Queueing Systems 9 (1991) 337–352

Invited paper

A tandem Jackson network with feedback to the first node

Jonathan Brandon and Uri Yechiali

Department of Statistics, School of Mathematical Sciences, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Received 26 February 1990; revised 12 November 1990

An N-node tandem queueing network with Bernoulli feedback to the end of the queue of the *first* node is considered. We first revisit the single-node $M/G/1$ queue with Bernoulli feedback, and derive a formula for $EL(n)$, the expected queue length seen by a customer at his nth feedback. We show that, as n becomes large, $EL(n)$ tends to $\rho/(1-\rho)$, ρ being the effective traffic intensity. We then treat the entire queueing network and calculate the mean value of S, the total sojourn time of a customer in the N-node system. Based on these results we study the problem of *optimally ordering the nodes so as to minimize ES*. We show that this is a special case of a general sequencing problem and derive sufficient conditions for an optimal ordering. A few extensions of the serial queueing model are also analyzed. We conclude with an appendix in which we derive an explicit formula for the correlation coefficient between the number of customers seen by an arbitrary arrival to an $M/G/1$ queue, and the number of customers he leaves behind him upon departure. For the $M/M/1$ queue this coefficient simply equals the traffic intensity ρ.

Keywords: Tandem queues, Jackson network, feedback, sequencing.

1. Introduction

The following model was studied in [13]: A single job is made up of N independent tasks, all of which must be successfully performed for the job to be completed. The tasks are performed sequentially and may be attempted in *any* order, where each attempt of task i requires a random time X_i and is successful with probability p_i ($1 \leqslant i \leqslant N$). Upon failure at any stage the job has to be started all over again, i.e., the job is fed back to the *first* stage. It was shown that $E(X)$, the expected total time to complete the job, is minimized if the tasks are indexed such that $EX_i/(1-p_i)$ is an increasing sequence, and that

$$E(X) = \sum_{i=1}^{N} \left[E(X_i) \Big/ \left(\prod_{j=i}^{N} p_j \right) \right]. \tag{1}$$

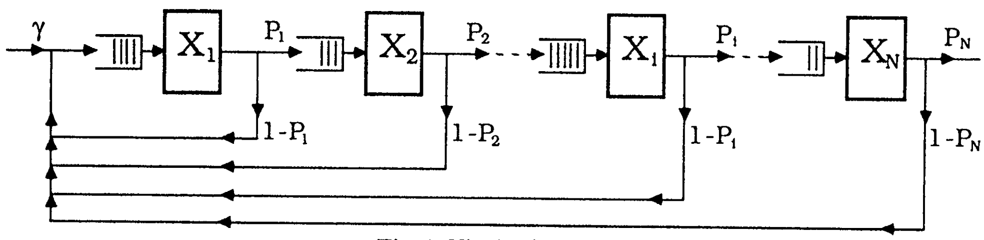

Fig. 1. The basic model.

An application of such a model could be, for example, a manufacturing process where at the end of each stage the product is tested and if it does not comply with a certain standard it is completely dismantled and the whole process starts again.

In this work we extend the above model to an N-node tandem *Jackson network* with a special set of switching probabilities, i.e., upon completion of service at node i each customer either moves forward to node $i+1$ (with probability p_i) or is *fed back* (with probability $1-p_i$) to the *end* of the queue of the *first* node. An illustration of the model is depicted in fig. 1, where γ is the rate of the *external* Poisson arrival.

Our goal is to characterize orderings of the nodes so as to minimize the expected *total* sojourn time of a customer in the system.

Jackson networks have been extensively studied in the literature and various general results have been obtained (cf. Gelenbe and Mitrani [8]). Nevertheless, we concentrate on the particular configuration of fig. 1, which may be considered as a generalization of the simplest form of Jackson network, this being a single-node queue with Bernoulli feedback.

Takács [12] studied the $M/G/1$ queue with Bernoulli feedback to the end of the queue, and derived the Laplace–Stieltjes transform (LST) of the sojourn time of a customer in the system. Disney, Konig and Schmidt [6] further analyzed Takács' model. They concentrated on the time spent by a customer waiting in queue at each visit, depending upon the epochs at which the queue is being observed (i.e. arrivals, departures, instants of feedback, joining of the queue or end of service). Laplace transforms were developed for each case. Van den Berg, Boxma and Groenendijk [4] considered an $M/G/1$ queue where the customer is fed back a *fixed* number of times. Following this, van den Berg and Boxma [1,2] treated the case where the probability of success changes with each feedback. Additional investigations of sojourn times of feedbacking customers in closely related models are reported in Doshi and Kaufman [7], Boxma and Cohen [5] and van den Berg and Boxma [3].

