Probability in the Engineering and Informational Sciences, **22**, 2008, 273–286. Printed in the U.S.A.
doi: 10.1017/S0269964808000168

INSENSITIVITY OF A FRONT-END WEB SYSTEM

GENJI YAMAZAKI
Department of Production & Information Systems
Tokyo Metropolitan Institute of Technology
Hino, Tokyo 191-0065, Japan

TAMOTSU SOGO
Doctoral Program in Intelligent Systems
Tokyo Metropolitan Institute of Technology
Hino, Tokyo 191-0065, Japan
E-mail: e25129@jp.ibm.com

In many companies, legacy systems have been used to serve customers arriving at service counters. The demand of a customer arriving at a counter is divided into R subdemands (SDs). Each SD is processed sequentially by the legacy system. When the final SD service is completed, the customer leaves the counter. On the other hand, Internet users desire that each SD be processed through the Internet without going to a service counter. We note that the applications for operating the web system differ from those of the legacy systems. For that reason, companies might use legacy systems for customers who come to the company through the Internet, without changing those legacy systems' applications. This web system, which is integrated with the legacy system, is called a *front-end web system*. Suppose that, for the web system, demands are generated according to a Poisson process. Then we show that the loss probability and macrostate distribution for the front-end web system are *insensitive* with respect to the distribution of the discrete random variable R, aside from the continuous distributions of other uncertain factors, under a restriction.

1. INTRODUCTION

Many companies' legacy systems have been used to serve customers who arrive at service counters. Recently, Internet users have been increasing rapidly. They typically seek to process their demands through the Internet without going to service counters.

© 2008 Cambridge University Press 0269-9648/08 $25.00

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

For that reason, the companies are interested in applying their legacy systems to serve the needs of customers through the Internet. Such a new web system must be integrated with the legacy system. We note that the applications for operating the web system differ from those of legacy systems. Suppose, for example, that a customer arrives at a counter at an ATM system in a bank. The demands of a customer are separable into R subdemands (SDs). The customer inputs the first SD using a terminal on the counter, called a *host point* (HP). The SD is processed using the legacy system. Then the response is sent to the customer via the HP. With some time lag, the customer then inputs the second SD using the same HP. This procedure is repeated until the customer receives the response to the last SD. Then he leaves the counter and the HP becomes available when the customer receives the final response; that is, the ATM system applications are constructed for each customer, but not for each SD. Suppose that the same customer attempts to process the demand through the Internet. First, he must to obtain a free terminal in the web system, called a *browser point* (BP). After catching the BP, say BP 1 (Fig. 1), he must seek a free HP via BP 1. When he finds such a HP, say HP 1 (Fig. 1), the first SD is sent to the legacy system via BP 1 and HP 1. The response is sent to the customer via HP 1 and BP 1. Then BP 1 becomes free, but HP 1 is retained for use by the customer. He must catch a new free BP to input the second SD. The second SD is then sent to the legacy system via the obtained BP and HP 1 when he finds a BP; that is, the web system applications are constructed for each SD.

Suppose that the web system is integrated with the legacy system using b BPs and h HPs without changing the legacy system applications. This integrated web system is called a *front-end web system*. The design issues of the front-end web system arise from the difference between the respective applications of the two systems. Are there simple and useful stochastic models for resolving those design issues and evaluating the performance measures of the front-end web system? The word *simple* means that it is easy to calculate the model performance. In this context, Sogo,

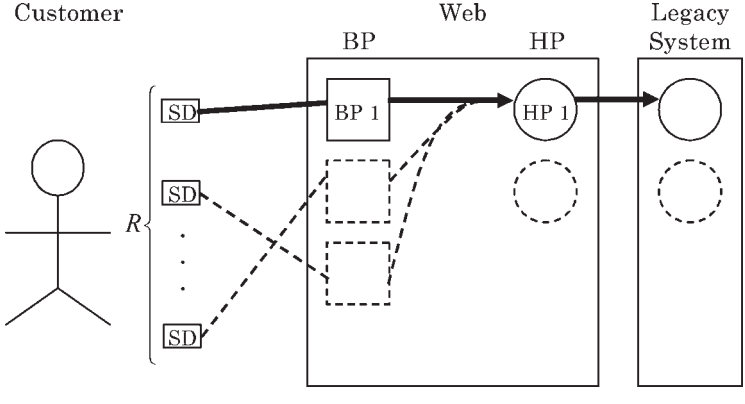


Figure 1. Processing the customer's demands through the Internet.

