PERTURBATION ANALYSIS OF THE M/M/1 QUEUE IN A MARKOVIAN ENVIRONMENT VIA THE MATRIX-GEOMETRIC METHOD

Cheng-Shang Chang and **Randolph Nelson**

IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Abstract

In this paper, we consider a family of $M(t)/M/1$ queues in which customers arrive according to nonhomogenous Poisson processes with intensity $\lambda_t(\epsilon) = \lambda_{\epsilon t}$, $0 < \epsilon < \infty$. We assume that $\lambda_t(\epsilon)$ is an irreducible finite-state Markov process. Based on the matrix-geometric method, we use perturbation analysis to obtain the second order approximations for the expected queue length for two cases where ϵ is small and where ϵ is large. Using these approximations, we show that the expected waiting times are *strictly* decreasing in ϵ when ϵ is small. In the case where ϵ is large, we show that the expected waiting times are *strictly* decreasing in ϵ if the intensity process is dynamically reversible. These results partially answer a question posed by Rolski.

Keywords: time dependent queues, perturbation analysis, matrix-geometric method, Ross's conjecture

1 INTRODUCTION

In this paper we address a question initially posed by Ross in 1978 [7] concerning comparison of performance for a family of $M(t)/GI/1$ queues. Consider a queueing system in which customers arrive according to a nonhomogenous Poisson process with intensity function $\lambda_t(\epsilon) = \lambda_{\epsilon t}$. Ross conjectured that the expected waiting time of the queue is lower bounded by that of the $M/GI/1$ queue with the same rate and that the expected waiting times of the queue are decreasing in ϵ if $\lambda_t(\epsilon)$ is a two-state Markov process. Both conjectures were shown to be true by Rolski [9] and Chang, Chao and Pinedo [1], respectively. Various other related work on these conjectures were reported in the literature and reviewed by Rolski [10] in a recent survey paper. In Rolski's survey paper (Conjecture (ii)), he posed the question when the expected waiting times of the queue are (strictly) decreasing in ϵ. One of our objectives in this paper is to answer the question in the case that $\lambda_t(\epsilon)$ is a finite-state Markov process and the service times are exponentially distributed.

To do this we consider a family of $M/M/1$ queues, parameterized by ϵ, in a random environment (see Neuts [5], Chapter 6). The random environment is described by a m-state, irreducible Markov process with generator ϵQ, where $0 < \epsilon < \infty$ and $Q = \{q(i,j), 1 \leq i,j \leq m\}$, is the matrix of transition rates of a m-state, irreducible Markov process. During sojourns in environment state j, $1 \leq j \leq m$, customers arrive to the $M/M/1$ queue according to a Poisson process with rate λ_j and are served with an exponential rate of μ_j. It is well known that such a queueing system forms a Quasi-Birth-Death (QBD) process and the steady state probabilities can be written in matrix-geometric form [5]. The rate matrix of such a solution, denoted by R, requires that one determine the solution of a quadratic equation in R. Numerical procedures are typically used for this solution. In this paper, instead of solving for the R-matrix numerically, we use perturbation analysis to obtain closed-form expressions for the R-matrix for small and large ϵ. Based on approximations of the R-matrix, we derive second-order approximations for the expected queue length. Using these approximations, we show that the expected queue length is *strictly* decreasing in ϵ for small ϵ under the assumptions (i-a) μ_i's are identical, (ii-a) not all λ_i's are identical and (iii-a) $\max(\lambda_i/\mu_i) < 1$. In the case where ϵ is large, we show that the queue lengths are *strictly* decreasing in ϵ if (i-b) μ_i's are identical, (ii-b) not all λ_i's are identical, (iii-b) $\sum_i \pi_i\lambda_i / \sum_i \pi_i\mu_i < 1$ with π_i being the stationary probability in state i and (iv-b) Q is dynamically reversible. The first assumptions in both cases indicate that the service times are independent of the states. The second assumptions are to avoid the trivial cases with constant Poisson arrival rates. The third assumptions are the stability conditions. We require a stronger stability condition in (iii-a) to allow for the case $\epsilon \to 0$. Here the holding times for states in the random environment increase without bound and we require that $\max(\lambda_i/\mu_i) < 1$ to insure that the expected queue length of the system is bounded. Note that π_i is independent of ϵ and also that condition (iii-a) implies condition (iii-b). Since the expected arrival rates are the same for all ϵ, we can use Little's formula [4] to obtain results for the expected waiting times. Thus, these results partially answer Rolski's question for the special case of exponential service times.

2 THE MODEL

We define the state $(i,j), i \geq 0, j = 1,2\ldots,m$, where i is the total number of customers in the $M/M/1$ queue and j is the state of the random environment. We will call the value of i (resp. j) the *level* (resp. *interlevel*) of state (i,j). The generator of this process, denoted by $\tilde{Q}(\epsilon)$, can be written as

$$\tilde{Q}(\epsilon) = \begin{bmatrix} \epsilon Q - \Delta(\boldsymbol{\lambda}) & \Delta(\boldsymbol{\lambda}) & 0 & 0 & \cdots \\ \Delta(\boldsymbol{\mu}) & \epsilon Q - \Delta(\boldsymbol{\lambda}+\boldsymbol{\mu}) & \Delta(\boldsymbol{\lambda}) & 0 & \cdots \\ 0 & \Delta(\boldsymbol{\mu}) & \epsilon Q - \Delta(\boldsymbol{\lambda}+\boldsymbol{\mu}) & \Delta(\boldsymbol{\lambda}) & \cdots \\ 0 & 0 & \Delta(\boldsymbol{\mu}) & \epsilon Q - \Delta(\boldsymbol{\lambda}+\boldsymbol{\mu}) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

where $\Delta(\mathbf{a})$ denotes, for $\mathbf{a} \equiv (a_1, a_2, \ldots, a_m)$, the diagonal matrix $\mathrm{diag}(a_1, a_2, \ldots, a_m)$ and Q is a generator of an irreducible Markov process. We define $\Delta_1(\mathbf{a}) \equiv I - \Delta(\mathbf{a})$.

