Journal of the Operations Research
Society of Japan
Vol. 31, No. 2, June 1988

OPTIMAL SPLITTING OF RENEWAL INPUT PROCESS TO A QUEUEING SYSTEM AND ITS APPLICATION TO A NETWORK

Masao Mori
Tokyo Institute of Technology

Hiroshi Shirakawa
Tokyo Institute of Technology

(Received July 22, 1987; Revised October 30, 1987)

Abstract Suppose customers over fraction p should be allocated to a single server queueing system by splitting a given input stream. In this paper we first show that the regular splitting rule, proposed by Hajek [1], minimizes the waiting time of a routed customer in the sense of convex stochastic ordering. However this arrival stream of routed customers is not in general a renewal process, so it is difficult to evaluate characteristics of the waiting time. Then we give upper and lower bounds for the waiting time distribution by using approximated renewal sequences. At last an application of the above evaluating method to a simple network queueing system will be demonstrated.

§1. Introduction

In a multi-queue system, it is a matter to decide to which queue an arriving customer should be delivered. Delivering rules may change depending on which kind of information on the state of the queueing system can be utilized, e.g. current queue lengths, waiting times in each queues and so on. Here we consider the case no information can be available excepting statistical behavior of arrival and service processes. In this context, Bernoulli splitting rule is mainly studied hitherto as seen in Markovian network models. However, we will concentrate on deterministic splitting rules, in which we deliver customers to each queues automatically depending on just the ordering of arriving customers. It is the first aim to find an optimal rule in order to minimize mean waiting time of a customer, given that ratios of delivered customers to each queues are fixed.

In a recent year, Hajek [1] considered the following simple splitting problem of input stream into a queueing system. Customers are arriving according to a renewal sequence and at least $100p\%$ of them are to be routed to a single server queue. Other customers may be sent to another queue or ignored. Here we put $r_k = 1$ if the kth arriving customer is routed to the queue and $r_k = 0$ otherwise. Thus the sequence $r = (r_k)_1^\infty$ represents a (deterministic) routing policy. The sequence $\tilde{r} = (\tilde{r}_k)_1^\infty$ is called "regular" if the components are given by

$$\tilde{r}_k = [kp + a] - [(k-1)p + a], \quad [\ ] = \lfloor\ \rfloor \text{ or } \lceil\ \rceil, \tag{1}$$

for some $a \in [0,1)$, where $\lfloor x \rfloor (\lceil x \rceil)$ denotes the maximum (minimum) integer not larger (smaller) than x.

Let $N_k(r)$ be the number of customers in the queueing system just before the arrival of the kth customer, when a splitting sequence r is used. For $GI/M/1$ queue, Hajek [1] has proved that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E\{\Psi(N_k(r))\} \geq \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E\{\Psi(N_k(\tilde{r}))\}, \tag{2}$$

for any sequence such that $\liminf_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} r_k \geq p$, where Ψ is any nondecreasing convex function. In the derivation of the result, Hajek tactically used the multimodularity property of $N_k(r^{k-1})$ as a function of finite sequence $r^{k-1} = (r_1, r_2, \cdots, r_{k-1})$. The result means that, among all deterministic splitting rules, the regular splitting rule $\tilde{r}$ minimizes the number of customers in the system in the convex ordering sense. In short, in case of $\Psi(x) = x$, we say, the mean system length is minimized by the regular sequence.

In Section 2, we introduce similar results for the waiting time process of $GI/G/1$ queues, which have been derived by Shirakawa, Mori and Kijima [7]. Here we only state some results necessary for the latter sections. Please refer to the paper for detail. Of course the stream of routed customer to the queue is not necessarily renewal type, so it is very difficult to evaluate the exact mean waiting time. In Section 3, upper and lower bounds for the mean waiting time for routed customers are discussed. Strict and detailed discussions on the derivation of bounds on the basis of sample path wise method should be reffered to a coming paper by Shirakawa et al [8]. In Section 4, an application to tree type network systems will be given. There we give a rough evaluation method of a network queue with splitting nodes by using an analogous way to QNA (see Whitt [10]).