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

Yamazaki, and Yamamoto [3] considered two special models. One is a model in which b is arbitrary, but $h = 1$. The other is a model in which h is arbitrary, but $b = 1$. Imposing a restriction, they showed that the *loss probabilities* and *macrostate distribution* (defined in Section 3) for the two models are *insensitive* w.r.t. the distribution of R, in addition to the continuous distributions of other uncertain factors when the demands (through the Internet) are generated according to a Poisson process. The definition of insensitivity is identical to that used in the assessment of queuing systems or networks. For insensitivity of queuing systems, see, for example, Kelly [1]. Using numerical experiments, Sogo et al. [3] conjectured that the property can be extended to a model in which both b and h are arbitrary under the restriction. Those results suggest that it is possible to resolve design issues using a model in which R is *random* (defined in Section 2) and the continuous distributions are exponential. This model is simplest from a viewpoint of numerical analysis.

This article presents the argument that the conjecture is true. An example is given that shows that the imposed restriction cannot be removed for insensitivity.

2. MODEL

We consider the following front-end web system. The respective numbers of BPs and HPs are b and h. Customers' demands (through the Internet) are generated according to a Poisson process at the rate of λ. Suppose that a demand is generated and that the demand is divisible into R subdemands (jobs), which are numbered 1 through R. When there are no free BPs on the generating epoch, the demand is lost (BP-loss). When there exist free BPs, the customer seeks a free HP with a free BP (e.g., BP 1). This *seeking* is designated as *service 0*. When the customer finds a free HP (e.g., HP 1), job 1 is sent immediately to the legacy system via BP 1 and HP 1; then it is processed. That *processing* is referred to as *service 1*. When there are no free HPs, the demand is also lost (HP-loss). When the service 1 of job 1 is completed, the response is sent immediately to the customer via HP 1 and BP 1. Subsequently, BP 1 becomes free, but HP 1 is retained for use by the customer. After that customer spends some time thinking, he tries to send job 2 to the legacy system via a new free BP. That *thinking* is referred to as *service 2*. If the customer cannot find a free BP by the first trial, he tries to find a free BP with some time lag. He repeats the trials until he finds a free BP. The *lag* between the two adjacent trials is denoted as *service 3*. When the customer finds a free BP, job 2 is sent immediately to the legacy system via the BP and HP 1 because the HP is retained for his use. This procedure is repeated until the customer receives the response to the final job (job R). When he receives that response, he leaves the system and HP 1 becomes free.

We assume the following:

(i) The R is a random variable (r.v.), independent of all else. The distribution is given by

$$P(R = k) = \bar{\alpha}_{k-1}\alpha_k, \qquad k = 1, 2, \ldots,$$

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

where

$$\bar{\alpha}_k = \prod_{j=1}^{k}(1-\alpha_j) \quad \text{and} \quad \bar{\alpha}_0 \equiv 1.$$

(ii) The service times for service 0 form of a sequence of independent and identically distributed (i.i.d.) r.v.s. The service times for service i, $i = 1, 2$, form sequences of i.i.d. r.v.s, not depending on the job's number. The service times for service 3 form a sequence of i.i.d. r.v.s, not depending on the job's and trial's numbers. The four sequences are mutually independent.

By S_i and G_i we denote the generic service time for service i and its distribution (G for general). We assume $P(S_i = 0) = 0$. The system is referred to as the (G_0, G_1, G_2, G_3; b, h) model. Let $ES_i = \mu_i^{-1}$, $\rho_i = \lambda/\mu_i$, and $ER = \bar{R}$. When R is a geometric r.v. (i.e., $\alpha_n = \alpha$, $n = 1, 2, \ldots$), the R is called *random*.

3. RESULTS

Let $N_i(t)$ be the number of jobs in service i, $i = 0, \ldots, 3$, at time t in the (G_0, G_1, G_2, G_3; b, h) model. Throughout the article, it is assumed that the joint process

$$(N_0(t), N_1(t), N_2(t), N_3(t))$$

is stationary. The most important performance measure for the (G_0, G_1, G_2, G_3; b, h) model is the fraction of demands lost. We note that this loss results from the B-loss and H-loss. The three loss probabilities, denoted by P_{loss}, $P_{\text{B-loss}}$ and $P_{\text{H-loss}}$, respectively, are defined as

$$\begin{aligned} P_{\text{loss}} &= \frac{\lambda - \lambda_d}{\lambda}, \\ P_{\text{B-loss}} &= \frac{\lambda - \lambda_{0,d}}{\lambda}, \\ P_{\text{H-loss}} &= \frac{\lambda_{0,d} - \lambda_d}{\lambda_{0,d}}, \end{aligned} \tag{3.1}$$

where λ_d and $\lambda_{0,d}$ are the departure rates of demands from the system and service 0, respectively.

Sogo et al. [3] considered the insensitivity of the model by assuming that the distributions of S_2 and S_3 are the same (i.e., $G_3 = G_2$). The model under the assumption is simply referred to as the (G_0, G_1, G_2; b, h) model. They showed that the distribution of the macroprocess

$$X(t) = (N_0(t), N_1(t), N_2(t) + N_3(t))$$

and the loss probabilities for the (G_0, G_1, G_2; 1, h) or (G_0, G_1, G_2; b, 1) model are *insensitive* w.r.t the distribution of R in addition to G_i, $i = 0, 1, 2$. The following

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

results, which are proved in Section 4, state that the insensitivity can be extended to the model in which both b and h are arbitrary.

