The work described herein was carried out at INRIA, Rocquencourt, France. The hospitality and support I received there are gratefully acknowledged.
Queueing Systems with a Procrastinating Server.

Micha Hofri

Department of Computer Science, Technion
Haifa, 32000 Israel.

Abstract

Two related problems are analyzed and discussed:

a) A queueing system that differs from the standard \( M/G/1 \) only in that at the end of a busy-period the server takes a sequence of vacations, inspecting the state of the queue at the end of each. When the length of the queue exceeds a predetermined level \( m \) it returns to serve the queue exhaustively.

b) Two queues, with Poisson arrivals and general service-time distributions are attended by a single server. When the server is positioned at a certain queue it will serve the latter exhaustively, and at busy-period end will only switch to the other if the queue length there exceeds in size a predetermined threshold \( m_4 \).

The treatment combines analytic and numerical methods. Only steady-state results are presented.

---

The work described herein was carried out at INRIA, Rocquencourt, France. The hospitality and support I received there are gratefully acknowledged.
A. Introduction.

1. We consider here two queueing problems that display the effects of "threshold switching". The interest in these devolves from the likelihood of such operating regimes being optimal for an important class of control problems [Ross, 1985]. The analysis appears to be of interest in its own right, hence the following presentation.

2. The first problem concerns a queue subject to arrivals that follow a Poisson process with constant rate \( \lambda \). These arrivals present service requirements that follow the process \( \{S_i, i \geq 0\} \), with the \( S_i \) all independent and having the cumulative distribution function (cdf) \( F(\cdot) \), probability density function (pdf) \( f(\cdot) \) and Laplace-Stieltjes transform (LST) \( S(\cdot) \). A single server handles them at the order of arrival (this last restriction is immaterial for most of our analysis). In discussing the evolution of the system it will be found convenient to adopt a language that allows the server judgement and volition. At the end of a busy-period the server departs for a "vacation" with a random duration \( U \). When \( U \) terminates the length of the queue, \( X \), is inspected. If \( X \leq m \), for some predetermined integer \( m > 0 \), the server goes on an additional vacation, also distributed as \( U \), and so on. Once the queue-length at vacation end reaches (or exceeds) \( m \), the server resumes service until the queue is exhausted, whereupon the vacation process is renewed. We shall evaluate the steady-state distribution of queue-length at a random point in time and request sojourn time, where the main interest is in their dependence on \( m \).

Such systems with vacations have been described extensively in the literature, the earliest one being probably [Skinner, 1967] (but see also [Gelenbe and Iasnogorodski, 1980] and additional references there). However, as far as we know only the case \( m = 1 \) was considered. Recently an interesting observation was made in [Fuhrman, 1984] on the way the number of customers present in the system can be related to the number in a standard \( M/G/1 \) system. In [Doshi, 1985] the case \( m = 1 \) with vacations that do not all have the same distribution (but independent) is treated. The analysis of this queue has been given in [Hofri, 1985]. We repeat most of it here as leads naturally to the second one.

3. The second problem to be considered is made up of two queues, with queue \( i \), \( i = 1,2 \) having Poisson arrivals at rate \( \lambda_i \), service requirements per customer distributed as \( S_i \) and all possible independencies are assumed to hold. One may imagine that the relation between this problem and the earlier one is that here one queue may be considered as the reason why the server takes vacations from handling the other. This analogy is not very useful, though, because this view entails a dependence between the vacation duration and the busy-period it follows, a dependence that we do not allow above.

To switch from serving one queue to the other - say queue \( i \) - the server has to spend a time distributed as \( A_i \), and it will only start this switch when it is at an empty queue (idling) and the other has the queue length \( X_i \) satisfying \( X_i \geq m_i > 0 \). The value of \( X_i \) is assumed to be available without delay.
Here, as in the first problem we shall be satisfied with the distribution of $X_i$ at departure epochs from queue $i$, which is equal to the distribution at a random point in time. The joint queue length distribution is not hard to obtain from it.

4. Analyses of systems with several queues, attended by a single server, have been frequently reported in the literature. The main reference for the case considered here, where the server treats each queue exhaustively, is probably still [Eisenberg, 1972]. There, as in any other analysis of a multiqueue system we have seen, the itinerary of the server does not depend on the state of the queues. The (mainly numerical) difficulty of the analysis presented here devolves precisely from the relaxation of this assumption, as we allow the server a random number of busy-periods between switching.

The analysis we do is closer in spirit to the one presented in [Eisenberg, 1971] for the regime called there 'absolute priority', but is unfortunately exacerbated by the more complex switching process induced by the thresholds.

5. The motivation for this analysis came from studies of the optimal control policies for such a system, when the objective is to minimize the combination of $E(c_1X_1 + c_2X_2)$, the expectation being taken at a random point in time and an additional cost associated with the switching activity. The one queue system may be of interest when the server incurs a cost every time it starts (or ends) a busy-period; it was actually done as a preamble, to tone up for the main problem. Still, we present some new results, especially equations (7) and (8). We have reasons to believe that under certain restrictions, the regime as described above is indeed the optimal one [Ross, 1985]. Thanks to the linearity of this object-function we could avoid considering the joint process, and availed ourselves of the standard analysis of the M/G/1 system - as far as it would go - in the manner shown below.