In section 2 we revisit the $M/G/1$-queue-with-feedback and derive a formula for $EL(n)$, the expected queue length seen by a customer at his nth feedback. We show that, as n becomes large, $EL(n)$ tends to $\rho/(1-\rho)$, where ρ is the effective traffic intensity. In section 3 we analyze the N-node tandem Jackson network with feedback to the *first* node, and calculate ES, the mean total sojourn time of a customer in the network, giving an intuitive interpretation to the result. In section 4 we consider the problem of *optimally* ordering the nodes so as to minimize ES. We show that this is a special case of a general sequencing problem and derive sufficient conditions for an optimal ordering. A few extensions are analyzed in section 5. We conclude the paper with an appendix in which we derive the correlation coefficient between the number of customers seen by an arbitrary arrival to an $M/G/1$ queue, and the number of customers he leaves behind him upon departure. For the $M/M/1$ queue this coefficient simply equals the traffic intensity ρ.

2. The $M/G/1$ queue with feedback revisited

Consider a single-server queue with Poisson arrival rate γ, service times V and feedback to the *end* of the queue with probability $(1-p)$, which we denote by $M/G^f/1$ (see Disney et al. [6]).

Our goal is to derive an explicit formula for the mean queue size $E[L(n)]$ seen by a customer at his nth feedback.

It is well known (cf. Takács [12]) that as far as the variable of interest is the number of customers in the system at an arbitrary point of time (denoted by L), then the system is equivalent to a regular $M/G/1$ queue with the service time $\hat{V}$ being a geometric sum of ordinary service times V_j's, i.e., $\hat{V}=\sum_{j=1}^{N} V_j$ where $P(N=k)=(1-p)^{k-1}p \quad (k=1, 2,\dots,)$. The Pollaczek–Khintchine formula yields

$$EL=\rho+\frac{\gamma^2 E[\hat{V}^2]}{2(1-\rho)},$$

where $\rho=\gamma E[\hat{V}]$, $E[\hat{V}]=E[V]/p$, $E[\hat{V}^2]=E[V^2]/p+\{2(1-p)E^2[V]/p^2\}$. Thus,

$$\begin{aligned} EL&=\frac{\gamma\{2E[V]+\gamma E[V^2]-2\gamma E^2[V]\}}{2(p-\gamma E[V])}\\ &=\frac{\rho}{1-\rho}+\frac{\gamma^2}{2}\frac{\{E[V^2]-2E^2[V]\}}{p(1-\rho)}. \end{aligned} \tag{2.1}$$

Consider an arbitrary customer C that arrives at time t_0. Let S be its *total* time in the system. Then,

$$\begin{aligned} E[S]=&\pi_0 E[V]+\sum_{i=1}^{\infty}\pi_i E[R_i]+E[L]E[V]\\ &+\sum_{i=1}^{\infty}(1-p)^n\{E[L(n)]+1\}E[V], \end{aligned} \tag{2.2}$$

where $L(n)$ is the number of customers found by C in the system *upon his nth feedback* ($n \geqslant 1$), and R_i is the *residual* service time of the customer being served when C arrives, *given* that there are i customers in the system ($R_0 \equiv V$).

We first calculate $\sum_{i=1}^{\infty}\pi_i E[R_i]$. Clearly, $\sum_{i=1}^{\infty}\pi_i E[R_i] = \rho E[R \mid L \geqslant 1]$, where R is the residual service time of the customer being served at time t_0. Let $\hat{R}_i = R_i + \sum_{j=1}^{\hat{N}} V_j$, where $\hat{N}$ is a shifted Geometric distribution such that $P(\hat{N} = k) = (1-p)^k p$ $(k = 0, 1, 2, \ldots)$. $\hat{R}_i$ is the residual *total* service time to be devoted to the customer being served at time t_0, *given* that i customers are present. Now,

$$E[R_i] = E[\hat{R}_i] - E[\hat{N}] \cdot E[V] = E[\hat{R}_i] - \left(\frac{1}{p} - 1\right) E[V]. \tag{2.3}$$

Following Mandelbaum and Yechiali [11] we have

$$\sum_{i=1}^{\infty} \pi_i E[\hat{R}_i] = \rho E[\hat{R} \mid L \geqslant 1] = \rho \frac{E[\hat{V}^2]}{2E[\hat{V}]},$$

where $\hat{R}$ is the *total* additional service time that will be devoted to the customer being served when C arrives. Hence, using (2.3),

$$\sum_{i=1}^{\infty} \pi_i E[R_i] = \rho \frac{E[\hat{V}^2]}{2E[\hat{V}]} - \left(\frac{1}{p} - 1\right) E[V](1 - \pi_0). \tag{2.4}$$

Substituting $E[\hat{V}]$, $E[\hat{V}^2]$ and $(1 - \pi_0) = \rho$, in (2.4) we obtain

$$\sum_{i=1}^{\infty} \pi_i E[R_i] = \gamma E[V^2]/2p. \tag{2.5}$$

Thus,

$$E[R \mid L \geqslant 1] = \sum_{i=1}^{\infty} \pi_i E[R_i]/\rho = \frac{E[V^2]}{2E[V]}. \tag{2.6}$$

That is, the expected remaining time until completion of the "current" service in an $M/G^f/1$ queue is the *same* as its corresponding value in a *regular* $M/G/1$ queue.