§2. Optimal Splitting for a $GI/G/1$ Queue

Consider a single server queueing system with a renewal arrival stream and with a general service time distribution, however all the customers are not served necessarily. At least $100p\%$ of them are routed to have services at the server under *FCFS* discipline, but the rests are ignored or delivered to other systems. Customers having services at the server are called routed customers here. Throughout this section, the following notations are used:

p a fraction of customers to be routed to the server ($0< p <1$).

$\boldsymbol{r} = (r_k)_1^\infty$ a splitting sequence.

$\boldsymbol{k} = (k_n)_0^\infty$, $k_0 = 0$ a sequence of indices of routed customers, which is uniquely determined by $\boldsymbol{r}$.

$\boldsymbol{u} = (u_n)_1^\infty$, $u_n = k_n - k_{n-1}$ the number of arriving customers between the $(n-1)$th and the nth routed customers to the server. ($\boldsymbol{u}$ is also one-to-one corresponding to $\boldsymbol{r}$).

$\boldsymbol{u}^m = (u_1, \cdots, u_m)$ a truncated sequence of $\boldsymbol{u}$.

$\boldsymbol{T} = (T_k)_1^\infty$ an *i.i.d.* random sequence of the interarrival time T_k between the $(k-1)$th and kth arriving customers.

$\boldsymbol{S} = (S_k)_0^\infty$ an *i.i.d.* random sequence of the (potential) service time S_k for the kth arriving customer ($k \geq 1$) and S_0 is the initial backlog of the system.

W_n the waiting time of the nth routed customer under *FCFS* discipline.

It is assumed in the subsequent sections that the queue under consideration is stable, i.e. $pE\{S_k\} < E\{T_k\}$.

For the routed customers, it is easily known that waiting times are recursively generated by

$$W_{n+1} = [W_n + S_{k_n} - \sum_{j=k_n+1}^{k_{n+1}} T_j]^+, \quad n \geq 0, \tag{3}$$

where $[a]^+ = \max(a, 0)$. For notational convenience, when $\boldsymbol{u}^m$ and the random sequences S and T are specified, we write $W_m = W_m(\boldsymbol{u}^m, S, T)$ in an abbreviated manner. Thus it is clear that the mean waiting time of W_m is represented as a function of $\boldsymbol{u}^m$. Let Ψ be any nondecreasing convex function, and put

$$J_m(\boldsymbol{u}^m) = \begin{cases} E\{\Psi(W_m(\boldsymbol{u}^m, S, T))\}, & \text{if } \boldsymbol{u}^m \in Z_+^m, \\ \infty, & \text{otherwise,} \end{cases} \tag{4}$$

where $Z_+^m = \{\boldsymbol{u}^m \in Z^m$; all components of $\boldsymbol{u}^m$ are strictly positive $\}$ and Z^m is the m dimensional lattice space.

Now we introduce the notion of the multimodular function on Z^m (Hajek [1]).

Definition (Multimodularity) A real valued function J on Z^m is said to be multimodular if

$$J(\boldsymbol{u} + \boldsymbol{f}_i) + J(\boldsymbol{u} + \boldsymbol{f}_j) \geq J(\boldsymbol{u}) + J(\boldsymbol{u} + \boldsymbol{f}_i + \boldsymbol{f}_j) \tag{5}$$

for all $\boldsymbol{u} \in Z^m$ and $0 \le i < j \le m$, where $\boldsymbol{f}_i$ $(0 \le i \le m)$ are vectors on Z^m given by

$$\begin{cases} \boldsymbol{f}_0 = (-1, 0, 0, \cdots, 0), \\ \boldsymbol{f}_1 = (1, -1, 0, \cdots, 0), \\ \qquad \cdots\cdots \\ \boldsymbol{f}_m = (0, \cdots\cdots, 0, 1). \end{cases} \tag{6}$$

Then we can show the next attractive property of $J_m(\boldsymbol{u}^m)$, the proof of which is given in [7] following several lemmas.

Theorem 1. $J_m(\boldsymbol{u}^m)$ is a nonincreasing (in the componentwise order) multimodular function.