B. One queue with vacations and a threshold.

1. We consider the system as defined in paragraph A.2. Denote by $J$ the number of customers in the queue when the service is resumed, and note that $J \geq m$ with probability 1. Proceeding exactly as in a class-room treatment of the standard M/G/1 system we observe the chain formed by the queue-length at points that correspond to service completion epochs. Since we later have to consider another embedded chain we shall grace the quantities that pertain to this one with the subscript $s$. The number of new arrivals, during a period distributed as $A$ will be denoted by $\tilde{A}$, and from the corresponding LST $A(\cdot)$ we obtain the pgf $\tilde{A}(z) = A(\lambda - \lambda z)$. For the steady-state probabilities we can now write
One server queues...

\[ \text{Prob (X at departure has length } x) = p_x(x) \]
\[ = \sum_{i=1}^{n+1} p_x(i) P(S = x - i + 1) + p_x(0) \sum_{i=1}^{n+1} P(I = i) P(S = x - i + 1) \quad x \geq 0 \]  

where the latter sum vanishes for \( x < m - 1 \). Hence

\[ g_x(z) = \sum_{x \geq 0} p_x(z)z^x = \frac{1}{z} S(z)(g_x(z) - p_x(0)) + p_x(0) \frac{S(z)}{z} I(z) \]

and we get the result

\[ g_x(z) = p_x(0) S(z) I(z) \frac{I(x)-1}{z-S(z)} \]  

Observe that by setting \( I(z) = z \) the familiar M/G/1 equation reappears. Substituting \( z = 1 \) we obtain \( p_x(0) \) as the reciprocal of the number of customers served in a busy-period:

\[ p_x(0) = \frac{1}{E[I]} \quad \rho = \lambda E(S) \]  

We get the nonsurprising stationarity condition \( \rho < 1 \).

2. It only remains to determine \( I(z) \). Possibly the simplest way is to consider the queue length embedded at a larger set of points, comprising both service and vacation terminations, but preserving the "type" of the points. Considering the probability that such a point chosen at random is a vacation-end and that \( z \) are then queued, we could write, with obvious notation

\[ p(x, U) = p(0, S) P(U=x) + \sum_{i=0}^{\text{min}(z, m-1)} p(i, U) P(U=x-i) \quad x \geq 0 \]  

and immediately

\[ g(z, U) = \sum_{x \geq 0} p(x, U)z^x \]
\[ = [p(0, S) + h(z)] \bar{U}(z), \quad h(z) = \sum_{i=0}^{m-1} p(i, U)z^i. \]

The component probabilities of \( h(z) \) can be determined numerically by restricting \( z \) to \( 0 \leq x < m \) in equation (4):

\[ p(x, U) = p(0, S) P(U=x) + \sum_{i=0}^{x} p(i, U) P(U=x-i) \quad 0 \leq x < m. \]  

The solution of (6) is immediate by successive substitutions, starting with \( p(0, U) \), up to a multiplicative factor - the \( p(0, S) \) that appears in the right-hand-side. Let \( h_{1}(z) = \sum_{i=0}^{m-1} p_{1}(i, U)z^i \) denote the solution obtained when the value of this factor is taken as 1. I.e. \( h(z) = p(0, S)h_{1}(z) \).

For the special case of exponentially distributed vacation durations, \( U \sim \exp(\beta) \), one
One server queues...

finds that the $p_{i}(z, U)$ are all equal (to $\eta/\lambda$), independently of $m^*$. Considering the relation between $g(z, U)$ and $I(z)$ (or rather, between the corresponding random variables) we get

$$I(z) = g(z, U) - h(z) = \tilde{g}(z) + h(z)[\tilde{g}(z) - 1]$$

where equation (5) was used to obtain the last equality.

Surprising as it may seem at first blush, we find that there is no need to evaluate the ratio between $p_{i}(z)$ and $p(z, S)$ in order to obtain $g_{i}(z)$. We do not have an adequate intuitive explanation for this convenient fact.

$$E(X) = (1-\rho)\tau'' + (\tilde{S}'' + 2\rho(1-\rho))\tau'''$$

Note: $\tilde{S}'' = \lambda^2E(S'')$

with all the derivatives being taken at $z = 1$. Higher moments are similarly available.

The LST of the distribution of the sojourn time $H$ of a customer in this system assuming FIFO order is given in terms of $g_{i}(z)$, just as for the standard $M/G/1$ system:

$$H(s) = g_{i}(1 - s/\lambda).$$

The above treatment is entirely straightforward to extend to the case when successive vacation durations are not identically distributed, in the same manner as done in [Doshi, 1985], for the case $m = 1$.

C. Two queues with conditional switching

1. Consider now the two-queue system as defined in paragraph A.3. In its discussion we shall mainly use the following notation:

$\lambda_{i}$ - arrival rate to queue $i$.

$S_{i}$ - service requirement at queue $i$.

$f_{i}(\cdot), F_{i}(\cdot)$ - density and distribution functions of $S_{i}$.