We now turn to calculate $EL(n)$ for $n \geqslant 1$.

$$E\{L(1) \mid L = i\} = \gamma\{E[R_i] + iE[V]\} + (1-p)i, \tag{2.7}$$

where the first term is the expected number of new arrivals, and the second term gives the expected number of feedbacks during the time between C's arrival and his first feedback. Thus,

$$\begin{aligned} E[L(1)] &= \pi_0 \gamma EV + \sum_{i=1}^{\infty} \pi_i E[L(1) \mid L = i] \\ &= \theta E[L] + \frac{2\gamma E[V](p - \gamma E[V]) + \gamma^2 E[V^2]}{2p} \\ &= \theta E[L] + \gamma E[V](1-\rho) + \frac{\gamma^2 E[V^2]}{2p}, \end{aligned} \tag{2.8}$$

where $\theta \equiv \gamma E[V] + (1-p)$. For the $(n+1)$st feedback, $n \geqslant 1$, we can write a recursive equation for $L(n+1)$ in terms of $L(n)$:

$$L(n+1) = N\left(\sum_{j=1}^{L(n)+1} V_j\right) + B(1-p, L(n)),$$

where $N(x)$ denotes the number of Poisson arrivals during a time interval of length x, V_j are all distributed like V, and $B(1-p, L(n))$ is a Binomial random variable resulting from $L(n)$ Bernoulli trials with individual probability of "success" $1-p$.

The generating function $G_{L(n+1)}(z) \equiv E[z^{L(n+1)}]$ is derived as

$$G_{L(n+1)}(z) = \tilde{V}[\gamma(1-z)] G_{L(n)}\big(\tilde{V}[\gamma(1-z)][p + (1-p)z]\big).$$

It readily follows that for $n \geqslant 2$,

$$E[L(n) \mid L(n-1)] = \gamma E[V][L(n-1)+1] + (1-p)L(n-1),$$

which can also be obtained directly by setting $E[V]$ instead of $E[R_i]$ in (2.7). Hence,

$$E[L(n)] = \theta E[L(n-1)] + \gamma E[V] \quad (n = 2, 3, 4, \ldots), \tag{2.9}$$

so that

$$E[L(n)] = \theta^{n-1} E[L(1)] + \left[\frac{1-\theta^{n-1}}{1-\theta}\right] \gamma E[V] \quad (n = 2, 3, 4, \ldots). \tag{2.10}$$

Therefore, with the use of (2.8), eq. (2.10) becomes

$$E[L(n)] = \theta^n E[L] + \theta^{n-1} \xi + \left(\frac{1-\theta^{n-1}}{1-\theta}\right) \gamma E[V], \tag{2.11}$$

where

$$\xi \equiv \gamma E[V](1-\rho) + \frac{\gamma^2 E[V^2]}{2p}.$$

In a stationary queue $\rho < 1$, so that $\theta = \gamma E[V] + (1-p) < 1$. Thus, after some algebra, we get that $\sum_{n=1}^{\infty}(1-p)^n E[L(n)] = ((1-p)/p)E[L]$, *as if* $EL(n) = EL$ for all n (which, of course, is *not* true in general). Another observation is that the probability of $n \geqslant 1$ feedbacks by C, given that he feeds back at all, is $(1-p)^n p/(1-p)$. Therefore, the mean system size seen by a fed back customer is given by

$$\sum_{n=1}^{\infty} \left[(1-\rho)^n p/(1-p)\right] EL(n) = EL.$$

Substituting (2.11) and (2.5) in equation (2.2), we obtain (see Takács [12])

$$\begin{aligned} E[S] &= (1-\rho)E[V] + \frac{\gamma E[V^2]}{2p} + E[L] \cdot E[V] + \frac{1-p}{p}(E[L]+1)E[V] \\ &= \frac{E[L]+1}{p} EV - \rho[EV - ER]. \end{aligned} \tag{2.12}$$

Equation (2.12) may be rewritten so as to give further insight, i.e.

$$E[S] = \rho\left[E[R] + \frac{E[L]}{\rho}E[V]\right] + (1-\rho)E[V] + \frac{1-p}{p}(E[L]+1)E[V]. \tag{2.13}$$

The first and second terms in (2.13) give the expected time until first service completion by C, given that he found the server busy or idle, respectively. The third term equals $\sum_{n=1}^{\infty}(1-p)^n[L(n)+1]E[V]$, which is the expected sojourn time due to feedbacks. Clearly, by Little's law, $E[L] = \gamma E[S]$, as can be verified by looking at eq. (2.1).

It is of interest to note that

$$E[L(n)] \xrightarrow[n \to \infty]{} \left[\frac{\gamma E[V]}{p - \gamma E[V]}\right] = \frac{\rho}{1-\rho}. \tag{2.14}$$

That is, after many feedbacks C observes a mean queue size that is *independent* of the second moment of the service time and is the *same* for all service time distributions possessing the same mean. The explanation is that the second moment affects mainly new arrivals (through the residual service time), while feedback customers find *no* customer in service.

It should also be pointed out that $E[L(n)] = \theta^{n-1}[\theta E[L] + \xi - \rho/(1-\rho)] + \rho/(1-\rho)$ is a *monotone* sequence: if $\alpha \equiv \theta EL + \xi - \rho/(1-\rho) = (\gamma^2/2p)$ $[E[V^2] - 2E^2[V]][1/(1-\rho) - p + 1]$ is positive (negative) then $E[L(n)]$ is decreasing (increasing). As $[1/(1-\rho) - p + 1] > 0$ the answer depends on the difference $\delta \equiv E[V^2] - 2E^2[V]$. For example, if V is deterministic, then $\delta = -E^2[V]$ and $E[L(n)]$ is increasing.