By using the above property, we got the following useful result :

Theorem 2. If $\boldsymbol{r}$ is any splitting sequence such that $\liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} r_k \ge p$ (that is $\limsup_{\ell\to\infty} \frac{1}{\ell}\sum_{n=1}^{\ell} u_n \le \frac{1}{p}$), then $\liminf_{\ell\to\infty} \frac{1}{\ell}\sum_{m=1}^{\ell} J_m(\boldsymbol{u}^m) \ge \lim_{\ell\to\infty} \frac{1}{\ell}\sum_{m=1}^{\ell} J_m(\tilde{\boldsymbol{u}}^m)$, where $\tilde{\boldsymbol{u}}$ is a regular sequence given by

$$\tilde{u}_n = \left\lceil \frac{n}{p} - \frac{a}{p} \right\rceil - \left\lceil \frac{n-1}{p} - \frac{a}{p} \right\rceil \ \left(\text{or } \tilde{u}_n = \left\lfloor \frac{n}{p} - \frac{a}{p} \right\rfloor - \left\lfloor \frac{n-1}{p} - \frac{a}{p} \right\rfloor \right). \tag{7}$$

The sequence $\tilde{\boldsymbol{u}} = (u_n)_1^\infty$ is corresponding to regular splitting sequence $\tilde{\boldsymbol{r}}$, i.e. $\tilde{r}_k = \lfloor kp + a \rfloor - \lfloor (k-1)p + a \rfloor$ (or $\tilde{r}_k = \lceil kp + a \rceil - \lceil (k-1)p + a \rceil$).

The above theorem states that the regular splitting rule is optimal to minimize the waiting time of a customer in the sense of convex ordering among all deterministic splitting rules with the same splitting ratio. When $\Psi(x) = x$, we can briefly say that the mean waiting time is minimized by the regular splitting rule.

§3. Upper and Lower Bounds of Mean Waiting Time

In this section we will try to evaluate the mean waiting time of the routed customer under the regular splitting rule. However, in general, interarrival times of the routed customers do not form a renewal sequence for regular splitting rules, so it is difficult to enumerate the mean waiting time in an exact sense. For example, in case of $p = \frac{2}{5}$, $\tilde{\boldsymbol{u}} = (3, 2, 3, 2, \cdots)$ by setting $a = 0$, which implies interarrival times process to the server is an alternative renewal.

Here, by approximating the arrival process, we will derive upper and lower bounds of the mean waiting time. However, it should be mentioned that, in the case that the original input stream is PH type (see Neuts [4]) and regular splitting is cyclic, the arrival stream of the routed customers can be also represented as PH type distribution with enlarged phase space and the mean waiting time can be obtained by a numerical computation. The method using PHtype technique is not discussed in this paper.

Upper Bound

We call $U = (U_n)_1^\infty$, the renewal splitting sequence if the number of arriving customers between $(n-1)$th and nth routed customers to the server, i.e. U_n, form a (discrete) renewal sequence. By elaborating the proof of Theorem 2, we derive the following results, the proofs of which are given in [5].

Theorem 3. Let $\tilde{u} = (\tilde{u}_n)_1^\infty$ be a regular sequence given in Theorem 2 and Ψ be any nondecreasing convex function. Then

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{n=1}^{\ell} E\{\Psi(W_n(\tilde{u}))\} \le \lim_{n\to\infty} E\{\Psi(W_n(U))\}, \tag{8}$$

for any renewal splitting sequence $U = (U_k)_1^\infty$ such that $E\{U_n\} = \frac{1}{p}$ and it is independent of $T = (T_k)_1^\infty$ and $S = (S_k)_0^\infty$.

Next we propose a special renewal splitting sequence, which minimizes the above performance measure of the queueing system. We set the renewal splitting sequence $\overline{U} = (\overline{U}_n)_1^\infty$ in the following manner.

$$P\{\overline{U}_n = j\} = \begin{cases} 1 + \left\lfloor \frac{1}{p} \right\rfloor - \frac{1}{p}, & \text{if } j = \left\lfloor \frac{1}{p} \right\rfloor, \\ \frac{1}{p} - \left\lfloor \frac{1}{p} \right\rfloor, & \text{if } j = \left\lfloor \frac{1}{p} \right\rfloor + 1, \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that $E\{\overline{U}_n\} = \frac{1}{p}$ and further that $\overline{U}_n \le_c U_n$ (we say $\overline{U}_n$ is smaller than U_n in the sense of convex order) for any integer valued renewal sequence with mean $E\{U_n\} = \frac{1}{p}$. Here $\overline{U}_n \le_c U_n$ is defined whenever $E\{f(\overline{U}_n)\} \le E\{f(U_n)\}$ holds for any nondecreasing convex function f (Stoyan [9]). The following lemma is a modified version of Stoyan's famous result to the waiting time process of routed customers.