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ be the stationary probability vector of Q. Then it follows from Neuts [5] that $\tilde{Q}(\epsilon)$ is ergodic if

$$\boldsymbol{\pi}\boldsymbol{\lambda}^t < \boldsymbol{\pi}\boldsymbol{\mu}^t. \tag{1}$$

The stationary probability vector, $\boldsymbol{x}(\epsilon) = [x_0(\epsilon), x_1(\epsilon), \ldots]$ of $\tilde{Q}(\epsilon)$ is given by

$$x_k(\epsilon) = \boldsymbol{\pi}(I - R(\epsilon))R^k(\epsilon), \quad \text{for } k \geq 0, \tag{2}$$

where the matrix $R(\epsilon)$ is the minimal solution of

$$R^2(\epsilon)\Delta(\boldsymbol{\mu}) + R(\epsilon)[\epsilon Q - \Delta(\boldsymbol{\lambda}+\boldsymbol{\mu})] + \Delta(\boldsymbol{\lambda}) = 0. \tag{3}$$

Using (2), the expected total number of customers in the queue is given by

$$EL(\epsilon) = \boldsymbol{\pi}(I - R(\epsilon))^{-1}R(\epsilon)\mathbf{e}^t, \tag{4}$$

where $\mathbf{e}$ throughout the paper denotes an appropriately dimensioned row vector of 1's.

In general, the matrix $R(\epsilon)$ cannot be solved explicitly. Here we will use perturbation analysis to obtain a closed-form second-order approximation for $R(\epsilon)$ and the expected queue length $EL(\epsilon)$ for both cases where ϵ is small and where ϵ is large.

We record here a factorization of (3) that will be used later in the paper. Rewriting (3) we have that

$$R^2(\epsilon)\Delta(\boldsymbol{\mu}) + R(\epsilon)[\epsilon Q - \Delta(\boldsymbol{\lambda}+\boldsymbol{\mu})] + \Delta(\boldsymbol{\lambda}) = (R(\epsilon) - I)(R(\epsilon)\Delta(\boldsymbol{\mu}) - \Delta(\boldsymbol{\lambda})) + \epsilon R(\epsilon)Q = 0. \tag{5}$$

Since we assume that $\mu_i > 0$ for all i, $\Delta^{-1}(\boldsymbol{\mu})$ exists. It then follows from (5) that

$$\begin{aligned} R(\epsilon) &= \Delta(\boldsymbol{\lambda})\Delta(\boldsymbol{\mu})^{-1} + \epsilon(I - R(\epsilon))^{-1}R(\epsilon)Q\Delta^{-1}(\boldsymbol{\mu}) && (6) \\ &= \Delta(\boldsymbol{\rho}) + \epsilon R_2, && (7) \end{aligned}$$

where

$$R_2 = (I - R(\epsilon))^{-1}R(\epsilon)Q\Delta^{-1}(\boldsymbol{\mu}). \tag{8}$$

3 PERTURBATION ANALYSIS FOR SMALL ϵ

In this section, we derive the second-order approximation for the matrix $R(\epsilon)$ and the expected queue length when ϵ is small. The expected sojourn time in environment state j, $1 \leq j \leq m$, is given by $\{-\epsilon q(j,j)\}^{-1}$ and thus increases without bound as $\epsilon \to 0$. For sufficiently small ϵ this implies that the statistics for the queue while the environment is in state j approach the stationary distribution of a $M/M/1$ queue with arrival and service rates of λ_j and μ_j, respectively. Thus, we define the process for $\epsilon = 0$ to be a stationary reducible Markov process. With probability π_j, it is a stationary $M/M/1$ queue with arrival and service rates of λ_j and μ_j. Let $R_0 = \Delta(\boldsymbol{\rho})$ and $L_0 = \Delta_1^{-1}(\boldsymbol{\rho})\Delta(\boldsymbol{\rho})$, where $\rho_i = \lambda_i/\mu_i$. Observe that the matrix R_0 is a solution of the R matrix in (3) when $\epsilon = 0$ and that L_0 is a diagonal matrix of expected queue lengths of the reducible Markov process at $\epsilon = 0$. Under the definition of the process for $\epsilon = 0$, we have $EL(0) = \boldsymbol{\pi} L_0 \mathbf{e}^t$. Our purpose in the next theorem is to characterize $R(\epsilon)$ and $EL(\epsilon)$ for small $\epsilon > 0$.

Theorem 3.1 *If* $\max_i(\lambda_i/\mu_i) < 1$ *and* $\mu_i > 0$ *for all* i*, then for small* $\epsilon > 0$ *the second-order approximation of* $R(\epsilon)$ *is*

$$R(\epsilon) = R_0 + \epsilon L_0 Q \Delta^{-1}(\boldsymbol{\mu}) + O(\epsilon^2). \tag{9}$$

The expected queue length is

$$\begin{aligned} EL(\epsilon) &= \boldsymbol{\pi} L_0 \mathbf{e}^t \\ &+ \epsilon \boldsymbol{\pi} \Delta_1^{-1}(\boldsymbol{\rho}) L_0 Q \Delta^{-1}(\boldsymbol{\mu})[I + L_0]\mathbf{e}^t + O(\epsilon^2). \end{aligned} \tag{10}$$

We first prove a technical lemma that will be used in the proof of Theorem 3.1.

