k-1}{k} E(s)$$

for $GI/E_p/k$ queue. This bound gives a much better evaluation than (5.13).

GI/M/k queue

In the system $GI/M/k$ queue, Makino [7] showed that

$$(5.19) \quad E(w^{(1)}) \geqq E(w^{(2)}) \geqq \cdots \geqq E(w)$$

and

$$(5.20) \quad E(w^{(1)}) + \frac{1}{k} E(s) \leqq E(w^{(2)}) + \frac{2}{k} E(s) \leqq \cdots \leqq E(w) + E(s)$$

where $E(w^{(i)}) + \frac{i}{k} E(s)$ is the mean sojourn time in the system for the modified system $GI/G/i$.

And in this case, the distribution functions of the stationary waiting times are represented as

$$(5.21) \quad P\{w^{(i)} \leqq x\} = 1 - \frac{A^{(i)}}{1-\omega} e^{-\mu(1-\omega)x} \qquad (x \geqq 0)$$

where ω is a constant independent of the number of servers i, $A^{(i)}$ is a constant and $\mu = 1/E(s)$ (see Takács [13]). Now we can write $E(w^{(i)}) = \frac{A^{(i)}}{(1-\omega)^2} \cdot E(s)$, so we have $A^{(i)} \geqq A^{(i+1)}$ from (5.19). Thus from this fact

$$(5.22) \quad w^{(1)} \overset{st.}{\geqq} w^{(2)} \overset{st.}{\geqq} \cdots \overset{st.}{\geqq} w$$

is directly implied. Recently Brumelle [2] has also showed the similar results as above.

G/D/k queue

In this case we need not assume that $\{t_n\}$ forms renewal process. Now let it be assumed that $y_0 = 0$, i.e. the system is empty at the time 0. It is noticed that customers are departing from the system in the order of their arrivals as well as from the queue in this case. Thus for the queue $G/D/i$, every j-th customers, i.e. the customers arrived at the times T_{im+j} ($m = 0, 1, 2, \ldots$ and $j = 0, 1, 2, \ldots, i-1$), may receive services from the same server. Then the service commencing time $C_n^{(i)} = T_n + w_n^{(i)}$ for these customers are represented as

$$(5.23)\quad C^{(i)}_{im+j} = \max_{0\leq l\leq m}\{(m-l)\frac{i}{k}b + T_{il+j}\} \qquad (j = 0, 1, 2, \ldots, i-1)$$

in the case of $G/D/1$ where b is the constant service time of the original system $G/D/k$.

By comparing (5.23), we can derive

$$(5.24)\quad C^{(i-1)}_n \geqq C^{(i)}_n \qquad (w.p.\ 1)$$

and

$$(5.25)\quad C^{(i-1)}_n + \frac{i-1}{k}b \leqq C^{(i)}_n + \frac{i}{k}b \qquad (w.p.\ 1)$$

for all n, from which

$$(5.26)\quad w^{(1)}_n \geqq w^{(2)}_n \geqq \ldots \geqq w_n \qquad (w.p.\ 1)$$

and

$$(5.27)\quad w^{(1)}_n + \frac{1}{k}b \leqq w^{(2)}_n + \frac{2}{k}b \leqq \ldots \leqq w_n + b \qquad (w.p.\ 1)$$

are easily shown by using the definition of $C^{(i)}_n$.

Here we are going to prove (5.24) and (5.25) only for the case of $i = 2$. For other cases we can prove similarly, but we abreviate them for brevity. At first we consider the customers arriving at T_{2m} $(m = 0, 1, 2, \ldots)$. Now in this case, we can rewrite (5.23) as

$$C^{(1)}_{2m} = \max_{0\leq l\leq 2m}\{(2m-l)\frac{1}{k}b + T_l\}$$

$$= \max_{0\leq l\leq m}\{(2m-2l)\frac{b}{k} + T_{2l},\ (2m-2l)\frac{b}{k} + \frac{b}{k} + T_{2l-1}\}$$

and

$$C_{2m}^{(2)} = \max_{0 \leq l \leq m} \{(2m - 2l)\frac{b}{k} + T_{2l}\} .$$

Thus by comparing each components in the parethesis in the right sides of both equations, we obtain

$$C_{2m}^{(1)} - \frac{b}{k} \leq C_{2m}^{(2)} \leq C_{2m}^{(1)} , \qquad (w.p.\ 1).$$

For the customers arriving at T_{2m+1}, we can also prove similarly.

Now from the arguments for *GI/M/k* queue and *G/D/k* queue, which are considered as two extreme models among the general *GI/G/k* queueing systems, it is conjectured that the inequalities of the types (5.19), (5.20) and (5.22) will hold for all *GI/G/k* queues.

Acknowledgement

The author wishes to thank Prof. Hidenori Morimura, Prof. Hajime Makabe and Dr. Yukio Takahashi for their guidance and encouragement.

References

[1] Brumelle, S.L., Some inequalities for parallel server queues, Opns. Res., 19 (1971) 402-413.

[2] Brumelle, S.L., Bound on the wait in a GI/M/k queue, Manag. Sci., 19 (1973) 773-777.

[3] Cox, D.R. and Bloomfield, P., A low traffic approximation for queues, J. Appl. Prob., 9 (1972) 832-840.

[4] Kiefer, J. and Wolfowitz, J., On the theory of queues with many servers, Trans. Amer. Math. Soc., 78 (1955) 1-18.

[5] Kingman, J.F.C., Some inequalities for the GI/G/1 queue, Biometrika, 49 (1962) 315-324.

[6] Kingman, J.F.C., Inequalities in the theory of queues, J. Roy. Statist. Soc., ser B, 32 (1970) 102-110.

[7] Makino, T., Investigation of the mean waiting time for queueing systems with many servers, Ann. Inst. Stat. Math., 21 (1968) 357-366.

[8] Marshall, K.T., Some inequalities in queueing, Opns. Res., 16 (1967) 651-658.

[9] Marshall, K.T., Bounds for some generalization of the GI/G/1 queue, Opns. Res., 16 (1968) 841-848.

[10] Mori, M., A note on a (J, X)-process with continuous state space and its applications to the queueing theory, Research Reports on Information Sciences, No. B-2, February, 1974. (Dept. Information Science, Tokyo Institute of Technology.)

[11] Stidham, S., On the optimality of single server queueing systems, Opns. Res., 18 (1970) 708-732.

[12] Suzuki, T. and Yoshida, Y., Inequalities for many-server queue and other queues, J. Opns. Res. Soc. Japan, 13 (1970) 69-77.

[13] Takács, L., Introduction to the theory of queues, (Oxford Univ. Press, New York, 1962).