Lemma. Let $U = (U_n)_1^\infty$ and $V = (V_n)_1^\infty$ be any renewal splitting sequences with rate p such that $U_n \le_c V_n$. Then

$$E\{\Psi(W_n(U))\} \le E\{\Psi(W_n(V))\}, \quad n \ge 1, \tag{9}$$

for any nondecreasing convex function Ψ.

That is, we say $W_n(U)$ is larger than $W_n(V)$ in the sense of convex order, abbreviated to $W_n(U) \leq_c W_n(V)$. By using Theorem 3 and the above lemma, the following main result will be directly implied.

Theorem 4. For any nondecreasing convex function Ψ, the renewal splitting sequence $\overline{U}$ given by (9) generates the least upper bound for a regular sequence among all renewal splitting sequences in the sense that

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{n=1}^{\ell} E\{\Psi(W_n(\tilde{u}))\} \leq \lim_{n\to\infty} E\{\Psi(W_n(\overline{U}))\} \leq \lim_{n\to\infty} E\{\Psi(W_n(U))\}, \tag{10}$$

where $U = (U_n)_1^\infty$ is any renewal splitting sequence with $E\{U_n\} = \frac{1}{p}$.

Thus for the mean waiting time, the least upper bound for a regular splitting sequence is given by the mean waiting time based on $\overline{U}$ among all renewal splitting sequences. We use the value as an upper bound of the mean waiting time for a regular splitting sequence.

Lower Bound

Concerning the lower bound, we restrict ourselves to queues where the interarrival time distribution $A(x)$ of the original input stream is *GPH* type, i.e. general phase type distribution. *GPH* is introduced by Shanthikumar [6], and the class of *GPH* is shown to be broader than the class of *PH* proposed by Neuts [4]. And also Otto [5] discusses *GPH* type queueing model in a different way, in which *GPH* is named as almost phase type (*APH*). The Laplace transform of a *GPH* distribution is given by

$$A^*(s) = E\left\{\left(\frac{\lambda}{\lambda+s}\right)^M\right\}, \tag{11}$$

where $\frac{1}{\lambda}$ is the mean sojourn time in each phase decaying in exponential law and M is a nonnegative integer valued random variable. That is, *GPH* is a mixture of *Erlang* distributions with the same scale parameter λ, which permits infinite mixtures. Let $(g_k)_0^\infty$ denotes the distribution of M, and we can write

$$A(x) = \sum_{k=0}^{\infty} g_k E_k(x;\lambda), \tag{12}$$

where $E_k(x;\lambda)$ is the k-phase *Erlang* distribution function of scale parameter λ.

Theorem 5. Suppose original input stream is a *GPH* renewal type with a distribution given by (12). Then for any nondecreasing convex function Ψ, we have

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{n=1}^{\ell} E\{\Psi(W_n(\tilde{u}))\} \geq \lim_{n\to\infty} E\{\Psi(W_n(\Gamma_{\frac{m}{p}}(\lambda)/G/1))\}. \tag{13}$$

Here $W_n(\Gamma_{\frac{m}{p}}(\lambda)/G/1)$ denotes the waiting time for the nth customer, under *FCFS* discipline, of the queueing system with *Gamma* renewal input of scale parameter λ and of shape parameter $\frac{m}{p}$ where $m = E\{M\}$ in (12).