THEOREM 1: *For the (G_0, G_1, G_2; b, h) model, we have*

$$P(l,m,n) \equiv P[N_0(t) = l, N_1(t) = m, N_2(t) + N_3(t) = n]$$
$$= \frac{\rho_0^l \beta_1^m \beta_2^n}{l!m!n!} P_0 \quad \text{for } l + m \le b \quad \text{and} \quad m + n \le h, \tag{3.2}$$

$$P_0 \equiv P(0,0,0) = \left[\sum_{j=0}^{\min(b,h)} \frac{\beta_1^j}{j!} A_{b-j}(\rho_0) A_{h-j}(\beta_2) \right]^{-1}, \tag{3.3}$$

where

$$A_n(x) = \sum_{r=0}^{n} \frac{x^r}{r!}, \quad \beta_1 = \rho_1 \bar{R}, \quad \text{and} \quad \beta_2 = \rho_2(\bar{R} - 1).$$

THEOREM 2: *For the (G_0, G_1, G_2; b, h) model, we have*

$$P_{\text{loss}} = 1 - P_0 \sum_{j=1}^{\min(b,h)} \frac{\beta_1^{j-1}}{(j-1)!} A_{b-j}(\rho_0) A_{h-j}(\beta_2) \tag{3.4}$$

and

$$P_{\text{B-loss}} = P_0 \sum_{j=0}^{\min(b,h)} \frac{\beta_1^j}{j!} \frac{\rho_0^{b-j}}{(b-j)!} A_{h-j}(\beta_2). \tag{3.5}$$

We note that $P_{\text{H-loss}}$ can be determined from (3.1), (3.4), and (3.5).

The assumption $G_2 = G_3$ in the theorems might be somewhat technical. Service 3 corresponds to the time between the two adjacent trials to catch a free BP. The S_3 can be viewed as a constant if the trials are done automatically. However, the assumption is essential for the insensitivity, as shown in the following example.

Example 1: The loss probability for (M, M, M, CX_2; 1, 1) model. We consider the following model with $b = h = 1$. The R is random. The distributions of S_i, $i = 0, 1, 2$, are exponential, denoted by M, and the distribution of S_3 is 2-Coxian, denoted by CX_2, which is a phase-type distribution. The Laplace–Stieltjes transform of CX_2 is defined as

$$\tilde{f}(s) = p \frac{\nu_1}{s + \nu_1} + (1 - p) \frac{\nu_1 \nu_2}{(s + \nu_1)(s + \nu_2)}.$$

For convenience, we set the three parameters as

$$\nu_1 = 2\mu_3, \quad \nu_2 = \mu_3 x, \quad p = 1 - x/2.$$

The mean and the square of coefficient of variation for S_3, then, are $1/\mu_3-$ and $1/x$ ($x \le 2$). The departure rate of demands for the (M, M, M, CX_2; 1, 1) model can be

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

expressed as

$$\lambda_d = \mu_1 \alpha P[N_0(t) = 0, N_1(t) = 1, N_2(t) = 0, N_3(t) = 0]$$
$$= \lambda \rho_1^{-1} \alpha P[N_0(t) = 0, N_1(t) = 1, N_2(t) = 0, N_3(t) = 0].$$

Combining this with (3.3) yields

$$P_{\text{loss}} = 1 - \rho_1^{-1} \alpha P[N_0(t) = 0, N_1(t) = 1, N_2(t) = 0, N_3(t) = 0].$$

Solving the system of balance equations for the model leads to

$$P[N_0(t) = 0, N_1(t) = 1, N_2(t) = 0, N_3(t) = 0]$$
$$= \frac{\rho_1(\rho_0 + \rho_2 + \rho_0\rho_2)(\rho_3 + \rho_0\rho_3 + 2\rho_0 x)}{a_0\rho_3(1 + \rho_0) + a_1\rho_0 x},$$

where

$$a_0 = (\rho_0 + \rho_2 + \rho_0\rho_2)[\rho_1 + (1 - \alpha)\rho_2 + (1 + \rho_0)\alpha]$$
$$+ \rho_0\rho_2(1 - \alpha)(2\rho_0 + \rho_3 + \rho_0\rho_3),$$
$$a_1 = a_0 + (\rho_0 + \rho_2 + \rho_0\rho_2)[\rho_1 + (1 - \alpha)\rho_2 + (1 + \rho_0)\alpha].$$

It can be checked that $P[*]$ is an increasing function of x and, hence, P_{loss} is a decreasing function of x.