$L(n)$ is closely related to the nth cycle-time of an arbitrary fed back customer in an $M/G^f/1$ queue. Denoting this time by CT_n, $n \geqslant 0$ (CT_0 is the initial pass through the system of a newly arrived customer), it has been shown by Doshi and Kaufman [7] that for the $M/M^f/1$ queue the correlation coefficient between CT_0 and CT_1 is given by $Corr(CT_0, CT_1) = \rho$. Van den Berg and Boxma [3] extended this result and showed that $Corr(CT_n, CT_{n+1}) = \rho$ for *all* $n \geqslant 0$. The correlation between CT_0 and CT_1 in the *general* service-time $M/G^f/1$ queue may be obtained from the results in [7]. In the appendix we derive an *explicit* formula of $Corr[L(0), L(1)]$ in the *regular* $M/G/1$ queue, and show that if V is Exponential then $Corr[L(0), L(1)] = \rho$.

From (2.9) it readily follows that $E[L(n)] = E[L]$, for all $n \geqslant 1$, if and only if $E[L] = \rho/(1-\rho)$. This is true (see (2.1)) if and only if $E[V^2] = 2E^2[V]$. An *example* of such a case is, of course, the Exponential distribution. Yet, for the $M/M/1$ queue this result can be obtained directly:

$$E[L(1) \mid L] = \gamma(L+1)E[V] = \rho(L+1),$$

so that

$$E[L(1)] = \rho(E[L]+1) = \rho\left[\frac{\rho}{1-\rho}+1\right] = \frac{\rho}{1-\rho} = E[L].$$

Therefore, $E[L(n)] = E[L]$ for *all* $n \geqslant 1$.

When the service time is Exponential, $E[R] = E[V]$, so that eq. (2.12) reduces to $E[S] = \{E[V][E[L]+1]\}/p$. This can be given the following interpretation: each time C joins the end of the queue he observes (on the average) $E[L]$ customers *ahead* of him. Together with his own service time he resides in the system $E[V][E[L]+1]$ units of time until his service is completed (in this round). As the average number of times the server is visited by C is $1/p$, the total sojourn time of an arbitrary customer C is given by the product of these two terms.

3. N-node tandem Jackson network with feedback to the first node

We now turn back to our original model, as depicted in fig. 1, and assume that X_i is Exponentially distributed with parameter μ_i. It is well-known [8,10] that the *effective* arrival rates at the various nodes can be found by solving the set of traffic equations

$$\boldsymbol{\lambda} = [I - R^{\mathrm{T}}]^{-1}\boldsymbol{\gamma}, \tag{3.1}$$

where λ_i is the effective arrival rate to node i, R is the transition-probability matrix of movements from node to node, γ_i is the external arrival rate to node i, and I is the identity matrix.

In our case,

$$\boldsymbol{\gamma}=\begin{bmatrix}\gamma\\0\\0\\\vdots\\0\\\vdots\\0\\\vdots\\0\end{bmatrix},\quad\text{and}\quad R=\begin{bmatrix}1-p_1 & p_1 & 0 & \cdots & \cdots\\ 1-p_2 & 0 & p_2 & \cdots & \cdots\\ \vdots & & & & \\ 1-p_{N-1} & 0 & 0 & \cdots & p_{N-1}\\ 1-p_N & 0 & 0 & \cdots & 0\end{bmatrix}.$$

Solving (3.1) we get

$$\lambda_i=\gamma/\sum_{j=i}^{N}p_j. \tag{3.2}$$

Assuming $\lambda_i<\mu_i$ for $i=1,\ldots,N$, we can use Jackson's theorem [8] and write down the steady-state distribution of the network, i.e.,

$$P(L_1=n_1,\ L_2=n_2,\ldots,L_n=n_N)=\prod_{i=1}^{N}P(L_i=n_i), \tag{3.3}$$

where $P(L_i=n_i)=(1-\lambda_i/\mu_i)(\lambda_i/\mu_i)^{n_i}$ is the steady-state probability of having n_i customers in node i, as if it is an $M/M/1$ queueing system with arrival and service intensities λ_i and μ_i $(i=1,2,\ldots,N)$, respectively. It follows that the average number of customers at node i equals

$$EL_i=\lambda_i/[\mu_i-\lambda_i]=\gamma/\left[\left(\mu_i\prod_{j=i}^{N}p_j\right)-\gamma\right]\quad(i=1,2,\ldots,N), \tag{3.4}$$

and the average *total* number of customers in the system is

$$EL=\sum_{i=1}^{N}EL_i=\gamma\left[\sum_{i=1}^{N}\left\{1/\left[\left(\mu_i\prod_{j=i}^{N}p_j\right)-\gamma\right]\right\}\right]. \tag{3.5}$$

Clearly, for $N=1$, $EL=\gamma/(p_1\mu_1-\gamma)$.