The proof needs somewhat intricated discussion (see [8] in detail), however the above statement might be easily understood in an intuitive manner. It is noticed that the nth interarrival time of customers routed to the server is represented as the $\tilde{u}_n$ sum of T_j's, where T_j's are distributed with a *GPH* distribution of (13) (see the equation (13)). So the Laplace transbe written as $[E\{(\frac{\lambda}{\lambda+s})^M\}]^{\tilde{u}_n}$. On the other hand, we have $(\frac{\lambda}{\lambda+s})^{\frac{m}{p}} = \{(\frac{\lambda}{\lambda+s})^m\}^{\frac{1}{p}}$ for a $\Gamma_{\frac{m}{p}}(\lambda)$ input, where $\frac{1}{p} = \lim_{\ell\to\infty} \frac{1}{\ell}\sum_{n=1}^{\ell} \tilde{u}_n$ and $m = E\{M\}$ are constant values, which are less in variability than random variables $\tilde{u}_n$ and M respectively. Thus the interarrival times of customers for the regular splitting rule are expected to be much larger in variability than *r.v.*s with the distribution $\Gamma_{\frac{m}{p}}(\lambda)$.

Remark 1. We consider the case that $(A^*(s))^{\frac{1}{p}}$ has a corresponding distribution function for p, where $A^*(s)$ denotes the Laplace transform of the distribution function of original arrival process. For example, if $A(x)$ is a degenerate distribution or an infinitely divisible distribution, the above makes sense for any $0 < p \le 1$. In this case, it can be shown that the lower bound is replaced by $\lim_{n\to\infty} E\{\Psi(W_n((A^*(s))^{\frac{1}{p}}/G/1))\}$, i.e. the performance measure for the $GI/G/1$ queueing system with interarrival distribution $(A^*(s))^{\frac{1}{p}}$ in Laplace Transform.

Remark 2. In order to enumerate the upper and lower bounds of the mean waiting time, we can use the very nice approximation formula given by Krämer and Langenbach-Berz [2].

§4. An Application to a Network Queue

Here we consider an open queueing network system with infinite waiting room at each service nodes, but the structure of network is tree type (Figure 1). This type of system is often observed as a production system and so on. In this paper we restrict models to the cases that the number of branching flows from each splitting point is 2. For the cases, if a splitting sequence for one queue makes a regular sequence, then the rest sequence of discarded customers (i.e. routed to another service node) is necessarily regular too. However, when the number of splitting flows $n \ge 3$, regular splitting sequences may not be constructed for some routes, depending on splitting ratio vector $p = (p_1, p_2, \cdots, p_n)$. For example, in case of $p = (\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$ r sequence can be constructed easily. But, in case of $p = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$, we cannot realize three regular flows to be routed to each queue without overlapping assignments. In order to overcome such difficulties, we will consider cyclic splitting sequences and propose methods evaluating lower and upper bounds of mean waiting times in a forthcoming paper.

In order to clarify our method of evaluating network queueing system, we consider here the model shown in Figure 1. We assume the original input stream to the network is a *Poisson* process with rate λ and service time distribution at service node i is exponential with rate μ_i. At each service node customers are served by a single server under *FCFS* discipline. And at each splitting points, customers are routed to queue i by regular splitting rule with splitting ratio p^i. Here the decomposition method proposed in *QNA* by Whitt [10] is adopted. That is to say, we decompose the network into a set of imaginary $GI/M/1$ queues. And we are going to enumerate approximating values of mean queue length at each service nodes by using the upper and lower bounds models stated in Section 3. However, we have to notice that the values enumerated in the above manners do not give the lower and upper bounds respectively in exact sense for latter stages, for the departure process from the former service node is not a renewal process, which makes it difficult to apply results in Section 3 in a strict sense.

But we are eagerly concerned in measuring the effect of introducing the simple control of arrival customers with a regular splitting rule. By using the above approximate values, we can compare the rule with the ordinary *Bernoulli* splitting rule. Numerical example will teach us how largely we can decrease the congestion in the network by adopting a regular splitting rule.

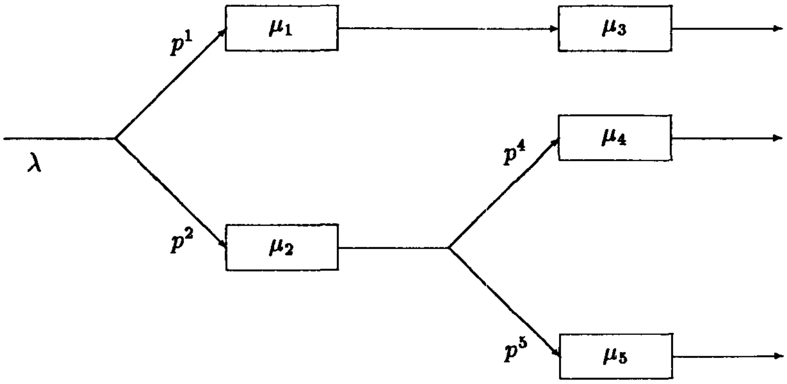

Figure 1. An example of tree type network ($p^1 + p^2 = 1, p^4 + p^5 = 1$)