$X_{i}$ - queue length at position $i$ of the system (including the one in service when appropriate).

$B_{i}$ - duration of a busy period initiated by one customer, at queue $i$.

$G_{i}(z)$ - pgf of the queue length at queue $i$, at departure epochs from that queue.

$A_{i}$ - the switching time to queue $i$ (a random variable).

$I_{i}$ - the number of customers with which a busy-period starts at queue $i$.

Note that the $p_{i}(z, U)$ are not probabilities (only $p(z, U)$ are), thus one need not be alarmed when
One server queues...

\( U_i(\cdot) \) -
LST (pgf) of the continuous (discrete) random variable \( U_i \).

\( U_j \) -
the number of arrivals to queue \( j \) during a period distributed as \( U_i \).
Clearly then: \( U_j(z) = U_i(\lambda_j - \lambda_j z) \).

\( V_i \) -
terarrival time for queue \( i \). \( V_i(s) = \lambda_i / (\lambda_i + s) \).

\( H_i \) -
\( V_i + B_i \).

\( m_i \) -
critical (or threshold) value of \( X_i \) for the server to switch over from
queue \( i' \). \( m_i > 0 \).

2. First we remark that the treatment as given here does not cover the case of zero
thresholds, which imply that a switch is initiated whenever the server is at an empty
queue, whether at the end of a busy-period or having just completed the reverse
switch. When \( m_1 = m_2 = 0 \) we simply get a special case of the queueing system treated in
[Eisenberg, 1972]. An approach that would account for one zero threshold appeared to
us less natural and is relegated to a future work (in preparation). Secondly, we note
that equations (1-3) developed in Section B apply here without change, except that we
need to subscript the variables and the functions to identify the queue considered, and
observe that \( I_i \) may now be either \( A_{i+} \) at least \( m_1 \) or 1, when a busy period does or
does not follow a switch, respectively. Thus

\[
G_i(z) = G(0)S_i(z)x_i(z) \frac{I_i(z) - 1}{z - S_i(z)},
\]

(10)

\[
G_i(0) = \frac{1 - \rho_i}{E(I_i)}. \rho_i = \lambda_i E(S_i).
\]

(11)

3. Computing \( I_i(z) \) however is here enormously more complicated because of the more
involved structure of the "vacation" process that couples the two queue-length
processes. In analogy with the procedure developed in Section B we consider the
queue-length processes embedded in points that are of two possible "types:
-end of a busy-period (type \( b \) point)
-beginning of switching by an idling server (type \( c \) point).

A type \( c \) point occurs following the event that a server finishes a busy-period (itself a
type \( b \) point) and observes that the queue at the other position is below the switching
threshold, and therefore idles. However, arrivals to that queue let it reach the
threshold before there is an arrival to the queue where the idling server is stationed,
and it starts to switch over. A state of this chain will be denoted by \((i, x, b \text{ or } c)\) with
\( i \in (1, 2) \) the position of the server, and \( x \) - the length of the other queue (\( X_i = 0 \) then).
Note that for a type \( c \) point only \( x = m_i \) is possible, with \( i' \) being the alternate of \( i \). The
transition probabilities for this chain may be read off the following equations, which we
write for clarity for a specific value of \( i = 1 \):

\[ A_1(1) > 1. \]

This chain is "minimal" in the sense that for low enough arrival rates it has no transient states."
One server queues...

\[ p(1,z,b) = \sum_{i=0}^{\infty} p(1,i,b)P(H_{12}=z-i) + p(2,m_1,c)P(A_{12}+B_{12}^{m_1+A_{11}} = z) \]

\[ + \sum_{j=m_1}^{\infty} p(2,j,b)P(A_{12}+B_{12}^{j+A_{11}} = z) \quad 0 \leq z < m_2 \]

where a superscript \( \ast r \) denotes \( r \)-fold convolution, and

\[ p(1,z,b) = \sum_{i_1=0}^{m_1-1} p(1,i_1,b) \sum_{i_2=0}^{m_2-i_1-1} P(V_{12}=i_2)P(B_{12} = z-i_1-i_2) \]

\[ + p(2,m_1,c)P(A_{12}+B_{12}^{m_1+A_{11}} = z) + \sum_{j=m_1}^{\infty} p(2,j,b)P(A_{12}+B_{12}^{j+A_{11}} = z) \quad x \geq m_2 \]

and lastly

\[ p(1,x,c) = \delta(x-m_2) \sum_{j=0}^{m_1-1} p(1,j,b)P(V_{21}^{m_1-j} = 0) \]  

Note that the \( A_{11} \) and \( A_{12} \) in equation (13) are dependent through the same realization of \( A_1 \). The corresponding equations and transition probabilities for \( i = 2 \) are obtained from the above by completely symmetrizing over the indices 1 and 2.