Lemma 3.2 *If* $\max_i(\lambda_i/\mu_i) < 1$ *and* $\mu_i > 0$ *for all* i*, then there exists a finite constant* c *such that* $\mathbf{e}(I - R(\epsilon))^{-1}R(\epsilon)\mathbf{e}^t \leq c$ *for all* $\epsilon > 0$*. Thus the matrix* $(I - R(\epsilon))^{-1}R(\epsilon)$ *is* $O(1)$*.*

Proof. Let $\rho_{max} = \max_i(\lambda_i/\mu_i)$ which, by assumption, satisfies $\rho_{max} < 1$. Now consider an auxiliary system that is identical to the original system except the arrival rate at state i is $\rho_{max}\mu_i \geq \lambda_i$, $i = 1, \ldots, m$. Since the arrival rate at each state in the auxiliary system is not less than that in the original system it follows from a standard sample path argument ([8, 11]) that the queue length process in the auxiliary system is not stochastically less than that of the original system. This implies that the expected queue length in the auxiliary system is not less than that in the original system. Since the traffic intensities at all states are identical in the auxiliary system, the expected queue length in the auxiliary system is simply $(1 - \rho_{max})^{-1}\rho_{max}$ (see Neuts [5], Remarks f., pp. 259-260). In conjunction with (4), we have

$$\boldsymbol{\pi}(I - R(\epsilon))^{-1}R(\epsilon)\mathbf{e}^t \leq (1 - \rho_{max})^{-1}\rho_{max}. \tag{11}$$

Let $\pi_{min} = \min_i(\pi_i)$ which, from the irreducibility of Q, implies that $\pi_{min} > 0$. From (11) and the fact that $R(\epsilon)$ is a nonnegative matrix, it then follows that for all ϵ

$$\mathbf{e}(I - R(\epsilon))^{-1}R(\epsilon)\mathbf{e}^t \leq ((1 - \rho_{max})\pi_{min})^{-1}\rho_{max} < \infty. \tag{12}$$

This implies that $(I - R(\epsilon))^{-1}R(\epsilon)$ is $O(1)$. ■

Proof of Theorem 3.1. To simplify the notation, we suppress the dependency of the rate matrix on ϵ and denote $R(\epsilon)$ by R. As a direct consequence of Lemma 3.2, R_2 defined by (8) is $O(1)$. Also from the assumption that $\max(\lambda_i/\mu_i) < 1$, $\Delta_1^{-1}(\boldsymbol{\rho})$ exists. Using (7) we can write

$$\begin{aligned} (I - R)^{-1} &= (\Delta_1(\boldsymbol{\rho}) - \epsilon R_2)^{-1} = [\Delta_1(\boldsymbol{\rho})(I - \epsilon\Delta_1^{-1}(\boldsymbol{\rho})R_2)]^{-1} & (13)\\ &= (I - \epsilon\Delta_1^{-1}(\boldsymbol{\rho})R_2)^{-1}\Delta_1^{-1}(\boldsymbol{\rho}). & (14)\end{aligned}$$

For sufficiently small ϵ, we have

$$(I - \epsilon\Delta_1^{-1}(\boldsymbol{\rho})R_2)^{-1} = [I + \epsilon\Delta_1^{-1}(\boldsymbol{\rho})R_2 + O(\epsilon^2)] \tag{15}$$

and thus, using (14), we can write

$$(I - R)^{-1}R = L_0 + \epsilon\Delta_1^{-1}(\boldsymbol{\rho})R_2[I + L_0] + O(\epsilon^2). \tag{16}$$

Inserting (16) into (6) yields (9).

It also follows from (16) that $R_2 = L_0Q\Delta^{-1}(\boldsymbol{\mu}) + O(\epsilon)$. Replacing R_2 in (16) by $L_0Q\Delta^{-1}(\boldsymbol{\mu})$ and using (4) yields (10). ■

We can exploit the probabilistic interpretation of the rate matrix (Neuts [5], page 35. remark b.) to interpret equation (9). To do this let $\nu_k \equiv \lambda_k + \mu_k - \epsilon q(k,k), 1 \le k \le m$, and note that for small ϵ, $\nu_k \approx \lambda_k + \mu_k$. We write an off diagonal element of R in (9) as

$$\begin{aligned} R(j,k) &= \frac{\rho_j}{1-\rho_j}\frac{\epsilon q(j,k)}{\mu_k} & (17)\\ &\approx \frac{\rho_j}{1-\rho_j}\;\frac{\epsilon q(j,k)}{\nu_j}\;\frac{\nu_j}{\nu_k}\,(1+\rho_k), \quad i \ne j. & (18)\end{aligned}$$

The interpretation for the rate matrix is simplified if we consider the process embedded at transition epochs. The rate matrix for this discrete time version of the process is denoted by R^0, and using equation (1.7.20) in Neuts [5] is given by

$$\begin{aligned} R^0(j,k) &= \frac{\nu_k}{\nu_j}R(j,k) & (19)\\ &\approx \frac{\rho_j}{1-\rho_j}\;\frac{\epsilon q(j,k)}{\nu_j}\,(1+\rho_k). & (20)\end{aligned}$$

The value of $R^0(j,k)$ is the number of visits to state $(i+1,k)$ before first return to a state below level $i+1$ given that the process is started in state (i,j). For ϵ very small, there is at most one interlevel transition before passing below level $i+1$. Given that an interlevel transition occurs from interlevel j to interlevel k (denoted here by $j \to k$), the expected number of visits to state $(i+1,k)$ is given by $(1+\rho_k)$. This can be explained as follows: the $j \to k$ transition must occur at a level higher than $i+1$ and since the process is a birth/death process with transition rates λ_k and μ_k it must visit state $(i+1,k)$ at least once before crossing below level $i+1$. The expected number of subsequent visits after the first is given by ρ_k and thus combining yields $(1+\rho_k)$. The remaining terms in (20) are the probability of making the interlevel transition $j \to k$. To see this note that the probability that any given transition is a $j \to k$ transition is given by $\epsilon q(j,k)/\nu_j$. The value of $\rho_j/(1-\rho_j)$ is the expected number of transitions on interlevel j before crossing below level $i+1$. To first order, then the probability of making a $j \to k$ transition is obtained by multiplying these two factors.