Remark: We derive the distribution of the process $N(t) = N_1(t) + N_2(t) + N_3(t)$ and P_{loss} for the $(G_0, G_1, G_2; \infty, h)$ model from the results for the $(G_0, G_1, G_2; b, h)$ model. For this, we view $P(l, m, n)$, P_0, and P_{loss} for the $(G_0, G_1, G_2; b, h)$ model as functions of b, denoted by $P(l, m, n; b)$, $P_0(b)$, and $P_{\text{loss}}(b)$, respectively. Equation (3.2) yields

$$P(m, n; b) \equiv P[(N_1(t) = m, N_2(t)) + N_3(t) = n; b]$$
$$= \sum_{l=0}^{b-m} P(l, m, n; b) = \frac{\beta_1^m \beta_2^n}{m!n!} A_{b-j}(\rho_0) P_0(b).$$

Because

$$\lim_{b\to\infty} A_{b-j}(\rho_0) = e^{\rho_0} \quad \text{for a fixed } j,$$

we can find that

$$P(m, n; \infty) \equiv \lim_{b\to\infty} P(m, n; b) = \frac{\beta_1^m \beta_2^n}{m!n!} \left[\sum_{j=0}^{h} \frac{\beta_1^j}{j!} A_{h-j}(\beta_2) \right]^{-1}. \tag{3.6}$$

Let us consider the process $N(t) = N_1(t) + N_2(t)$ for the $(G_0, G_1, G_2; \infty, h)$ model,

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

which can be interpreted as the process of the number of jobs in the system. Then

$$P[N(t) = k; \infty] = \sum_{m,n;m+n=k} P(m, n; \infty) \quad \text{for} \quad k \leq h. \tag{3.7}$$

The sum is over m and n subject to $m + n = k$. Calculation yields

$$\sum_{j=0}^{h} \frac{x^j}{j!} A_{h-j}(y) = A_h(x + y)$$

and

$$\sum_{m,n;m+n=k} \frac{x^m y^n}{m!n!} = \frac{(x + y)^k}{k!.}.$$

Substituting (3.6) into (3.7) and using these equalities lead to

$$P(N(t) = k; \infty) = \frac{\rho^k}{k!} [A_h(\rho)]^{-1},$$

where

$$\rho = \lambda \bar{S}, \qquad \bar{S} \equiv \bar{R}/\mu_1 + (\bar{R} - 1)/\mu_2.$$

$\{P(N(t)=k; \infty)\}$ is the distribution for the number of jobs in an $M/G/h/h$ loss model whose mean service time is $\bar{S}$.

Similarly, we can find that

$$P_{\text{loss}}(\infty) \equiv \lim_{b \to \infty} P_{\text{loss}}(b) = \frac{\rho^h}{h!} [A_h(\rho)]^{-1},$$

which is *Erlang's loss formula* for the $M/G/h/h$ model.

4. PROOFS OF THEOREMS 1 AND 2

The assumption $G_3(\mu_3) = G_2(\mu_2)$ in Theorems 1 and 2 permits the following interpretation. Suppose that a customer receives the response of a SD, say job i. After he spends time for thinking, he tries to catch a free BP for sending job $i + 1$ to the legacy system. If he cannot find such a BP, he enters in thinking (service 2) again. This is repeated until he finds a free BP. The $N_2(t) + N_3(t)$ in X(t) in Section 3 can be replaced with $N_2(t)$ in this sense. Hence, the macrostate is redefined as

$$X(t) = (N_0(t), N_1(t), N_2(t)).$$

We will prove (3.2) and (3.3) for the $X(t)$ by using the result by Miyazawa [2]. For this, we need to modify the (G_0, G_1, G_2; b, h) model as follows. The jobs in service i, $i = 0$, 1, 2, are ordered and, hence, service i contains jobs in

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

positions $1, 2, \ldots, n_i$, where n_i is the total number of jobs in service i. Then the model operates the following manner:

1. The job in each position at service i is processed by each single server.
2. When job 1 arrives at service 0, the job moves into position l ($l = 1, 2, \ldots, n_0 + 1$) with probability $1/(n_0 + 1)$ if $n_0 + n_1 < b$; jobs previously in positions $l, \ldots, n_0$ move to positions $l + 1, \ldots, n_0 + 1$, respectively. The whole demand is lost if $n_0 + n_1 = b$. When service of the job (job 1) in position l is completed, jobs in positions $l + 1, \ldots, n_0$ move to positions $l, \ldots, n_0 - 1$, respectively. The completed job moves to service 1 if $n_1 + n_2 < h$. If $n_1 + n_2 = h$, the demand including this job is lost.
3. When a job arrives at service 1, the job enters position l ($l = 1, \ldots, n_1 + 1$) with probability $1/(n_1 + 1)$; jobs previously in positions $l, \ldots, n_1$ move to positions $l + 1, \ldots, n_1 + 1$, respectively. Suppose that the job's number in position l is r and that service of the job is completed. Then the job moves to service 2 as job r or becomes the final job (SD) of a demand with probability $1 - \alpha_r$ or α_r. The latter means that the customer sending the demand leaves the system. Jobs previously in positions $l + 1, \ldots, n_1$ move to positions $l, \ldots, n_1 - 1$, respectively.
4. When a job arrives at service 2, the job enters position l ($l = 1, \ldots, n_2$) with probability $1/(n_2 + 1)$. Suppose that the job's number in position l is q and that service of the job is completed. Then the job moves to service 1 as job $q + 1$ if $n_0 + n_1 < b$; jobs in positions $l + 1, \ldots, n_2$ move to positions $l, \ldots, n_2 - 1$, respectively. If $n_0 + n_1 = b$, the completed job remains the same position and starts to receive new service 2. This is repeated until $n_0 + n_1$ is less than b; that is, the job finds a free BP. When the job finds the BP, it moves to service 1 as job $q + 1$.