The mean total sojourn time of an arbitrary customer C in the system is given by applying Little's rule $E[S]=EL/\gamma$ to (3.5), i.e.,

$$E[S]=\sum_{i=1}^{N}\left\{1/\left[\mu_i\sum_{j=i}^{N}p_j-\gamma\right]\right\}. \tag{3.6}$$

Using (3.4) this can be written as

$$E[S]=\sum_{i=1}^{N}\left[(EL_i+1)/\left(\mu_i\prod_{j=i}^{N}p_j\right)\right]. \tag{3.7}$$

Clearly, the average number of times C visits node i is $(\prod_{j=i}^{N} p_j)^{-1}$, and on the average the time he spends in node i at *each* visit is $(EL_i + 1)/\mu_i$. Hence, the mean total sojourn time of C in the system is given by the sum of these products. Observe also that the mean residence time of C in node i at *each* visit is given by EL_i/λ_i, whereas the *total* time he spends at i is

$$(EL_i + 1)/\left(\mu_i \prod_{j=i}^{N} p_j\right) = EL_i/\gamma.$$

Finally, if in Yechiali's model [13] we replace EX_i with EL_i/λ_i we obtain

$$E[S] = \sum_{i=1}^{N} (EL_i/\lambda_i)/\prod_{j=i}^{N} p_j. \tag{3.8}$$

4. Minimizing sojourn time

With the results obtained in the previous section our goal now is to find a rule for ordering the nodes so as to minimize the expected sojourn time of a customer in the system. We are looking for something similar to the simple index-rule ($EX_i/(1-p_i)$ increasing) derived in [13] in as much as the optimal placing of each node can be determined by its relationship to its immediate neighbours. Unfortunately, there is no general simple index-form rule that nullifies the need for a more involved combinatorial analysis. One simple case can be solved by applying a direct interchange argument to eq. (3.6). Considering the stations in order $\pi_0 = (1, 2, \ldots, k, k+1, \ldots, N)$ and comparing it to the order $\pi_1 = (1, 2, \ldots, k-1, k+1, k, k+2, \ldots, N)$, where stations k and $k+1$ are interchanged, it readily follows that if $\mu_k = \mu$ for all k, then the optimal ordering is determined by *increasing* values of p_k. Naturally, one would expect such a result since if service times are the same it is better to try first the stations with higher risk.

Another observation is that in the case where the time spent for service is of *no* consequence (i.e., only times spent with the servers are important) then the problem reduces quite simply to Yechiali's model.

The problem of minimizing $E[S]$, as given by eq. (3.6), turns out to be a special case of the following *general* sequencing problem: Given N pairs of real numbers (x_i, y_i), where $x_i \in [0, +\infty)$, and $y_i \in (-\infty, +\infty)$, find an ordering of the pairs so as to

$$\text{Minimize} \quad \left\{\sum_{i=1}^{N} \Psi(z_i)\right\}, \tag{4.1}$$

where Ψ is a monotone function and

$$z_i = \sum_{j=1}^{i} x_j + y_i. \tag{4.2}$$

Table 1
Sufficient conditions for optimality of π_0

	Non-decreasing Ψ	Non-increasing Ψ
Concave Ψ	$x_k \leqslant x_{k+1}$ (i) and $y_k \leqslant y_{k+1}$	$x_k \geqslant x_{k+1}$ (ii) and $y_k \leqslant y_{k+1}$
Linear Ψ	$x_k \leqslant x_{k+1}$ (iii) order of y's unimportant	$x_k \geqslant x_{k+1}$ (iv) order of y's unimportant
Convex Ψ	$x_k \leqslant x_{k+1}$ (v) and $y_k \geqslant y_{k+1}$	$x_k \geqslant x_{k+1}$ (vi) and $y_k \geqslant y_{k+1}$

Comparing the orderings π_0 and π_1 it follows that π_0 is better than π_1 if and only if

$$\Psi\left(\sum_{j=1}^{k} x_j + y_k\right) + \Psi\left(\sum_{j=1}^{k+1} x_j + y_{k+1}\right)$$
$$\leqslant \Psi\left(\sum_{j=1}^{k-1} x_j + x_{k+1} + y_{k+1}\right) + \Psi\left(\sum_{j=1}^{k+1} x_j + y_k\right). \tag{4.3}$$

We consider cases where Ψ is either concave or convex. This leads to a set of six possible characterizations of Ψ, and for each case the sufficient conditions for π_0 to be an optimal sequence are summarized in table 1.

To show that these are indeed the conditions we observe that for *all* six cases

$$\Psi\left(\sum_{j=1}^{k+1} x_j + y_{k+1}\right) - \Psi\left(\sum_{j=1}^{k+1} x_j + y_{k+1} - x_k\right)$$
$$\leqslant \Psi\left(\sum_{j=1}^{k+1} x_j + y_{k+1}\right) - \Psi\left(\sum_{j=1}^{k} x_j + y_{k+1}\right)$$
$$\leqslant \Psi\left(\sum_{j=1}^{k+1} x_j + y_k\right) - \Psi\left(\sum_{j=1}^{k} x_j + y_k\right),$$

which is equivalent to (4.3).