We also note that the extension of our perturbation analysis to the $M/PH/1$ queue in a Markovian environment is difficult. This is due to the fact that there does not exist a simple factorization in (5). Moreover, the inverse matrix in (6) may not exist.

In the following theorem, we show a structural result for the expected queue length when ϵ is small.

Theorem 3.3 *If (i) all μ_i's are identical, i.e., $\mu_i = \mu$ for all i and (ii) not all λ_i's are identical, then under the assumptions in Theorem 3.1 the expected queue length is strictly decreasing in ϵ for small ϵ.*

We use the following lemma in the proof of Theorem 3.3.

Lemma 3.4 *Let M be an irreducible generator matrix of order m with zero row and column sums and let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be vectors arranged in decreasing order, i.e., $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_m$ and $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_m$. Then $\boldsymbol{\alpha} M \boldsymbol{\beta}^t \leq 0$. If, furthermore, $\alpha_i = f_1(z_i)$ and $\beta_i = f_2(z_i)$ for a m-vector $\mathbf{z}$ and two strictly increasing functions f_1 and f_2, then $\boldsymbol{\alpha} M \boldsymbol{\beta}^t = 0$ implies $z_1 = z_2 = \ldots = z_m$.*

Proof. Let $\mathbf{e}_j$ be a row vector with the first j elements equal to 1 and the other elements equal to 0. We first note that since M has zero row and column sums that

$$\mathbf{e}_i M \mathbf{e}_j^t \leq 0, \quad 1 \leq i,j \leq m. \tag{21}$$

Furthermore we note that irreducibility of M implies that

$$\mathbf{e}_i M \mathbf{e}_i^t < 0, \quad 1 \leq i < m. \tag{22}$$

We now write $\boldsymbol{\alpha} = \sum_{j=1}^m a_j \mathbf{e}_j$ with $a_j \equiv \alpha_j - \alpha_{j+1}$ $(\alpha_{m+1} = 0)$ and similarly, $\boldsymbol{\beta} = \sum_{j=1}^m b_j \mathbf{e}_j$, with $b_j \equiv \beta_j - \beta_{j+1}$ $(\beta_{m+1} = 0)$. Note that by assumption $a_j, b_j \geq 0,\ 1 \leq j \leq m$. Using this we write

$$\boldsymbol{\alpha} M \boldsymbol{\beta}^t = \sum_{i=1}^m \sum_{j=1}^m a_i b_j \mathbf{e}_i M \mathbf{e}_j^t \leq 0. \tag{23}$$

The inequality in (23) follows from (21) and the fact that $a_i b_j \geq 0$. Now consider the case that $\alpha_i = f_1(z_i)$ and $\beta_i = f_2(z_i)$ for a m-vector $\mathbf{z} = (z_1, \ldots, z_m)$ and two strictly increasing functions f_1 and f_2. Suppose all the z values are not equal and thus that $z_j > z_{j+1}$ for some $1 \leq j < m$. Since f_1 and f_2 are strictly increasing, $\alpha_j > \alpha_{j+1}$ and $\beta_j > \beta_{j+1}$. Thus, from (22), (23) and the fact that $a_j b_j > 0$ we have that

$$\boldsymbol{\alpha} M \boldsymbol{\beta} \leq a_j b_j \mathbf{e}_j M \mathbf{e}_j^t < 0. \tag{24}$$

■

Proof of Theorem 3.3. It suffices to show that the second term in (10) is less than 0 if all μ_i's are identical and not all λ_i's are identical. Without loss of generality, we may assume that $\mu_i = \mu = 1$ and thus $\Delta(\boldsymbol{\mu}) = I$. Also we may rearrange the states such that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_m$. Since Q is a generator of a Markov process, $Q\mathbf{e}^t = 0$ and therefore $\boldsymbol{\pi}\Delta_1^{-1}(\boldsymbol{\rho})L_0Q\mathbf{e}^t = 0$. To establish the result we thus only need to show that

$$\boldsymbol{\pi}\Delta_1^{-1}(\boldsymbol{\rho})L_0QL_0\mathbf{e}^t < 0. \tag{25}$$

We rewrite (25) as $\boldsymbol{\alpha}M\boldsymbol{\beta}^t$ where $M \equiv \Delta(\boldsymbol{\pi})Q$ and

$$\boldsymbol{\alpha} \equiv (\frac{\rho_1}{(1-\rho_1)^2}, \ldots, \frac{\rho_m}{(1-\rho_m)^2}) \tag{26}$$

and

$$\boldsymbol{\beta} \equiv (\frac{\rho_1}{1-\rho_1}, \ldots, \frac{\rho_m}{1-\rho_m}). \tag{27}$$

Note that M, by definition, is an irreducible generator matrix with zero row and column sums The result then follows from Lemma 3.4.

■

4 PERTURBATION ANALYSIS FOR LARGE ϵ

In this section, we derive the second-order approximation for the matrix $R(\epsilon)$ and the expected queue length when ϵ is large.