Approximations by upper bound model

For the service node i $(= 1, 2)$, we use the model $(\beta_i E_{\ell_i}(\lambda) + (1-\beta_i)E_{\ell_i+1}(\lambda))/M(\mu_i)$ /1 queue, i.e. the interarrival time distribution is a mixture of *Erlang* distributions, where $\ell_i = \left\lfloor \frac{1}{p^i} \right\rfloor$ and $\beta_i = 1 + \left\lfloor \frac{1}{p^i} \right\rfloor - \frac{1}{p^i}$. The inter-departure time distribution from this service node is represented as

$$\tilde{D}_i \equiv (g_0 E_0 + g_1 E_1(\lambda) + \cdots + g_{\ell_i+1} E_{\ell_i+1}(\lambda)) * M(\mu), \tag{14}$$

where E_0 is the distribution with unit mass at 0 and "$*$" denotes convolution operation.

This departure process will be discussed in Appendix. Thus, for the service node 3, we use the model $\tilde{D}_1/M(\mu_3)/1$ queue. And for the node i $(= 4, 5)$, we consider the model $(\beta_i \tilde{D}_2^{*\ell_i} + (1-\beta_i)\tilde{D}_2^{*(\ell_i+1)})/M(\mu_i)/1$ queue, where $\tilde{D}^{*k}$ is the k-fold convolution of $\tilde{D}$. The values of mean queue length enumerated for these models are shown in Table 2 at the row notated as "*U.A.*". This enumeration procedure is tiresome for latter stage service nodes.

As the second approximation procedure, we introduce the spirit of *QNA* [10] much more and use Krämer and Langenbach-Berz approximation formula for mean queue size. Here the coefficient of variation of the departure process from some service node is estimated as

$$C_d^2 = (1-\rho^2)C_a^2 + \rho^2 C_b^2, \tag{15}$$

where C_a and C_b are the coefficients of variation of the arrival and service processes for the node. And the coefficient of variation of arrival process of splitted customers (with splitting ratio p) is given by

$$C_a^2 = \frac{C_{ao}^2}{\ell + 1 - \beta} + \frac{\beta(1-\beta)}{(\ell + 1 - \beta)^2}, \tag{16}$$

where *beta* and ℓ are given in upper bound model and C_{ao} is the coefficient of variation of the original input stream to the splitting point. (17) is easily derived from the relation for the arrival process of upper bound model that

$$A = \beta A_o^{*\ell} + (1-\beta)A_o^{*(\ell+1)}, \tag{17}$$

stated in Section 3, where A_o is the distribution of the original arrival process before splitting. The approximate values obtained by this method are shown at the row "*K.L.U.*" in Table 2, which give fairly good evaluation for "*U.A.*" values.

Approximations by lower bound model

Concerning the lower bound model, we can get the lower bound in a sense stated in Section 3 for the first stage service nodes i (= 1, 2), by using the model $\Gamma_{\frac{1}{p_i}}(\lambda)/M(\mu_i)/1$, where $\Gamma_\alpha(\lambda)$ is the *Gamma* distribution with shape parameter α and scale parameter λ. However, the inter-departure distribution from the service node i cannot be represented in a explicit form, which makes it difficult to apply the lower bound evaluation method directly to the latter stage's service nodes.

So we consider the QNA method analogous to the second upper bound approximation method. The only different point from the upper bound approximation method is to use

$$C_a^2 = pC_{ao}^2 \tag{18}$$

for a splitted flow, in stead of using (17). The above relation (19) is implied by the statements in Remark 1 below Theorem 5. The approximate values calculated by using this lower model are shown at the row "*K.L.L.*" in Table 2, which seem to give us nice estimations by comparing with the values obtained through simulation.