It is clear that the stationary distribution of $X(t)$ and loss probabilities for the modified model are identical with those for the original (G_0, G_1, G_2; b, h) model. In this section, the modified model is referred to as the (G_0, G_1, G_2) model simply. For the model, let

$$\bar{N}_i(t) = (N_i(t); J_{i,1}, \ldots, J_{i,N_i(t)})$$

be the state of service i, $i = 1, 2$, at time t, where $J_{i,l}$ denotes the job's number in position l, $l = 1, \ldots, N_i(t)$. Note that we do not need to specify the job's number in position l, $l = 1, 2, \ldots, N_0(t)$, for service 0 because the numbers are all 1. Assuming that the microstate process

$$Y(t) = (N_0(t), \bar{N}_1(t), \bar{N}_2(t))$$

is stationary, we have the following result, which means that the stationary distribution of $Y(t)$ is insensitive w.r.t. the service time distributions.

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

LEMMA 1: *The distribution of Y(t) for the* (G_0, G_1, G_2) *model is expressed as*

$$\begin{aligned} &P[n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})] \\ &\quad \equiv P[N_0(t) = n_0, (N_1(t); J_{1,1}, \ldots, J_{1,N_1(t)}) \\ &\quad = (n_1; r_1, \ldots, r_{n_1}), (N_2(t); J_{2,1}, \ldots, J_{2,N_2(t)}) = (n_2; q_1, \ldots, q_{n_2})] \\ &\quad = \frac{\rho_0^{n_0}}{n_0!}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!} P_0 \prod_{l=1}^{n_1} \bar{\alpha}_{r_l - 1} \prod_{l=1}^{n_2} \bar{\alpha}_{q_l} \quad \textit{for } n_0 + n_1 \le b \textit{ and } n_1 + n_2 \le h, \end{aligned} \tag{4.1}$$

where $P_0 = P[N_0(t) = 0, N_1(t) = 0, N_2 = 0]$.

PROOF: Along with the (G_0, G_1, G_2) model, we need to construct the corresponding *exponential model*. The exponential model is obtained by replacing G_0, G_1, and G_2 in the original model with exponential distributions with the same means, respectively. Let S be the state space of $Y(t)$. For the exponential model, suppose that (1) *partial balance equations* (PBEs) hold for each $s \in S$; (2) there exists a proper probability distribution satisfying the system of PBEs and the global balance equations. The result by Miyazawa [2], then, states that the probability distribution gives the stationary distribution of $Y(t)$ for the (G_0, G_1, G_2) model.

We first show (1). For each state $s \in S$, we define two sets A_s^e and A_s^g. A_s^e has only one element, a generation of new demand, whose life (service) time is according to an exponential distribution with mean λ^{-1} because the generating process is Poisson (e for exponential). A_s^g is the set of positions in service i, $i = 0, 1, 2$, whose service time distributions are general in the (G_0, G_1, G_2) model (g for general). For example, consider the state

$$s_0 = (n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})) \in S.$$

By p_l^i, we denote position l, $l = 1, \ldots, n_i$, in service i, $i = 0, 1, 2$. Then we have

$$A_{s_0}^g = \{p_1^0, \ldots, p_{n_0}^0; p_1^1, \ldots, p_{n_1}^1; p_1^2, \ldots, p_{n_2}^2\}.$$

For an $a \in A_{s_0}^g$, let $\mu_{(s_0,a)\rightarrow}$ be the transition rate from s_0 to another state by completing service of a. Suppose that there are other m_a states, say $s_1, \ldots, s_{m_a}$, such that the state $s_j \in S, j = 1, \ldots, m_a$, changes to state s_0 by completing $a_j \in A_{s_j}^e \cup A_{s_j}^g$ and starting new service in the position a. Let $\mu_{(s_0,a)\leftarrow(s_j,a_j)}$ be this transition rate. The PBE for a is given as

$$\mu_{(s_0,a)\rightarrow} P(s_0) = \sum_{j=1}^{m_a} \sum_{a_j \in A_{s_j}^e \cup A_{s_j}^g} \mu_{(s_0,a)\leftarrow(s_j,a_j)} P(s_j).$$

The PBEs for all elements of $A_{s_0}^g$ are called the PBEs for s_0. For the s_0, let

$$n_0 + n_1 < b, \qquad n_1 + n_2 < h, \quad \text{and} \quad r_l > 1, \qquad l = 1, \ldots, n_1.$$

Then the PBEs for s_0 can be written as follows.