Theorem 4.1 *If (i) all the μ_i's are identical and equal to μ and (ii) $\bar{\rho} = \boldsymbol{\pi}\boldsymbol{\lambda}^t/\boldsymbol{\pi}\boldsymbol{\mu}^t < 1$ then for large ϵ the second-order approximation of $R(\epsilon)$ is*

$$R(\epsilon) = R_3 + \epsilon^{-1}(I - R_3)R_4 + O(\epsilon^{-2}), \tag{28}$$

where

$$R_3 = \boldsymbol{\rho}^t\boldsymbol{\pi} = \begin{bmatrix} \rho_1\pi_1 & \rho_1\pi_2 & \cdots & \rho_1\pi_m \\ \rho_2\pi_1 & \rho_2\pi_2 & \cdots & \rho_2\pi_m \\ \vdots & \vdots & \vdots & \vdots \\ \rho_m\pi_1 & \rho_m\pi_2 & \cdots & \rho_m\pi_m \end{bmatrix}, \tag{29}$$

and R_4 is the unique solution of the following equations

$$R_4Q = -\Delta(\boldsymbol{\lambda}) + R_3\Delta(\boldsymbol{\mu}), \tag{30}$$

$$R_4\mathbf{e}^t = 0. \tag{31}$$

Under the same assumptions, the second-order approximation for the expected queue length is

$$EL(\epsilon) = \frac{\bar{\rho}}{1-\bar{\rho}} + \epsilon^{-1}\frac{1}{1-\bar{\rho}}\boldsymbol{\pi}R_4\boldsymbol{\rho}^t + O(\epsilon^{-2}). \tag{32}$$

For large ϵ, one would expect that the arrival process approaches to a Poisson process with rate $\boldsymbol{\pi}\boldsymbol{\lambda}^t$ and that the queue length approaches to an $M/M/1$ queue. Note that the first term in (32) is the expected queue length of an $M/M/1$ queue with arrival rate $\boldsymbol{\pi}\boldsymbol{\lambda}^t$.

Before we prove Theorem 4.1, we discuss some properties of the matrix R_3. We first note that it is $O(1)$. Moreover, the spectral radius of R_3 is $\bar{\rho}$ since the rank of R_3 is one. From the assumption that $\bar{\rho} < 1$, it follows that the matrix $I - R_3$ is invertible. Note that

$$R_3^2 = \boldsymbol{\rho}^t\boldsymbol{\pi}\boldsymbol{\rho}^t\boldsymbol{\pi} = \bar{\rho}\boldsymbol{\rho}^t\boldsymbol{\pi} = \bar{\rho}R_3. \tag{33}$$

Thus,

$$\begin{aligned}(I - R_3)^{-1} &= I + R_3 + R_3^2 + R_3^3 + \ldots \\ &= I + R_3 + \bar{\rho}R_3 + \bar{\rho}^2 R_3 + \ldots \\ &= I + \frac{1}{1-\bar{\rho}}R_3. \end{aligned}\tag{34}$$

Also, we have that

$$R_3 Q = \boldsymbol{\rho}^t \pi Q = 0 \tag{35}$$

and that

$$R_3\mathbf{e}^t = \boldsymbol{\rho}^t\boldsymbol{\pi}\mathbf{e}^t = \boldsymbol{\rho}^t. \tag{36}$$

Proof. The proof is simplified if we make the following definitions. Let $R_5 = (I - R_3)R_4$, $R_6 = R_7 - \epsilon^{-1}R_5$ and $R_7 = R(\epsilon) - R_3$. Since we assume that all μ_i's are equal, it follows from (7) and $Q\mathbf{e}^t = 0$ that

$$R(\epsilon)\mathbf{e}^t = \boldsymbol{\rho}^t \leq \rho_{max}\mathbf{e}^t, \tag{37}$$

where $\rho_{max} = \max_i(\lambda_i/\mu)$. This implies that all the elements in $R(\epsilon)$ are less than ρ_{max} for all $\epsilon > 0$. Thus, $R(\epsilon)$ is $O(1)$. Since R_3 is $O(1)$, it follows that matrix R_7 is at most $O(1)$. We will strengthen this conclusion after preliminary algebra and show that matrix R_7 is $O(\epsilon^{-1})$. Observe here that we can equivalently express R_7 as $R_7 = \epsilon^{-1}R_5 + R_6$.

To show that (28) holds, it suffices to show that R_6 is $O(\epsilon^{-2})$. Replacing $R(\epsilon)$ by $R_3 + R_7$ in (5) and using (35) and (30) yields

$$R_6 Q = -\epsilon^{-1}\Big(R_7(R_3\Delta(\boldsymbol{\mu}) - \Delta(\boldsymbol{\lambda})) + (R_3 - I)R_7\Delta(\boldsymbol{\mu}) + R_7^2\Delta(\boldsymbol{\mu})\Big). \tag{38}$$

From (37), (36) and (31), we also have that

$$R_6\mathbf{e}^t = 0. \tag{39}$$

Since Q is an irreducible generator, the matrix $I + \nu^{-1}Q$ is an irreducible stochastic matrix, where $\nu = \max_i(|q(i,i)|)$. Analogous to the argument in Neuts [6], pp. 125, the matrix $\nu^{-1}Q - \mathbf{e}^t\boldsymbol{\pi}$ is invertible (see Horn [2], Page 498, Lemma 8.2.7(g) for a proof). Thus,

$$R_6 = -\epsilon^{-1}\nu^{-1}\Big(R_7(R_3\Delta(\boldsymbol{\mu}) - \Delta(\boldsymbol{\lambda})) + (R_3 - I)R_7\Delta(\boldsymbol{\mu}) + R_7^2\Delta(\boldsymbol{\mu})\Big)(\nu^{-1}Q - \mathbf{e}^t\boldsymbol{\pi})^{-1}. \tag{40}$$

Equation (40) is an expression for R_6 in terms of itself (expressed through matrix R_7). A series of arguments now establishes that R_6 is $O(\epsilon^{-2})$. Recall that matrices R_7 and R_3 have already been shown to be $O(1)$. It thus follows from (40) that R_6 is at most $O(\epsilon^{-1})$. This, however, implies that R_7 is $O(\epsilon^{-1})$ since, as noted above, it is equivalently expressed as $R_7 = \epsilon^{-1}R_5 + R_6$. From this and equation (40) we infer that R_6 is $O(\epsilon^{-2})$. We have thus established (28).