Before showing that minimizing ES, as given by (3.6), is equivalent to (4.1) we describe a few examples where one encounters the sequencing problem (4.1) and (4.2).

Example 1

Consider the problem where N items, having processing times x_i ($i = 1, 2, \ldots, N$), are to be processed sequentially on a single machine. Suppose that upon completion, item i requires an additional service of length y_i from an independent server which is always available. The objective is to minimize the sum of final completion times. Setting $\Psi(z_i)=z_i$ this is case (iii) whose solution is to arrange the jobs in increasing order of x_i, with no regard to the y_i values. That is, the optimal rule is *shortest processing time first*.

Example 2

Same problem as in example 1, but with discounting on the completing time, i.e., $\Psi(z_i)=[1-\exp(-\alpha z_i)]/\alpha$. Clearly, Ψ is concave and non-decreasing, which corresponds to case (i).

Example 3

Each of N jobs of length x_i has to be ready by time d_i, otherwise a linear charge $C_d \cdot \max(\Sigma_{j=1}^{i} x_j - d_i, 0)$ is levied for tardiness ($C_d > 0$). Setting $y_i = -d_i$, and $\Psi(z_i) = C_d \cdot \max(z_i, 0)$, we see that Ψ is a convex non-decreasing function (case (v)).

We now show that our original problem is equivalent to (4.1). Set $Q = \Pi_{j=1}^{N} p_j$, $x_j = -\log(p_j)$, $y_i = \log(\mu_i p_i)$, so that $z_i = \Sigma_{j=1}^{i} x_j + y_i$. Evidently, (3.6) becomes

$$E[S] = \sum_{i=1}^{N} \{1/[Q \exp(z_i) - \gamma]\} \equiv \sum_{i=1}^{N} \Psi(z_i). \tag{4.4}$$

Differentiating with respect to z_i we get $\Psi'(z_i) = -Q \exp(z_i)/[Q \exp(z_i) - \gamma]^2 < 0$, and $\Psi''(z_i) = Q \exp(z_i)[Q \exp(z_i) + \gamma]/[Q \exp(z_i) - \gamma]^3$. Since we assume that $\mu_i > \lambda_i$, i.e., $Q \exp(z_i) = \mu_i \Pi_{j=i}^{N} p_j > \gamma$, it follows that $\Psi''(z_i) > 0$. That is, Ψ is a convex monotone decreasing function (case (vi)). The conditions $x_k \geqslant x_{k+1}$ and $y_k \geqslant y_{k+1}$ are now translated into

$$p_k \leqslant p_{k+1} \quad \text{and} \quad \mu_k p_k \geqslant \mu_{k+1} p_{k+1}. \tag{4.5}$$

Result (4.5) establishes conditions for another special case: if $p_k = p$ for all k, then it is optimal to sequence the stations by *decreasing* values of μ_k. Intuitively, if feedback probabilities are equal, one would arrange the stations in increasing values of processing times (μ_k^{-1}), since the first stations are repeated most.

5. Extensions

Following [13] the results above can be extended to a case where feedback occurs either *internally* to node i itself, with probability q_i, say, or all the way to the *first* stage, with probability f_i, such that $f_i + q_i = 1 - p_i$ (see fig. 2).

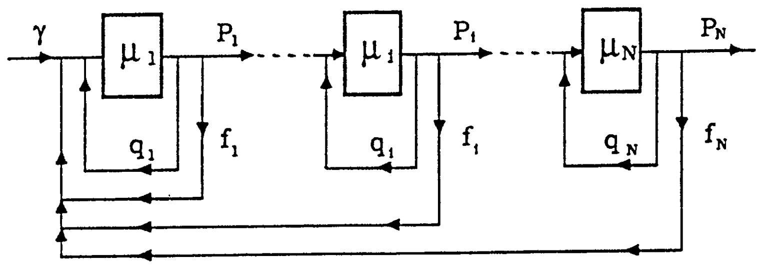

Fig. 2. Two-way feedback.

If each *internal* feedback is to the *head* of the queue then, as the total time that customer C spends in *service* at node i is Exponential with parameter $\hat{\mu}_i = \mu_i(1-q_i)$, we can set $\hat{p}_i = p_i/(p_i + f_i) = p_i/(1-q_i)$ and make the relevant substitutions in the previous model, so that *all* the results of sections 3 and 4 hold with $\hat{p}_i$ and $\hat{\mu}_i$ replacing p_i and μ_i, respectively. Note also that $\hat{p}_k\hat{\mu}_k = p_k\mu_k$. If upon repeating a node customer C rejoins the *end* of the queue, then the set of equations (3.1) takes the form

$$p_1\lambda_1 - \sum_{j=2}^{N} f_j\lambda_j = \gamma, \tag{5.1}$$

$$p_{i-1}\lambda_{i-1} - (1-q_i)\lambda_i = 0, \quad i = 2, 3, \ldots, N.$$

The solution of (5.1) is

$$\lambda_i = (\gamma/p_i) \prod_{k=i+1}^{N} [(1-q_k)/p_k]. \tag{5.2}$$

As before, $EL_i = \lambda_i/[\mu_i - \lambda_i]$, $EL = \Sigma_{i=1}^{N} EL_i$, $E[S] = EL/\gamma$. Setting now

$$Q = \prod_{j=1}^{N} [p_j/(1-q_j)], \quad x_j = -\log[p_j/(1-q_j)], \quad y_i = \log(\mu_i p_i),$$

eq. (4.4) holds in this case too, and all the results of the previous section apply here as well. In particular, the sufficient conditions for optimality, $x_k \geqslant x_{k+1}$ and $y_k \geqslant y_{k+1}$, are now translated, respectively, into

$$p_k/(1-q_k) \leqslant p_{k+1}/(1-q_{k+1}) \quad \text{and} \quad \mu_k p_k \geqslant \mu_{k+1} p_{k+1}.$$