We repeat to notice that, by this approach, an upper and a lower bound are available only for the first stage nodes, but just approximate values are given for latter stage nodes. And the relative error of these approximations are not so small yet. This is because "*renewal*" property collapses much for latter stages.

However, Table 2 tells that a control of arrival process by using regular splitting rule brings significant effect compared with using probabilistic routing rule. And we see that the queue sizes for nodes with small splitting ratio decrease much compared with the case of *Bernoulli* splitting rule.

Remark 3. "*K.L.U.*" and "*K.L.L.*" approximation methods are applicable to much more general queueing network systems with general interarrival and service time distributions and having merging points of customer flows by using the method of QNA.

Table 1. Parameter sets for numerical evaluation.

Case	λ	μ_1	μ_2	μ_3	μ_4	μ_5	p^1	p^4
1	1.0	1.0	1.0	1.0	0.5	0.5	0.5	0.5
2	1.0	1.4	0.6	1.4	0.42	0.18	0.7	0.7

In each cases, traffic intencities for each service nodesame value $\rho = 0.5$.

Table 2. Values of Average Waiting Number of Customers for Q_i ($A.W.Q_i.$) and Average Waiting Number of Customers in the Network ($A.W.N.$).

	Routing		$A.W.Q_1.$	$A.W.Q_2.$	$A.W.Q_3.$	$A.W.Q_4.$	$A.W.Q_5.$	$A.W.N.$
1	*B.S.*		0.5	0.5	0.5	0.5	0.5	2.5
	R.S.	*U.A.*	0.309	0.309	0.400	0.257	0.257	1.527
		Simu.	0.301 ±0.0095	0.304 ±0.0085	0.378 ±0.0095	0.237 ±0.0092	0.245 ±0.0106	1.465 –
		K.L.A.	0.309	0.309	0.383	0.258	0.258	1.517
2	*B.S.*		0.5	0.5	0.5	0.5	0.5	2.5
	R.S.	*U.A.*	0.441	0.243	0.466	0.337	0.198	1.685
		Simu.	0.401 ±0.0102	0.244 ±0.0112	0.429 ±0.0116	0.273 ±0.0131	0.168 ±0.0115	1.515 –
		K.L.U.	0.449	0.261	0.463	0.320	0.196	1.689
		K.L.L.	0.384	0.232	0.435	0.266	0.186	1.504

In case 1, upper and lower models are definitely the same. *Simu.* : 20 repetitions of 9600 time units of simulations are excused, and 90% confidence interval are given in the Table. *B.S.*=Bernoulli splitting rule. *R.S.*=Regular splitting rule. *U.A.*=Upper bound approximation. *K.L.A.*=Krämer and Langenbach-Berz (K.L.) approximation. *K.L.U.*=K.L. approximation of upper bound. *K.L.L.*=K.L. approximation of lower bound.

Concluding Remarks

In this paper the waiting behaviour of splitted customers is studied at first. Briefly to say, the regular splitting rule is optimal to minimize the mean waiting time of a routed customer to the server among all deterministic splitting rules. And in the section 3, upper and lower bounds of the mean waiting time for the rule are derived.

In the section 4, for simple queueing networks with splitting nodes, a method of evaluating system performances approximately for the rule is proposed on the basis of the QNA method.

In a routing problem for a network, a nicely designed dynamic routing policy, which may depend on current states of queue sizes at each nodes, can decrease congestions much. However, the overhead time arising in communicating the current data of the system states and in enumeration of selecting routes may be considerably large and cannot be ignored. In fact, it is reported by Maruya [3], through a simulation experiment, that the joining to the shortest queue rule become worse than the regular splitting rule, when a considerably large time lag occurs in getting information on queue sizes before each parallel servers. Therefore, we think, it is worthwhile to study on effects of introducing such a simple control policy like the regular splitting rule into routing problems of queueing networks.

In this paper we just propose the method to evaluate system performances roughly under the condition that splitting ratios are given at each splitting nodes. However, a significant problem how we should set such splitting ratios remains yet to consider minimizing congestions in a network.

Acknowledgement

We thank Mr. Machida for his sincere assist in doing numerical experiments. And also we sincerely appreciate refrees' valuable comments to revise the paper.