To establish (32) we proceed analogously to the derivation of (16). Following that line of reasoning it is easy to show (with $R(\epsilon) = R_3 + \epsilon^{-1}R_5 + O(\epsilon^{-2})$) that for sufficiently large ϵ

$$(I - R(\epsilon))^{-1}R(\epsilon) = (I - R_3)^{-1}R_3 + \epsilon^{-1}[(I - R_3)^{-1}R_5 + (I - R_3)^{-1}R_5(I - R_3)^{-1}R_3] + O(\epsilon^{-2}). \quad (41)$$

Inserting $R_5 = (I - R_3)R_4$ in (41) and using (4) yields

$$EL(\epsilon) = \boldsymbol{\pi}(I - R_3)^{-1}R_3\mathbf{e}^t + \epsilon^{-1}\boldsymbol{\pi}R_4\mathbf{e}^t + \epsilon^{-1}\boldsymbol{\pi}R_4(I - R_3)^{-1}R_3\mathbf{e}^t + O(\epsilon^{-2}). \quad (42)$$

Note that $\boldsymbol{\pi}R_4\mathbf{e}^t = 0$ from (31). Moreover, it follows from (34), (33) and (36) that

$$\boldsymbol{\pi}(I - R_3)^{-1}R_3\mathbf{e}^t = \frac{\bar{\rho}}{1 - \bar{\rho}}. \quad (43)$$

Similarly,

$$\boldsymbol{\pi}R_4(I - R_3)^{-1}R_3\mathbf{e}^t = \frac{1}{1 - \bar{\rho}}\boldsymbol{\pi}R_4\boldsymbol{\rho}^t. \quad (44)$$

■

We can also give a probabilistic interpretation of R_3 as follows. As in the previous section, let $\nu_k \equiv \lambda_k + \mu_k - \epsilon q(k,k), 1 \le k \le m$, and note that for large ϵ, $\nu_k \approx -\epsilon q(k,k)$. Similarly, we denote the rate matrix for the discrete time version of the process by R^0, and use equation (1.7.20) in Neuts [5]. Thus,

$$\begin{aligned} R^0(j,k) &= \frac{\nu_k}{\nu_j}R(j,k) && (45)\\ &\approx \frac{-\epsilon q(k,k)}{\nu_j}R_3(j,k) && (46)\\ &= \frac{\lambda_j}{\nu_j}\frac{1}{\mu}\pi_k[-\epsilon q(k,k)]. && (47) \end{aligned}$$

The value of $R^0(j,k)$ is the expected number of visits to state $(i+1,k)$ before the first visit to level i given that the process starts in state (i,j). Observe that $\frac{\lambda_j}{\nu_j}$ is the probability of that the process goes up one level at the first step. To count the expected number of visits made to state $(i+1,k)$ after this initial transition before returning to level i we make a simple observation. Since ϵ is large, for any level $j, j \ge i+1$, the fraction of time the process spends in interlevel state k given that it is in level j is given by π_k. This observation implies that transitions between levels can be characterized by a birth-death process where we coalesce transitions according to their interlevel probabilities. Specifically, increasing transitions $(j \to j+1)$ between levels occur at rate $\lambda = \boldsymbol{\pi}\boldsymbol{\lambda}^t$ and decreasing transitions $(j+1 \to j)$

occur at rate μ. This implies that the number of visits to level $i+1$ before first visiting level i is geometrically distributed with parameter $\mu/(\lambda+\mu)$ and a simple calculation shows that the expected time spent in level $i+1$ is given by $1/\mu$. Thus, π_k/μ is the expected time that the process stays in state $(i+1,k)$ before first visiting level i. Since $\frac{1}{-\epsilon q(k,k)}$ is the expected time spent in state $(i+1,k)$ before it leaves this state, $(\pi_k/\mu)[-\epsilon q(k,k)]$ is the expected number of visits to state $(i+1,k)$ before visiting level i starting from the state $(i+1,j)$. Multiplying this by the probability of initially going up to state $(i+1,j)$, previously stated, completes the argument.

Analogous to Theorem 3.3, we show a structural result for the expected queue length when ϵ is large.

Theorem 4.2 *If (i) not all λ_i's are identical and (ii) Q is a generator of a dynamically reversible and irreducible Markov process, then under the assumptions in Theorem 4.1 the expected queue length is strictly decreasing in ϵ for large ϵ.*

We need the following lemma in the proof of Theorem 4.2.

Lemma 4.3 *If Q is a generator of a dynamically reversible and irreducible Markov process, then $\mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1}\mathbf{z}^t \leq 0$ for any m-vector $\mathbf{z}$. Moreover, $\mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1}\mathbf{z}^t = 0$ if and only if $\mathbf{z}$ is proportional to $\boldsymbol{\pi}$. Thus, if $\mathbf{z}\mathbf{e}^t = 0$ and $\mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1}\mathbf{z}^t = 0$, then $\mathbf{z} = \mathbf{0}$.*

Recall that a stationary Markov process is dynamically reversible (Kelly [3]) if there exists a permutation of $\{1,2,\ldots,m\}$, denoted by j^+, $j=1,2,\ldots,m$, such that (i) $\pi_j = \pi_{j^+}$ (ii) $q(j,j) = q(j^+,j^+)$ and (iii) $\pi_i q(i,j) = \pi_{j^+} q(j^+,i^+)$, where $q(i,j)$ is the element of Q in the i-th row and the j-th column. As a special case of dynamic reversibility, a Markov process is reversible if the permutation is identical, i.e., $j^+ = j$ for all j. In such a case, $Q\Delta(\boldsymbol{\pi})^{-1}$ is symmetric ([3], Exercise 1.4.2). Thus, Lemma 4.3 implies that $-Q\Delta(\boldsymbol{\pi})^{-1}$ is positive semidefinite (see [2]) if Q is reversible.