4. Define

\[ g(1,z,b) = \sum_{x=0}^{\infty} p(1,x,b)z^x \]  

and observe that

\[ P(V_{12} = k) = \frac{\lambda_1 \lambda_2^k}{(\lambda_1 + \lambda_2)^{k+1}}, \quad P(V_{21}^m = 0) = \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^m \]  

Substituting (12)-(14) and (16) into (15) and rearranging we obtain

\[ g(1,z,b) = A_1(\lambda_1(1-B_{12}(z)) + \lambda_2(1-z)) \cdot [p(2,m_1,c)B_{12}^{m_1}(z)] \]

\[ + g(2,B_{12}(z),b) = \sum_{j=0}^{m_1-1} p(2,j,b)B_{12}(z) \]

\[ + \frac{\lambda_1 B_{12}(z)}{\lambda_1 + \lambda_2 - \lambda_2 z} \sum_{i=0}^{m_1-1} p(1,i,b)[1-\left( \frac{\lambda_2 z}{\lambda_1 + \lambda_2} \right)^{m_1-i}] \]  

and the dual equation:

\[ g(2,z,b) = A_2(\lambda_2(1-B_{21}(z)) + \lambda_1(1-z)) \cdot [p(1,m_2,c)B_{21}^{m_2}(z)] \]

\[ + g(1,B_{21}(z),b) = \sum_{j=0}^{m_2-1} p(1,j,b)B_{21}(z) \]

\[ + \frac{\lambda_2 B_{21}(z)}{\lambda_2 + \lambda_1 - \lambda_1 z} \sum_{i=0}^{m_2-1} p(2,i,b)[1-\left( \frac{\lambda_1 z}{\lambda_1 + \lambda_2} \right)^{m_2-i}] \]  

As in Section B, denote by \( h_i(z) \) the low-order part of \( g(i,z,b) \). Specifically

\[ h_i(z) = \sum_{z=0}^{m_i-1} p(i,z,b)z^x. \]

Note that substituting \( z = 1 \) in (17.1) (or (17.2)) yields

\[ p(1,m_2,c) + g(1,1,b) - h_1(1) = p(2,m_1,c) + g(2,1,b) - h_2(1) \]  

which merely states that under steady-state conditions, the frequency of switching from queue 1 to queue 2 is equal to the frequency of switching in the reverse direction.
This obvious observation will provide us with an equation for the "boundary" probabilities \( p(i,x,b), x < \alpha_2 \). Another equation is supplied by the natural normalization

\[
p(1,m_2,c) + p(2,m_1,c) + g(1,1,b) + g(2,1,b) = 1
\]  

(19)

5. We turn now to the solution of equations (17). To keep the size of the expressions manageable we shall introduce a raft of functions, as we go along. We start by abbreviating \( g(i,z,b) \) to \( g_i(z) \), \( p(i,z,b) \) to \( p_i(z) \) and defining

\[
a_i(z) = A_i(\lambda_1(1-B_i(z)) + \lambda_2(1-z)).
\]

Substituting (14) into (17:1) we obtain:

\[
g_1(z) = a_1(z)g_2(B_12(z)) + \sum_{i=0}^{m_2-1} p_1(i)\beta_1,i(z) + \sum_{i=0}^{m_1-1} p_2(i)\gamma_1,i(z)
\]  

(20.1)

with

\[
\beta_{1,i}(z) = \frac{\lambda_1 z^i B_{12}(z)}{\lambda_1 + \lambda_2 - \lambda_2 z} \left[1 - \left(\frac{\lambda_2 z}{\lambda_1 + \lambda_2}\right)^{m_2-i}\right],
\]

\[
\gamma_{1,i} = a_1(z)\left[\frac{\lambda_1}{\lambda_1 + \lambda_2}\right]^{m_1-i} - B_{12}(z).\]

A dual equation is likewise obtained from (16:2)

\[
g_2(z) = a_2(z)g_1(B_12(z)) + \sum_{i=0}^{m_1-1} p_1(i)\beta_{2,i}(z) + \sum_{i=0}^{m_2-1} p_2(i)\gamma_{2,i}(z)
\]  

(20.2)

with the \( \beta_{2,i}(z) \) and \( \gamma_{2,i}(z) \) being obtained from their correspondents in equation (21) by the exchange 1 <-> 2.

Using (20.2) to evaluate \( g_2(B_{12}(z)) \) we get from (20.1) an equation involving \( g_1(\cdot) \) only, with \( m_1 + m_2 \) "boundary" state probabilities:

\[
g_1(z) = \delta_1(z)g_1(\eta_1(z)) + \sum_{i=0}^{m_2-1} p_1(i)\phi_{1,i}(z) + \sum_{i=0}^{m_1-1} p_2(i)\xi_{1,i}(z)
\]  

(22)

with

\[
\delta_1(z) = a_1(z)a_2(B_{12}(z))
\]

\[
\eta_1(z) = B_{21}(B_{12}(z))
\]

\[
\phi_{1,i}(z) = \beta_{1,i}(z) + a_1(z)\gamma_{2,i}(B_{12}(z))
\]

\[
\xi_{1,i}(z) = \gamma_{1,i}(z) + a_1(z)\beta_{2,i}(B_{12}(z))
\]

Clearly a dual equation for \( g_2(\cdot) \) is produced when one symmetrizes over 1 and 2 in equations (22-23).