Another simple extension to our first model is the addition of a "general

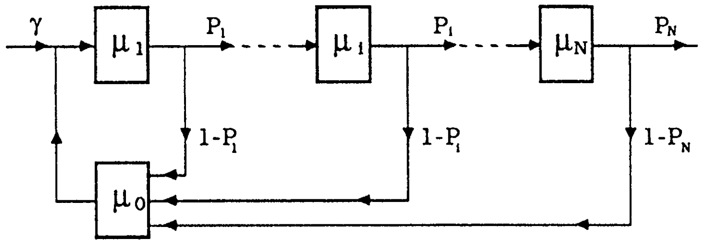

Fig. 3. A model with a "general feedback" node.

feedback" node, i.e., a node through which all fed back traffic must pass (node 0 in fig. 3). For example, in some production processes if the product is found defective in stage i, it is moved to a special "station" where it is dismantled so that its parts may be recycled.

It can be verified that (3.2) holds in this case for $1 \leqslant i \leqslant N$, and that

$$\lambda_0 = \lambda_1 - \gamma = \lambda_1 \left(1 - \prod_{i=1}^{N} p_i\right). \tag{5.3}$$

The expected total sojourn time is now obtained by adding eq. (5.4) below to eq. (3.6), where

$$EL_0/\gamma = \left[1 - \prod_{i=1}^{N} p_i\right] \Big/ \left[\left(\prod_{i=1}^{N} p_i\right)(\mu_0 + \gamma) - \gamma\right]. \tag{5.4}$$

Equation (5.4) is obtained either directly by setting $EL_0 = \lambda_0/(\mu_0 - \lambda_0)$, or by calculating $EL_0/\gamma = [(EL_0 + 1)/\mu_0][1/Q - 1]$, where $Q = \Pi_{i=1}^{N} p_i$. The last equation follows since $1/Q$ is the expected number of times customer C passes through channel N, and $[1/Q - 1]$ is the expected number of times he visits the general feedback node, where his mean sojourn time per visit is $[(EL_0 + 1)/\mu_0]$. It should be noted that (5.4) is independent of the *order* of the nodes (as the sojourn time in node i is independent of the order of nodes $i+1$ to N in the original model). Therefore, an optimal order in the *original* model will hold here too. A combination of the above two extensions (fig. 2 and fig. 3) can of course be made with our previous analysis quite simply applied.

Appendix

The correlation between the number of customers found in the system by an arbitrary arrival to an $M/G/1$ queue, and the number of customers he leaves behind him upon departure

Consider a customer C who joins a stationary *regular* $M/G/1$ queue with arrival rate γ and service times V. Let I denote the number of customers in front of him, and let J denote the number of customers that C leaves behind him upon service completion. We wish to calculate the correlation coefficient between these two random variables. If $I = i$, then C resides in the system a total time of $S = R_i + \Sigma_{j=1}^{i} V_j$, where R_i is the residual service time of the customer being served given that there are $I = i$ customers present ($R_0 \equiv V$), and $V_j \sim V$. Since all customers that are left upon departure have arrived during the time S, we readily have,

$$E[I \cdot J] = \sum_i \sum_j ij\pi_i \int_0^\infty \exp(-\gamma t) \frac{(\gamma t)^j}{j!} \mathrm{d}\left(\left(V^{*i} * R_i\right)(t)\right), \tag{A.1}$$

where $\{\pi_i\}_{i=0}^{\infty}$ is the stationary distribution of L, the number of customers in the system, and $(V^{*i} * R_i)(t)$ is the convolution between i service periods and the residual time of the customer in service. Now, $E[I \cdot J] = \gamma \sum_i i\pi_i \int_0^\infty t \, \mathrm{d}(V^{*i} * R_i)(t)$, and since R_i is independent of the V_j's, this reduces to

$$E[I \cdot J] = \gamma\left\{E[L^2]E[V] + \sum_{i=0}^{\infty} i\pi_i E[R_i]\right\}. \tag{A.2}$$

It was shown by Mandelbaum and Yechiali [11] that

$$E[R_i] = \frac{1-\rho}{\gamma\pi_i}\left[1 - \sum_{k=0}^{i} \pi_k\right],$$

where $\rho = \gamma EV$. Hence,

$$\begin{aligned} E[I \cdot J] &= \rho E[L^2] + (1-\rho)\sum_{i=0}^{\infty} i\left(\sum_{k=i+1}^{\infty} \pi_k\right) \\ &= \rho E[L^2] + (1-\rho)\sum_{k=2}^{\infty} \pi_k \frac{k(k-1)}{2} \\ &= \rho E(L^2) + \frac{1-\rho}{2} E[L(L-1)] = \tfrac{1}{2}\{(1+\rho)E[L^2] - (1-\rho)EL\}. \end{aligned} \tag{A.3}$$