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

For $p_l^0 \in A_{s_0}^{\mathrm{g}}$, $l = 1, \ldots, n_0$,

$$\mu_0 P[n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})] = \frac{\lambda}{n_0} P[n_0 - 1, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})]. \quad \textbf{(4.2)}$$

The right-hand side (RHS) in (4.2) means that a new demand generates and its job 1 enters in p_l^0 with probability $1/n_0$.

For p_l^1, $l = 1, \ldots, n_1$,

$$\mu_1 P[n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})] = \sum_{j=1}^{n_2+1} \frac{\mu_2}{n_1} P[n_0, (n_1 - 1; r_1, \ldots, r_{l-1}, r_{l+1}, \ldots, r_{n_1}), \times (n_2 + 1; q_1, \ldots, q_{n_2}; r_l - 1 \to p_j^2)], \quad \textbf{(4.3)}$$

where

$$\begin{aligned}
&(n_1 - 1; r_1, \ldots, r_{l-1}, r_{l+1}, \ldots, r_{n_1}) \\
&\quad = (N_1(t) = n_1 - 1; J_{1,1} = r_1, \ldots, J_{1,l-1} = r_{l-1}, \\
&\qquad J_{1,l} = r_{l+1}, \ldots, J_{1,n_1-1} = r_{n_1}), \\
&(n_2 + 1; q_1, \ldots, q_{n_2}; r_l - 1 \to p_j^2) \\
&\quad = (N_2(t) = n_2 + 1; J_{2,1} = q_1, \ldots, J_{2,j-1} = q_{j-1}, \\
&\qquad J_{2,j} = r_l - 1, J_{2,j+1} = q_j, \ldots, J_{2,n_2+1} = q_{n_2}).
\end{aligned}$$

The RHS in (4.3) means that service of the job in p_j^2 is completed and the job enters in p_l^1 with probability $1/n_1$.

For p_l^2, $l = 1, \ldots, n_2$,

$$\mu_2 P[n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})] = \sum_{j=1}^{n_1+1} \frac{\mu_1}{n_2} (1 - \alpha_{q_l}) P[n_0, (n_1 + 1; r_1, \ldots, r_{n_1}; q_l \to p_j^1), (n_2 - 1; q_1, \ldots, q_{l-1}, q_{l+1}, \ldots, q_{n_2})], \quad \textbf{(4.4)}$$

where the definition of $(n_1 + 1;\ r_1, \ldots,\ r_{n_1};\ q_l \to p_j^1)$ or $(n_2 - 1;\ q_1, \ldots,\ q_{l-1}, q_{l+1}, \ldots, q_{n_2})$ is similar to that of $(n_2+1;\ q_1, \ldots,\ q_{n_2};\ r_1-1 \to p_j^2)$ or $(n_1-1;\ r_1, \ldots, r_{l-1}, r_{l+1}, \ldots, r_{n_1})$. The RHS means that when service of the job in p_j^1 is completed, the job (job q_l) moves to service 2 with probability $1-\alpha_{q_l}$ and enters in p_l^2 with probability $1/n_2$.

For the s_0, the job's numbers in service 1 are all greater than 1. When some of them are equal to 1, the PBEs for the corresponding positions are different from

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

(4.3). For example, consider the same state as s_0 except for $r_1 = 1$, denoted by

$$s_0' = (n_0, (n_1; 1, r_2, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})],$$

where $r_l > 1$, $l = 2, \ldots, n_1$. The PBE for $p_l^0 \in A_{s_0'}^g$, or $p_l^2 \in A_{s_0'}^g$, is similar to (4.2) or (4.4). The PBE for p_l^1, $l = 2, \ldots, n_1$, is similar to (4.3), but the PBE for p_1^1 becomes

$$\mu_1 \Pr[n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})]$$
$$= \sum_{j=1}^{n_0+1} \frac{\mu_0}{n_1} Pr[n_0 + 1, (n_1 - 1; r_2, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})].$$

This reflects the fact that only job 1 moves from service 0 to service 1.

Similarly, we can write the PBEs for all other states belonging to S. Solving for $P(s)$'s in terms of $P(0)$ from the system of PBEs yields (4.1).