6. So far we have proceeded with blithe disregard to the question of stability or stationarity of the two queue system, expecting - as is often the case - that ergodicity will produce equations that inter alia display, or at least determine the conditions under which stationarity will in fact prevail.
We have not been disappointed. Considering equation (22) one may ask: under what conditions does this equation define a function $g_1(z)$ which is analytic for $|z|<1$ and continuous for $|z|\leq 1$, as a proper pgf should be? Since when $g_1(z)$ is analytic we may use (22) for recursive substitution, one may ask instead what conditions are required to guarantee that such a procedure, continued ad infinitum, shall yield an analytic result. To this question we have an answer. Let $g^j(z)$ denote the $j$th functional iterate of $g_1(z)$, i.e.

$$g^0(z) = z$$
$$g^j(z) = g_1(g^{j-1}(z))$$

(24)

Carrying out the recursive substitution procedure we formally obtain

$$g_1(z) = \lim_{j \to \infty} g_1(g^j(z)) \prod_{k=0}^{m-1} \delta_k(y^k(z))$$
$$+ \sum_{i=0}^{m-1} p_{1,i} (\sum_{j=0}^{m-1} g^{j-1}(z) \prod_{k=0}^{m-1} \delta_k(y^k(z)))$$
$$+ \sum_{i=0}^{m-1} p_{2,i} (\sum_{j=0}^{m-1} g^{j-1}(z) \prod_{k=0}^{m-1} \delta_k(y^k(z)))$$

(25)

Now the conditions can be stated explicitly: Equation (22) is viable iff the following hold:

a) $g^j(z)$, for $|z|<1$ and $j \to \infty$ converges to a point where $g_1(z)$ is defined. Call this point $a$.

Then we also need:

b) $\delta_1(a) < 1$;

c) $\delta_{1,i}(a)$, $\xi_{1,i}(a) < \infty$, and if $\delta_1(a) = 1$, they must both vanish there.

Finally, the limits must be approached fast enough to assure convergence to a continuous function for $|z|\leq 1$. The last requirement will turn out to be the key one.

In the Appendix we show that these requirements indeed hold when $\rho = \lambda_1 E(S_1) + \lambda_2 E(S_2) < 1$. Then $a=1$ and the convergence is geometrically fast.

This condition has a very simple intuitive meaning. The definition in equation (23) of $\eta_1(z)$ can be read as

$$\eta_1(z) = E(z^{\text{arr}_1})$$

this translates to the verbal statement that $\eta_1(z)$ is the pgf of the number of arrivals to queue 1, during a sequence of busy-periods of queue 2 with multiplicity equal to a random variable with the distribution of the number of arrivals to queue 2 during a (simple) busy-period of queue 1. Clearly the expected value derived from this pgf must be less than one to assure stability. Note that the switching times play no role in the stability condition, which is natural for the case of exhaustive service. We may therefore now state

Theorem: When $\rho = \lambda_1 E(S_1) + \lambda_2 E(S_2) < 1$, the chain embedded at the $b-c$ points defined in paragraph C.2 reaches a steady state with the pgf
The fraction of queue $i$ busy-periods that begin when an idling server receives an arrival and then $I_i = 1$ is $\frac{1}{g_i(1)}[z(h_i - p(i, m_i, c)) + z^{m_i + A_u}p(i', m_i, c) + \sum_{j=m_i}^{m_i + A_u} p_t(j)]$. The dual $g_2(z)$ is obtained by symmetrizing over the indices 1 and 2.

7. To complete the computation two tasks remain: relating $g_i(\cdot)$ to $I_i(\cdot)$ as required in equation (10), and the evaluation of the $2 + m_1 + m_2$ unknowns appearing in (26) and its dual. The first task is simple, the other is not. We shall tackle them in order.

The fraction of queue $i$ busy-periods that begin when an idling server receives an arrival and then $I_i = 1$ is $\left(\frac{h_i(1) - p(i, m_i, c)}{g_i(1)}\right)$; the fraction that begin after the server completed a switch it started when idling and then $I_i = m_i + A_u$ is $\frac{p(i', m_i, c)}{g_i(1)}$, and finally there are the busy-periods that follow a switch started immediately following a busy-period of queue $i$; the value of $I_i$ is then $j + A_u$ and their frequency $p_t(j)/g_i(1)$ for $j \geq m_i$. Thus, for a given value of $A_u$

$$I_i(z) = \frac{1}{g_i(1)}[z(h_i - p(i, m_i, c)) + z^{m_i + A_u}p(i', m_i, c) + \sum_{j=m_i}^{m_i + A_u} p_t(j)]$$

Removing the conditioning on $A_u$ yields

$$I_i(z) = \frac{1}{g_i(1)}[z(h_i - p(i, m_i, c)) + A_u(z)\{p(i', m_i, c)z^{m_i} + g_i(z) - h_i(z)\}]$$

Note that the $g_i(z)$ allow us to evaluate the mean queue lengths and waiting time distributions precisely as in the one queue case; i.e. via the suitably subscripted equations (8) and (9).