Proof. Note that

$$\mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1}\mathbf{z}^t = \sum_{i=1}^{m} z_i^2 \frac{q(i,i)}{\pi_i} + \sum_{i=1}^{m}\sum_{j\neq i} z_i z_j \frac{q(i,j)}{\pi_j}. \tag{48}$$

Since Q is a generator, $q(i,i) = -\sum_{j\neq i} q(i,j)$. Thus,

$$\sum_{i=1}^{m} z_i^2 \frac{q(i,i)}{\pi_i} = -\sum_{i=1}^{m}\sum_{j\neq i} z_i^2 \frac{q(i,j)}{\pi_i}. \tag{49}$$

From the dynamically reversible properties, $\pi_{i^+} = \pi_i$ and $q(i,i) = q(i^+,i^+)$, it follows that

$$\sum_{i=1}^{m} z_i^2 \frac{q(i,i)}{\pi_i} = \sum_{i=1}^{m} z_i^2 \frac{q(i^+,i^+)}{\pi_{i^+}} = -\sum_{i=1}^{m}\sum_{j\neq i} z_i^2 \frac{q(i^+,j^+)}{\pi_{i^+}}. \tag{50}$$

By interchanging the sums and the indices i and j, we have from (50) that

$$\sum_{i=1}^{m} z_i^2 \frac{q(i,i)}{\pi_i} = -\sum_{j=1}^{m}\sum_{i\neq j} z_i^2 \frac{q(i^+,j^+)}{\pi_{i^+}} = -\sum_{i=1}^{m}\sum_{j\neq i} z_j^2 \frac{q(j^+,i^+)}{\pi_{j^+}}. \tag{51}$$

Replacing

$$\sum_{i=1}^{m} z_i^2 \frac{q(i,i)}{\pi_i} = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j\neq i} z_i^2 \frac{q(i,j)}{\pi_i} - \frac{1}{2}\sum_{i=1}^{m}\sum_{j\neq i} z_j^2 \frac{q(j^+,i^+)}{\pi_{j^+}} \tag{52}$$

in (48) yields

$$\mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1}\mathbf{z}^t = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j\neq i}\left[z_i^2 \frac{q(i,j)}{\pi_i} + z_j^2 \frac{q(j^+,i^+)}{\pi_{j^+}} - 2z_i z_j \frac{q(i,j)}{\pi_j}\right]. \tag{53}$$

Since Q is dynamically reversible, $\pi_i q(i,j) = \pi_{j^+} q(j^+,i^+)$ and $\pi_{j^+} = \pi_j$. The expression in the square bracket of (53) can be rewritten as

$$\left[z_i\sqrt{\frac{q(i,j)}{\pi_i}} - z_j\sqrt{\frac{q(j^+,i^+)}{\pi_{j^+}}}\right]^2. \tag{54}$$

Thus, $\mathbf{z}Q\Delta^{-1}(\boldsymbol{\pi})\mathbf{z}^t \leq 0$. Moreover, it equals zero if and only if for $i \neq j$

$$z_i\sqrt{\frac{q(i,j)}{\pi_i}} = z_j\sqrt{\frac{q(j^+,i^+)}{\pi_{j^+}}}. \tag{55}$$

Using the dynamic reversibility $\pi_i q(i,j) = \pi_{j^+} q(j^+,i^+)$ and $\pi_j = \pi_{j^+}$ again, the equality in (55) is equivalent to $z_i/\pi_i = z_j/\pi_j$ if $q(i,j) > 0$. Since Q is irreducible, there is always a path from i to j such that $q(k,l) > 0$ for all intermediate states k and l. Thus, it follows that (55) is equivalent to $z_i/\pi_i = z_j/\pi_j = \gamma$ for all $i \neq j$ and some constant γ. ■

Proof of Theorem 4.2. It suffices to show the second term in (32) is larger than 0 under the assumptions given in Theorem 4.2. Again, we assume that $\mu_i = 1$ for all i and thus $\Delta(\boldsymbol{\mu}) = I$ and $\boldsymbol{\rho} = \boldsymbol{\lambda}$. Thus, we need to show that

$$\boldsymbol{\pi} R_4 \boldsymbol{\lambda}^t > 0 \tag{56}$$

given that R_4 satisfies (30) and (31). Let $\mathbf{z} = \boldsymbol{\pi} R_4$. It then follows from (30), (31) and (29) that

$$\mathbf{z}Q = -\boldsymbol{\pi}\Delta(\boldsymbol{\lambda}) + \boldsymbol{\pi} R_3 I = -\boldsymbol{\pi}\Delta(\boldsymbol{\lambda}) + (\boldsymbol{\pi}\boldsymbol{\lambda}^t)\boldsymbol{\pi} \tag{57}$$

and that

$$\mathbf{z}\mathbf{e}^t = \boldsymbol{\pi} R_4 \mathbf{e}^t = 0. \tag{58}$$

Note that $\boldsymbol{\pi}\Delta(\boldsymbol{\lambda}) = \boldsymbol{\lambda}\Delta(\boldsymbol{\pi})$. Thus, we have from (57) and (58) that

$$\mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1}\mathbf{z}^t = -\boldsymbol{\lambda}\mathbf{z}^t + (\boldsymbol{\pi}\boldsymbol{\lambda}^t)\mathbf{e}\mathbf{z}^t = -\mathbf{z}\boldsymbol{\lambda}^t. \tag{59}$$