The process $Y(t)$ for exponential model is a Markov process. The global balance equations in item (1) earlier means the (usual) balance equations for the Markov process. For example, the balance equation for the above $s_0 \in S$ can be written as7

$$(\lambda + n_0\mu_0 + n_1\mu_1 + n_2\mu_2)P[n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})]$$
$$= \lambda P[n_0 - 1, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})]$$
$$+ \sum_{j=1}^{n_2} \sum_{l=1}^{n_1+1} \frac{\mu_1}{n_2}(1 - \alpha_{q_j}) P[n_0, (n_1 + 1; r_1, \ldots, r_{n_1}; q_j \to p_l^1),$$

$$(n_2 - 1; q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_{n_2})]$$
$$+ \sum_{r=1}^{\infty} \sum_{l=1}^{n_1+1} \mu_1 \alpha_r P[n_0, (n_1 + 1; r_1, \ldots, r_{n_1}; r \to p_l^1), (n_2; q_1, \ldots, q_{n_2})]$$
$$+ \sum_{j=1}^{n_1} \sum_{l=1}^{n_2+1} \frac{\mu_2}{n_1} P[n_0, (n_1 - 1; r_1, \ldots, r_{j-1}, r_{j+1}, \ldots, r_{n_1}),$$
$$(n_2 + 1; q_1, \ldots, q_{n_2}; r_j - 1 \to p_l^2)].$$

Substituting $P(s)$'s, $s \in S$, given by (4.1) into this equation, we can check that the $P(s)$'s satisfies this equation. Similarly, we can show that the distribution given by (4.1) satisfies the balance equations for all other states; that is, the distribution satisfies item (2). It remains to find the normalizing constant P_0. This will be done in the following proof of Theorem 1.

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

PROOF OF THEOREM 1: For the (G_0, G_1, G_2) model, we have

$$\begin{aligned} P(n_0, n_1, n_2) &\equiv P[N_0(t) = n_0, N_1(t) = n_1, N_2(t) = n_2] \\ &= \sum_{r_1=1}^{\infty} \cdots \sum_{r_{n_1}=1}^{\infty} \sum_{q_1=1}^{\infty} \cdots \sum_{q_{n_2}=1}^{\infty} P[n_0, (n_1; r_1, \ldots, r_{n_1}), \\ &\quad (n_2; q_1, \ldots, q_{n_2})]. \end{aligned} \tag{4.5}$$

Because

$$\bar{R} = \sum_{k=1}^{\infty} kP(R = k) = \sum_{k=0}^{\infty} P(R > k),$$

we have

$$\sum_{r_l=1}^{\infty} \bar{\alpha}_{r_l - 1} = \bar{R} \tag{4.6}$$

and

$$\sum_{q_l=1}^{\infty} \bar{\alpha}_{q_l} = \bar{R} - 1. \tag{4.7}$$

Substituting (4.1) into (4.5) and using (4.6) and (4.7) leads to (3.2).

The remainder of proof is to determine the normalizing constant P_0. The constant can be obtained by setting

$$\sum_{n_0} \sum_{n_1} \sum_{n_2} P(n_0, n_1, n_2) = 1. \tag{4.8}$$

Note that $n_0 + n_1 \leq b$, $n_1 + n_2 \leq h$, and, hence, $n_1 \leq \min(b, h)$. For fixed n_0 and n_1, (3.2) yields

$$\begin{aligned} P(n_0, n_1) &\equiv P[N_0(t) = n_0, N_1(t) = n_1] \\ &= \sum_{n_2=0}^{h-n_1} P(n_0, n_1, n_2) = \frac{\rho_0^{n_0} \beta_1^{n_1}}{n_0! n_1!} A_{h-n_1}(\beta_2) P_0. \end{aligned}$$

From this, we have

$$\begin{aligned} P(n_1) &\equiv P[N_1(t) = n_1] \\ &= \sum_{n_0=0}^{b-n_1} P(n_0, n_1) = \frac{\beta_1^{n_1}}{n_1!} A_{b-n_1}(\rho_0) A_{h-n_1}(\beta_2) P_0. \end{aligned}$$

Combining this with (4.8) yields (3.3). ■

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

PROOF OF THEOREM 2: For the (G_0, G_1, G_2) model, let $\lambda_d(*)$ be the departure rate of demands (or customers) from the system when the state of the model is $*$. We note that each demand leaves the model only when service 1 of the final job (SD) is completed. For $n_1 \geq 1$, suppose

$$Y(t) = (n_0, (n_1; r_1, \ldots, r_{n_1}), (n_2; q_1, \ldots, q_{n_2})) \equiv s_0.$$

Let $R_l^1(t)$ be the remaining service time of job r_l in p_l^1 at time t, $l = 1, \ldots, n_1$. Then the result by Miyazawa [2] states that $R_l^1(t)$ is independent of all else and the distribution function is given by

$$P[R_l^1(t) \geq x)] = \mu_1 \int_0^x [1 - G_1(x)]\, dx,$$

where $G_1(x)$ is the distribution function of S_1. This means that the completion rate of a job in service 1 at time t is μ_1, not depending on the job number and the service time distribution. Noting that job r_l is the final SD of a demand with probability α_{r_l}, we have

$$\lambda_d(s_0) = \mu_1(\alpha_{r_1} + \alpha_{r_2} + \cdots + \alpha_{r_{n_1}})P(s_0).$$