8. The determination of $g_i(1)$ and $p_t(j)$, $0 \leq j < m_i$, is in principle quite straightforward: equations (18) and (19), equation (26) (respectively, its dual), its first $m_2$ derivatives (respectively, $m_1$), all the derivatives evaluated at $z = 0$, provide the required number of linear equations in those unknowns. The difficulty resides in differentiating anything as unattractive as the right-hand-side of equation (26) several times. Still, the search for alternative representations having failed (it would have helped, for example, if we could replace equations (20) by equivalent, and solvable, integral equations) this was carried out. Some aspects of this task, as well as numerical examples are in the next Section. We remark that devising the procedure to evaluate these boundary probabilities formed the larger part of the entire effort.

9. As was mentioned in A.5, a purpose of the analysis was to evaluate this policy under a cost structure where the system incurs constant-rate holding charges for any customer in the queues or being serviced, as well as a one-time charge for every switch. The first component averages out to $c_1E(X_1) + c_2E(X_2)$ which is given by the analogue of equation (9). To adapt the second component to this scale of average cost per time unit we need the mean frequency of switching. This can be approached in a number of ways, the simplest being probably as follows:
Consider the chain defined over the \( b \) and \( c \) points in C.2. Let \( \pi \) be the probability that the state at a transition point of the chain, selected at random, leads to an immediate switch to queue \( i \) (every switch starts at a transition point of this chain). Such a switch occurs if the selected point found the chain in the state \((i', m_i, c)\) or \((i', x, b)\) with \( x \geq m_i \). Otherwise one or more transitions will be realized before such a switch takes place. Thus

\[
\pi_i = p(i', m_i, c) + g_i(1) - h_i(1)
\]

To convert this to frequency we need the mean time between transition points of the chain, \( L \). This is straightforward to compute by counting the possible transitions, the only one presenting some difficulty being the time to the next transition following a \( b \)-point where the server idles. One obtains

\[
L = \sum_{i=1}^{\infty} \left( (g_i(1) - h_i(1))E(A_i)[1 + \lambda_i E(B_i)] + g_i(1) - h_i(1) \right)
+ \sum_{j=0}^{m_c-1} p_i(j) \left[ \frac{\alpha_i^{m_c-j}E(A_i)}{\lambda_i} + \frac{\alpha_i}{\lambda_i} \right] E(B_i)
+ \frac{1}{m_c-j} E(B_i)(1 - \alpha_i^{m_c-j}) + \frac{1}{\lambda_i} \left[ 1 - (m_c-j+1)\alpha_i^{m_c-j} + (m_c-j)\alpha_i^{m_c-j+1} \right]
\]  

The required frequency is clearly given by \( \pi_i / L \), for which all the necessary quantities are given parameters or have been computed.

D. Numerical procedures, examples and discussion.

1. As mentioned in C.8 the method we used to evaluate \( g_i(1) \) and \( p_i(j) \) was to combine equations (18) and (19) with derivatives of equation (26) and its dual at \( z = 0 \). Two obstacles had to be overcome, beyond the sheer complexity of the expressions:

   a) The expressions contain infinite sums and products. As shown in the appendix, these converge exponentially. In practice we iterated \( \eta_i(z) \), starting at \( z = 0 \) until it approached \( z = 1 \) to within \( z = 10^{-8} \), and used the order of the iteration thus determined as the upper limit to all the (infinite) sums and products. It was found that changing \( \log_{10} \) in the range -7 to -13 had no effect on the first six significant decimal digits in all the numbers of interest.

   b) The basic function, \( \eta(z) \), is available in terms of the LSTs of the busy-period length distributions. These are determined by the equations

\[
B_i(u) = S_i(u + \lambda_i (1-B_i(u)))
\]  

In the general case there is no closed-form solution to equation (30). A formal solution is given by
where a bracketed superscript \([j]\) denotes \(j\)th order differentiation. From the numerical point of view our experience with equation (31) was disappointing. Although the series converges nicely in principle, the higher order terms - at least when \(S(u)\) is a rational or an exponential function - are made of very large and extremely small factors, and overflow or underflow usually intervened before satisfactory convergence was achieved. This can probably be overcome by careful programming, specific for each type of \(S(u)\), but it turned out that the following iterative scheme provided a stable and as prompt convergence. Shedding the subscript \(i\) we write:

\[
B_0(u) = S(u + \lambda) \\
B_{j+1}(u) = S(u + \lambda(1 - B_j(u)))
\]  

(32)

2. To obtain derivatives of \(B(u)\) and the other functions that abound here one uses the Faa-di-Bruno formula:

\[
f(g(z))^{[r]} = \sum_{j=1}^{r} f^{[j]}(g(z)) \sum_{E} F_{E,r} \prod_{i=1}^{r} [g^{[i]}(z)]^{k_i}
\]  

(33)

with the coefficients \(F\) given by

\[
F_{E,r} = \frac{r!}{\prod_{i=1}^{r} k_i! (i!)^{k_i}}
\]

and the sum over \(E\) ranges on integer-valued \(r\)-dimensional vectors that satisfy

\[
\sum_{i=1}^{r} k_i = j, \quad \sum_{i=1}^{r} \Delta k_i = r, \quad k_i \geq 0.
\]

Applying (33) to equation (30) iteratively, and noting that \(B^{[r]}(u)\) appears just once in the right-hand-side (with coefficient 1), the derivatives of \(B(u)\) to any order were easy to compute. Equation (33) was used extensively throughout the computation. The entire set of the required vectors \(E\) and coefficients \(F\) was prepared in a first phase, to order of \(\max(m_1, m_2) - 1\).