It then follows from Lemma 4.3 that $\mathbf{z}\boldsymbol{\lambda}^t \geq 0$. If $\mathbf{z}\boldsymbol{\lambda}^t = 0$, then we have also from Lemma 4.3 that $\mathbf{z} = 0$ taking into account of (58). In conjunction with (57),

$$0 = \mathbf{z}Q\Delta(\boldsymbol{\pi})^{-1} = -\boldsymbol{\lambda} + (\boldsymbol{\pi}\boldsymbol{\lambda}^t)\mathbf{e}. \tag{60}$$

This implies λ_i's are all identical which violates our assumption. Thus, $\mathbf{z}\boldsymbol{\lambda}^t > 0$. ■

Though the matrix R_4 can be easily obtained by solving a set of linear equations, we show in the following example that a closed-form expression for the second-order approximation of the expected queue length could be derived.

Example 4.4 *Assume that μ_i=1 for all i. If the transition rate $q(i,j)$ from state i to j is α_j for all $i \neq j$, then under the assumptions in Theorem 4.1 the expected queue length is*

$$EL(\epsilon) = \frac{\bar{\rho}}{1-\bar{\rho}} + \epsilon^{-1}\frac{1}{\sum_{j=1}^m \alpha_j}\frac{\sigma^2}{1-\bar{\rho}} + O(\epsilon^{-2}), \tag{61}$$

where $\bar{\rho} = \sum_{j=1}^m \pi_i\lambda_i$ and $\sigma = (\sum_{j=1}^m \pi_i\lambda_i^2 - \bar{\rho}^2)^{1/2}$.

We note that the queue lengths in Example 4.4 are shown to be decreasing in ϵ (see [1]).

Proof. It suffices to show that

$$\boldsymbol{\pi}R_4\boldsymbol{\lambda}^t = (\alpha\mathbf{e}^t)^{-1}\sigma^2, \tag{62}$$

where $\alpha = (\alpha_1, \ldots, \alpha_m)$. Analogous to the proof in Theorem 4.2, let $\mathbf{z} = \boldsymbol{\pi}R_4$. Thus, we have from (57) and (58) that $\mathbf{z}$ is the unique solution of $\mathbf{z}Q = \boldsymbol{\pi}(\bar{\rho}I - \Delta(\boldsymbol{\lambda}))$ and $\mathbf{z}\mathbf{e}^t = 0$. Now we guess that $\mathbf{z} = -(\alpha\mathbf{e}^t)^{-1}\boldsymbol{\pi}(\bar{\rho}I - \Delta(\boldsymbol{\lambda}))$. One can easily show that $\mathbf{z}\boldsymbol{\lambda}^t = (\alpha\mathbf{e}^t)^{-1}\sigma^2$ and that $\mathbf{z}\mathbf{e}^t = -(\alpha\mathbf{e}^t)^{-1}(\bar{\rho} - \bar{\rho}) = 0$. Note that $\mathbf{z}Q = (\alpha\mathbf{e}^t)^{-1}\boldsymbol{\pi}\Delta(\boldsymbol{\lambda})Q$ since $\boldsymbol{\pi}Q = 0$. Thus, we only need to show that

$$(\alpha\mathbf{e}^t)^{-1}\boldsymbol{\pi}\Delta(\boldsymbol{\lambda})Q = \bar{\rho}\boldsymbol{\pi} - \boldsymbol{\pi}\Delta(\boldsymbol{\lambda}), \tag{63}$$

or equivalently that

$$\boldsymbol{\pi}\Delta(\boldsymbol{\lambda})(I + (\alpha\mathbf{e}^t)^{-1}Q) = \bar{\rho}\boldsymbol{\pi}. \tag{64}$$

Note that $\pi_i = \alpha_i/(\alpha\mathbf{e}^t)$ in this example. It is easy to verify that

$$I + (\alpha\mathbf{e}^t)^{-1}Q = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_m \\ \pi_1 & \pi_2 & \cdots & \pi_m \\ \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_m \end{bmatrix}.$$

Using this, the equation (64) can be easily verified. ■

Acknowledgments: The authors would like to thank the referees for their comments made after a careful reading of the manuscript.

References

[1] Chang, C.S., Chao, X.L., and Pinedo, M., "Monotonicity Results for Queues with Doubly Stochastic Poisson Arrivals: Ross's Conjecture," *Adv. Appl. Prob* 23, pp. 210-228, 1991.

[2] Horn, R.A. and Johnson, C.R., *Matrix Analysis*, Cambridge, 1987.

[3] Kelly, F.P., *Reversibility and Stochastic Networks*, John Wiley & Sons, 1979.

[4] Little, J.D.C., "A Proof of the Queueing Formula $L = \lambda W$," *Operat. Res.* 9, pp. 383-387, 1961.

[5] Neuts, M.F., *Matrix Geometric Solutions in Stochastic Models*, John Hopkins University Press, 1981.

[6] Neuts, M.F., *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, Inc., 1989.

[7] Ross, S.,"Average Delay in Queues with Non-stationary Poisson Arrivals," *J. Appl. Prob.* 15, pp. 602-609, 1978.

[8] Ross, S., *Stochastic Processes*, Wiley, 1983.

[9] Rolski, T.,"Queues with Non-stationary Input Stream: Ross's Conjecture," *Adv. Appl. Prob.* 13, pp. 603-618, 1981

[10] Rolski, T.,"Queues with Nonstationary Arrivals," *Queueing Systems* 5, pp. 113-130, 1989.

[11] Stoyan, D., *Comparison Methods for Queues and Other Stochastic Models*. Berlin: J. Wiley & Sons, 1983.