Appendix

Inter-departure Time Distribution of $GI/M/1$ Queue

We first consider the idle time distribution $I(t)$ of $GI/M/1$ queue with interarrival distribution $A(x)$ and with service rate μ. It is assumed that the system is in steady state. Let T_n be a random variable of interarrival time between the $(n-1)$th and the nth customers, σ_j be the sum of j service times and $\tilde{I}$ be a generic random variable of an idle time. Then we have

$$\begin{aligned} &P\{\text{no customer stays just before the } n\text{-th arrival and } \tilde{I} \leq t\} \\ &= \sum_{k=0}^{\infty} P\{k \text{ customers stay just before the } (n-1)\text{th arrival and } T_n - \sigma_{k+1} \leq t\} \\ &= \sum_{k=0}^{\infty} \pi_k \int_0^{\infty} A(t+x)\frac{(\mu x)^k}{k!}\mu e^{-\mu x}dx, \end{aligned}$$

where $(\pi_k)_0^\infty$ is the stationary queue size distribution observing just before arrivals. By using geometric property of $(\pi_k)_0^\infty$, i.e. $\pi_k = \pi_0(1-\pi_0)^k$, we easily derive

$$I(t) = \int_0^\infty A(t+x)e^{-\pi_0 \mu x}\mu dx. \qquad (19)$$

Notice that this distribution is fairly different from the stationary residual life time distribution of the interarrival time. In heavy traffic case, i.e. $\pi_0 \to 0$ and $\mu \to \frac{1}{a}$ (mean inter-arrival time), they are very close. Thus the inter-departure time distribution $D(t)$ is given by

$$D^*(s) = (1-\pi_0+\pi_0 I^*(s))\frac{\mu}{\mu+s} \qquad (20)$$

in Laplace Transform. In $E_\ell/M/1$ queue, $I^*(s)$ is represented as

$$I^*(s) = \sum_{k=1}^{\ell} \left(\frac{\mu}{\lambda + \pi_0\mu}\right) \left(\frac{\lambda}{\lambda + \pi_0\mu}\right)^{\ell-k} \left(\frac{\lambda}{\lambda + s}\right)^{k} \tag{21}$$

which is a finite mixture of $E_k(\lambda)$. The fact implies that the idle time distribution of $GPH/M/1$ queue become a GPH distribution too.

The coefficient of variation, C_d, of the departure process is given by

$$C_d^2 = \frac{1}{\ell} + 2\rho\left(1 - \frac{1-\rho}{\pi_0}\right) \tag{22}$$

for $E_\ell/M/1$ queues.

References

[1] Hajek, B. : Extremal Splitting of Point Processes, *Math. of O.R.*, **10** (1985), 543–556.

[2] Krämer, W. and Langenbach-Berz, M. : Approximate Formulae for the Delay in the Queueing System $GI/G/1$, In :*Proc. 8th Int. Teletraffic Congress.* (1976), 235/1–8.

[3] Maruya,M. : Parallel Queues with Delayed Information, Master Thesis, Department of System Sciences, Tokyo Institute of Technology (1987).

[4] Neuts,M.F. *Matrix-Geometric Solutions in Stochastic Models,* Johns Hopkins University Press, Baltimore, London, 1981.

[5] Otto,T.J. : On the Stationary Waiting-time Distribution in the GI/G/1 Queue, I: Transform Methods and Almost-Phase-Type Distributions, *Adv. Appl. Prob.*, **19** (1987), 240–265.

[6] Shanthikumar, J, G. : Bilateral Phase-Type Distributions, *Naval Res. Logi. Qurt.*, **32** (1983), 119--136.

[7] Shirakawa, H., Mori, M. and Kijima, M. : Further Properties of Extremal Sequences in Queues, to appear in *Stochastic Models.*

[8] Shirakawa, H., Mori, M. and Kijima, M. : Evaluation of Regular Splitting Queues, in preparation.

[9] Stoyan, D. *Comparison Methods for Queues and Other Stochastic Models,* John Wiley & Sons, 1983.

[10] Whitt, W. : The Queueing Network Analyzer, *The Bell System Tech. Jour.*, **62** (1983), 2779–2815.

Masao MORI : Department of Industrial
Engineering and Management,
Tokyo Institute of Technology, Ookayama,
2-12-1, Meguro, Tokyo, 152, Japan.