3. For the special case of exponentially distributed service times \(B(u)\) does have a closed-form representation

\[
B(u) = \frac{(\mu + \lambda + u - \sqrt{(\mu + \lambda + u)^2 - 4\lambda \mu})}{2\lambda}
\]

(34)

and in this case the iteration (32) is not required. Indeed, most of the numerical experiments were conducted for this case, as there is no apparent reason for the qualitative behavior of the system to depend in any essential way on the shape of the distribution of the service times.

All the computations proved to be numerically well-behaved. The computational times were highly variable, occasionally substantial, with most of the time expended in computing the high order derivatives of the functions that make up equation (26), through iterations of (33). The computation times were also quite sensitive to \(\rho\), once
it exceeded 0.8 or 0.9, as the convergence of the sums and products is then much slower (see the appendix). An example: for $\rho = 0.9$, with all the parameters fixed except that $m_4$ were varied in the ranges $1 \leq m_1, m_2 \leq 8$, $1 \leq m_3 \leq 3$ the computation required 20 seconds on a DEC-785, most of it in the preprocessing phase, involving some 100 arrays. The results were compared against a separate computation for $m_4 = 1$, which requires no differentiation, and with simulation for higher-valued thresholds.

4. Examples.

The "parameter space" that describes this problem is quite large: $\lambda_i$, $m_i$, and four distributions for the $A_i$, $S_i$ with their parameters. The same embarras de richesse holds for the computable quantities. The following were chosen then rather arbitrarily. $\lambda_i$ and $\mu_q$ were varied as indicated in the first four table captions. $S_i$ were chosen exponential, $\mu_q$ was fixed at 1.0, $A_i$ were picked as constant and their values both fixed at 5.0. In table 5 all were held constant except $E[A_i]$, as the sum $E[A_1 + A_2]$ was partitioned between the two senses of the switch in various ways. The columns "swfr" give the values of the switching frequency computed from equation (29).

<table>
<thead>
<tr>
<th>$m_1$</th>
<th>$m_2=1$</th>
<th>$m_2=5$</th>
<th>$m_2=9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E[X_1]$</td>
<td>$E[X_2]$</td>
<td>$E[X_1]$</td>
</tr>
<tr>
<td>1</td>
<td>0.0305</td>
<td>0.2444</td>
<td>0.2851</td>
</tr>
<tr>
<td>2</td>
<td>0.2660</td>
<td>0.7032</td>
<td>0.2486</td>
</tr>
<tr>
<td>3</td>
<td>0.2490</td>
<td>1.0644</td>
<td>0.2089</td>
</tr>
<tr>
<td>4</td>
<td>0.2370</td>
<td>1.5637</td>
<td>0.2040</td>
</tr>
<tr>
<td>5</td>
<td>0.0227</td>
<td>2.0410</td>
<td>0.1929</td>
</tr>
<tr>
<td>6</td>
<td>0.0219</td>
<td>2.5326</td>
<td>0.1854</td>
</tr>
<tr>
<td>7</td>
<td>0.0211</td>
<td>3.0310</td>
<td>0.1802</td>
</tr>
<tr>
<td>8</td>
<td>0.0205</td>
<td>3.5342</td>
<td>0.1770</td>
</tr>
<tr>
<td>9</td>
<td>0.0198</td>
<td>4.0406</td>
<td>0.1749</td>
</tr>
<tr>
<td>10</td>
<td>0.0193</td>
<td>4.5491</td>
<td>0.1736</td>
</tr>
<tr>
<td>11</td>
<td>0.0198</td>
<td>5.0590</td>
<td>0.1731</td>
</tr>
<tr>
<td>12</td>
<td>0.0193</td>
<td>5.5697</td>
<td>0.1731</td>
</tr>
</tbody>
</table>