Combining this with (4.1) yields

$$\begin{aligned}
&\lambda_d[(N_0(t) = n_0, \bar{N}_1(t) = (n_1; r_1, \ldots, r_{n_1}), N_2(t) = n_2)] \\
&\quad = \mu_1(\alpha_{r_1} + \alpha_{r_2} + \cdots + \alpha_{r_{n_1}})P_0 \frac{\rho_0^{n_0}\rho_1^{n_1}\rho_2^{n_2}}{n_0!n_1!n_2!} \prod_{l=1}^{n_1} \bar{\alpha}_{r_l - 1} \sum_{q_1=1}^{\infty} \cdots \sum_{q_{n_2}=1}^{\infty} \prod_{l=1}^{n_2} \bar{\alpha}_{q_l}.
\end{aligned}$$

Taking account of (4.7), this can be rewritten as

$$\begin{aligned}
&\lambda_d[(N_0(t) = n_0, \bar{N}_1(t) = (n_1; r_1, \ldots, r_{n_1}), N_2(t) = n_2)] \\
&\quad = \mu_1(\alpha_{r_1} + \alpha_{r_2} + \cdots + \alpha_{r_{n_1}})P_0 \frac{\rho_0^{n_0}\rho_1^{n_1}\beta_2^{n_2}}{n_0!n_1!n_2!} \prod_{l=1}^{n_1} \bar{\alpha}_{r_l - 1}.
\end{aligned}$$

Calculation yields

$$\begin{aligned}
\lambda_d[(n_0, n_1, n_2)] &\equiv \lambda_d[(N_0(t) = n_0, N_1(t) = n_1, N_2(t) = n_2)] \\
&= \sum_{r_1=1}^{\infty} \cdots \sum_{r_{n_1}=1}^{\infty} \lambda_d[(N_0(t) = n_0, \\
&\quad \bar{N}_1(t) = (n_1; r_1, \ldots, r_{n_1}), N_2(t) = n_2)] \\
&= \mu_1 \frac{\rho_0^{n_0}\rho_1^{n_1}\beta_2^{n_2}}{n_0!n_1!n_2!} n_1 \bar{R}^{n_1 - 1} P_0 \\
&= \lambda \frac{\rho_0^{n_0}\beta_1^{n_1 - 1}\beta_2^{n_2}}{n_0!(n_1 - 1)!n_2!} P_0.
\end{aligned}$$

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press

In the second equality, (4.6) and

$$\sum_{r_l=1}^{\infty} \bar{\alpha}_{r_l-1}\alpha_{r_l} = \sum_{k=1}^{\infty} P(R=k) = 1$$

were used. Note that λ_d in (3.1) can be written as

$$\lambda_d = \sum_{n_1=1}^{\min(b,h)} \sum_{n_0=0}^{b-n_1} \sum_{n_2=0}^{h-n_1} \lambda_d[(n_0,n_1,n_2)]. \tag{4.9}$$

Calculation yields

$$[\text{RHS in (4.9)}] = \lambda P_0 \sum_{n_1=1}^{\min(b,h)} \frac{\beta_1^{n_1-1}}{(n_1-1)!} A_{b-n_1}(\rho_0) A_{h-n_1}(\beta_2).$$

Combining this with (3.2) yields (3.4).

Because of PASTA, the B-loss probability can be written as

$$\begin{aligned} P_{\text{B-loss}} &= P[N_0(t)+N_1(t)=b] \\ &= \sum_{n_1=0}^{\min(b,h)} \sum_{n_2=0}^{h-n_1} P[N_0(t)=b-n_1, N_1(t)=n_1, N_2(t)=n_2]. \end{aligned}$$

From this and (3.2), (3.5) follows. ■

Acknowledgments

We thank Professor M. Miyazawa, Science University of Tokyo, for valuable discussions. This research was conducted while the first author was visiting the University of California, Berkeley. The first author was supported in a part by grant from Tokyo Metropolitan Government and by a grant (No. 15500047) from the Japan Society for the Promotion of Science.

References

1. Kelly, F.P. (1979). *Reversibility and stochastic networks*. New York: Wiley.
2. Miyazawa, M. (1993). Insensitivity and product-form decomposability of reallocatable GSMP. *Advances in Applied Probability*, 25:415–437.
3. Sogo, T., Yamazaki, G., & Yamamoto, H. (2003) Modeling for design of a front-end web system, the Transactions of the Institute of Electronics, Information and Communication Engineers. B, J86-B, 20, 2035-2040 (in Japanese).

https://doi.org/10.1017/S0269964808000168 Published online by Cambridge University Press