As the marginal distributions of I and J are the same as that of L, the correlation between I and J now becomes

$$\begin{aligned} Corr[I, J] &= \frac{E[I \cdot J] - E^2[L]}{E[L^2] - E^2[L]} \\ &= \frac{\{(1+\rho)E[L^2] - (1-\rho)E[L]\}/2 - E^2[L]}{E[L^2] - E^2[L]}. \end{aligned} \tag{A.4}$$

It is well known [7] that for the regular $M/G/1$ queue the generating function of L is given by $\Pi(z) \equiv \sum_{j=0}^{\infty} \pi_j z^j = \tilde{S}[\gamma(1-z)]$, where $\tilde{S}[\cdot]$ is the LST of S, the sojourn time of a customer in the system. Therefore,

$$E[L(L-1)] = \gamma^2 E[S^2]. \tag{A.5}$$

It is also known (cf. Kella and Yechiali [9]) that

$$E[W_q^2] = \frac{\gamma E[V^3]}{3(1-\rho)} + \frac{\lambda E[V^2]}{1-\rho} E[W_q] \tag{A.6}$$

where W_q denotes the waiting time before service, and

$$E[W_q] = \lambda E[V^2]/(2(1-\rho)) \tag{A.7}$$

(cf. Kleinrock [10]). Using $S = W_q + V$ together with eqs. (A.5), (A.6) and (A.7), we arrive at

$$E[L^2] = \frac{\lambda^3 E(V^3)}{3(1-\rho)} + \left[\frac{\lambda^2 E[V^2]}{1-\rho} + 1\right] E[L] + \lambda^2 E[V^2], \tag{A.8}$$

where $E[L] = \lambda E[W_q] + \rho$. Substituting $E[L]$ and $E[L^2]$ in (A.4) gives an explicit formula for $Corr[I \cdot J]$.

For the $M/M/1$ queue, due to the memoryless properties of V, the conditional expectation of J is linear in I, i.e.

$$E[J \mid I = i] = \gamma(ER_i + iE[V]) = \rho(i+1).$$

Therefore,

$$Corr[I, J] = \rho \frac{\sigma(I)}{\sigma(J)} = \rho, \tag{A.9}$$

where $\sigma(X)$ denotes the standard deviation of a random variable X. Clearly, (A.9) could be calculated from (A.4) by substituting $E[L] = \rho/(1-\rho)$ and $E[L^2] = \rho(1+\rho)/(1-\rho)^2$.

References

[1] J.L. van den Berg and O.J. Boxma, Sojourn times in feedback queues, in: *Operations Research Proceedings 1988*, eds. D. Pressmar et al. (Springer, Berlin, 1989) pp. 247–257.

[2] J.L. van den Berg and O.J. Boxma, Sojourn times in feedback and processor sharing queues, in: *Proc. 12th Int. Teletraffic Congress 1988*, ed. M. Bonatti (North-Holland, Amsterdam, 1989) pp. 1467–1475.

[3] J.L. van den Berg and O.J. Boxma, The $M/G/1$ queue with processor sharing and its relation to a feedback queue, this issue.

[4] J.L. van den Berg, O.J. Boxma and W.P. Groenendijk, Sojourn times in the $M/G/1$ queue with deterministic feedback, Commun. Statist. – Stochastic Models 5 (1989) 115–129.

[5] O.J. Boxma and J.W. Cohen, The $M/G/1$ queue with permanent customers, IEEE J. Sel. Areas Commun. (1991).

[6] R.L. Disney, D. König and V. Schmidt, Stationary queue-length and waiting-time distributions in single-server feedback queues, Adv. Appl. Prob. 16 (1984) 437–446.

[7] B.T. Doshi and J.S. Kaufman, Sojourn time in an $M/G/1$ queue with Bernoulli feedback, in: *Queueing Theory and its Applications* (*Liber Amicorum for J.W. Cohen*), eds. O.J. Boxma and R. Syski (North-Holland, Amsterdam, 1988) pp. 207–233.

[8] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems* (Academic Press, New York, 1980).

[9] O. Kella and U. Yechiali, Priorities in $M/G/1$ queue with server vacations, Nav. Res. Log. 35 (1988) 23–34.

[10] L. Kleinrock, *Queueing Systems*, vol. 1 (Wiley, New York, 1975).

[11] A. Mandelbaum and U. Yechiali, The conditional residual service time in the $M/G/1$ queue, Technical Report, Dept. of Statistics, Tel-Aviv University (1979) unpublished.

[12] L. Takács, A single-server queue with feedback, Bell Sys. Tech. J. 42 (1963) 505–519.

[13] U. Yechiali, Sequencing an N-stage process with feedback, Prob. Eng. Infor. Sci. 2 (1988) 263–265.