Table I: $\lambda_1 = 0.1$, $\lambda_2 = 0.05$, $\mu_2 = 1.00$

<table>
<thead>
<tr>
<th>$m_1$</th>
<th>$m_2=1$</th>
<th>$m_2=5$</th>
<th>$m_2=9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E[X_1]$</td>
<td>$E[X_2]$</td>
<td>$E[X_1]$</td>
</tr>
<tr>
<td>1</td>
<td>0.0509</td>
<td>2.3500</td>
<td>1.4052</td>
</tr>
<tr>
<td>2</td>
<td>0.0477</td>
<td>2.3765</td>
<td>1.3765</td>
</tr>
<tr>
<td>3</td>
<td>0.0438</td>
<td>2.5838</td>
<td>1.3450</td>
</tr>
<tr>
<td>4</td>
<td>0.0402</td>
<td>2.9160</td>
<td>1.3289</td>
</tr>
<tr>
<td>5</td>
<td>0.0371</td>
<td>3.3279</td>
<td>1.3125</td>
</tr>
<tr>
<td>6</td>
<td>0.0348</td>
<td>3.7934</td>
<td>1.3130</td>
</tr>
<tr>
<td>7</td>
<td>0.0325</td>
<td>4.2613</td>
<td>1.3216</td>
</tr>
<tr>
<td>8</td>
<td>0.0306</td>
<td>4.7505</td>
<td>1.3382</td>
</tr>
<tr>
<td>9</td>
<td>0.0290</td>
<td>5.2454</td>
<td>1.3552</td>
</tr>
<tr>
<td>10</td>
<td>0.0275</td>
<td>5.7432</td>
<td>1.3775</td>
</tr>
<tr>
<td>11</td>
<td>0.0262</td>
<td>6.2423</td>
<td>1.4023</td>
</tr>
<tr>
<td>12</td>
<td>0.0251</td>
<td>6.7421</td>
<td>1.4290</td>
</tr>
</tbody>
</table>

Table II: $\lambda_1 = 0.3$, $\lambda_2 = 0.15$, $\mu_2 = 1.00$
5. The same analysis and computational procedure would suffice for a system of any number of queues served exhaustively by a single server, but with one limitation: the itinerary of the server may not depend on the state of the queues, only the instants when it does switch. Specifically, for an itinerary \((i, i', i'', \cdots)\), if, when queue \(i\) is exhausted \(X_i = m_i\) and \(X_i > m_i\), the server should idle and not switch. Giving up this...
constraint leads to a process which is considerably more complicated to analyze, even though it is of the same general structure. With this constraint the analysis presented above is adequate, but the numerical effort is considerably larger. Thus, for three queues we would have
\[ \delta_1(z) = a_1(z)a_2(B_{12}(z))a_3(B_{23}(B_{12}(z))). \]
and the analogue of equation (26) will be correspondingly heavier.

6. It is an interesting challenge, to which we have not found an answer, to devise an alternative representation for the information carried by the chain introduced in C.3, that will provide easier access to the distribution of \( I_4(z) \), particularly for more than two queues; or at least for the first few moments. We note that computing higher moments of the queue lengths than we did here, from the \( G_4(z) \) of equation (11), requires a marginal amount of additional work with the set-up described here.
Appendix

We wish to prove the conditions specified in C.6 for the existence of the steady-state of the chain on the $b$ and $c$ points. Consider first the behavior of $\eta^{[1]}(z)$. (We need not treat separately the dual function).

When $\rho_i = \lambda_i E(S_i) < 1$ the busy-period in queue $i$ has a proper limiting distribution with the LST $B_i(\cdot)$, and $B_\infty$ is well defined too. Thus $\eta_1(z)$ is an analytic function for $|z| < 1$ (at least) and continuous for $|z| \leq 1$. Observe that being a pgf it is completely monotonic - all its derivatives are positive for $0 \leq z \leq 1$. When $\eta_1(z)$ is iterated we get a sequence of analytic functions, which could converge to any solution of the equation $z = \eta_1(z)$. This equation need not have a unique solution. Indeed, if $\eta_1'(1) > 1$ there must be a real solution in $(0,1)$, since $\eta_1(0) > 0$, $\eta_1(1) = 1$, in addition to $z=1$ which is a solution as well.

Observe however that if $\eta^{[1]}(z) \rightarrow a < 1$, then also $\delta_1(\eta^{[1]}(z))$, being a pgf, would converge to a number smaller than 1. Since for $z=1$ the "convergence",is to the value 1, equation (26) would define a function which is discontinuous at $z=1$. We must reject this possibility, and disallow $\eta_1'(1) > 1$, and demand that $z=1$ be the unique fixed point for the iteration of $\eta^{[1]}(z)$.

This would be satisfied if $\eta_1'(1) \leq 1$, but in order to have the sums and products in equation (26) converge we must impose a stronger condition: we have, developing around $z=1$

$$\eta^{[1]}(z) = \eta^{[1]}(1) + (z-1)\eta^{[1]}(1) + O(z-1)^2$$

or

$$|\xi_j| = |\eta^{[1]}(z)-1| \leq |z-1| |\eta^{[1]}(1)|$$

so that $\eta_1'(1) < 1$ is the required condition to assure convergence. However, $\eta_1'(1) = \rho_1 \rho_2 / (1-\rho_1)(1-\rho_2)$, hence the condition is:

$$\rho = \rho_1 + \rho_2 < 1.$$  

This will also guarantee exponential convergence for the coefficient of $g_1(1)$: Let $P_k(z) = \prod_{j=0}^{k-1} \delta_1(\eta_j^{[1]}(z))$, then

$$\frac{P_{j+1}(z)}{P_j(z)} = \delta_1(\eta_j^{[1]}(z))$$

$$= \delta_1(1-\xi_j(z)) = 1 - \delta_1(1) \xi_j(z) + O(\xi_j(z))$$

which approaches the limit geometrically fast. The same computation holds for the sums in equation (26).
References


M. Hofri: An M/G/1 Queue with Vacations and a Threshold. Department of Computer Science, The Technion Haifa 32000 Israel. TR#775, August 1985.
