Rate of convergence to stationarity of the system $M/M/N/N+R$

Erik A. van Doorn

Department of Applied Mathematics, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
E-mail: e.a.vandoorn@utwente.nl
Fax: +31 (0)53 4893069, Phone: +31 (0)53 4893387

18 January 2011

Abstract. We consider the $M/M/N/N+R$ service system, characterized by N servers, R waiting positions, Poisson arrivals and exponential service times. We discuss representations and bounds for the rate of convergence to stationarity of the number of customers in the system, and study its behaviour as a function of R, N and the arrival rate λ, allowing λ to be a function of N.

Keywords and phrases: decay rate, delay and loss system, many-server queue, orthogonal polynomials

2000 Mathematics Subject Classification: Primary 60K25, Secondary 90B22

1 Introduction

We consider the $M/M/N/N+R$ service system, characterized by Poisson arrivals, exponential service times, $N \geq 1$ servers and $R \geq 0$ waiting places. With $\lambda > 0$ denoting the arrival rate and $\mu > 0$ the service rate per server, the number of customers in this system is a birth-death process $\mathcal{X} \equiv \{X(t),\ t \geq 0\}$ taking values in $S \equiv \{0, 1, \ldots, N+R\}$, with birth and death rates

$$\lambda_j = \lambda, \quad 0 \leq j < N+R, \quad \text{and} \quad \mu_j = \min\{j, N\}\mu, \quad 0 < j \leq N+R,$$

respectively. We write $p_j(t) \equiv \Pr\{X(t) = j\},\ j \in S$, and let the vector $\boldsymbol{p}(t) \equiv (p_0(t), p_1(t), \ldots, p_{N+R}(t))$ represent the state distribution at time $t \geq 0$. The stationary distribution of $\mathcal{X}$ will be represented by the vector $\boldsymbol{\pi} \equiv (\pi_0, \pi_1, \ldots, \pi_{N+R})$, where

$$\pi_j = \begin{cases} c\dfrac{a^j}{j!}, & 0 \leq j \leq N \\ c\dfrac{a^j}{N!N^{j-N}}, & N < j \leq N+R, \end{cases} \tag{1}$$

$a \equiv \lambda/\mu$, and c is a normalizing constant. For any initial distribution $\boldsymbol{p}(0)$ the vector $\boldsymbol{p}(t)$ converges to $\boldsymbol{\pi}$ as $t \to \infty$.

In what follows we will be concerned with the speed of convergence to stationarity of the $M/M/N/N+R$ service system, represented by the rate of convergence to zero of

$$d_{tv}(\boldsymbol{p}(t), \boldsymbol{\pi}) \equiv \sup_{A \subset S} \left\{ \left| \sum_{j \in A} p_j(t) - \sum_{j \in A} \pi_j \right| \right\} = \frac{1}{2} \sum_{j \in S} |p_j(t) - \pi_j|,$$

the *total variation distance* between $\boldsymbol{p}(t)$ and $\boldsymbol{\pi}$. That is, we focus on

$$\beta = \sup\{b > 0 :\ d_{tv}(\boldsymbol{p}(t), \boldsymbol{\pi}) = \mathcal{O}(e^{-bt}) \ \text{ as } \ t \to \infty \ \text{ for all } \ \boldsymbol{p}(0)\}, \tag{2}$$

and will refer to this quantity as the rate of convergence (or *decay rate*) of the $M/M/N/N+R$ service system. The reciprocal of β is sometimes called the *relaxation time* of the system (see, for example, Keilson and Ramaswamy [14]). Since the behaviour of β as a function of $\lambda,\ N$ and R will be of interest to us, we will often indicate this dependence by writing $\beta(\lambda, N, R)$ instead of β.

The plan of the paper is as follows. Representations and bounds for $\beta \equiv \beta(\lambda, N, R)$ will be discussed in Section 2. Then, in Section 3, we investigate how β behaves as a function of the arrival rate λ for constant N and R. The behaviour of β as a function of N and R is studied in Section 4 under the assumption that λ is constant. In Section 5 we discuss asymptotic results for β as $N \to \infty$, assuming a constant *traffic intensity* ρ (so that $\lambda \equiv \lambda(N) = \rho\mu N$) and $\rho \neq 1$. Asymptotic results for the borderline case $\lambda = \mu N$, and, more generally, $\lambda \sim \mu N$ as $N \to \infty$, are discussed in Section 6.

Pivotal in our approach are the identifications of β as the smallest zero of a polynomial that can be expressed in terms of orthogonal polynomials (Theorem 1), and as the smallest zero of a polynomial that is itself an element of an orthogonal-polynomial sequence (Theorem 2). An appeal to orthogonal-polynomial theory subsequently enables us to draw conclusions about the behaviour of $\beta(\lambda, N, R)$ as a function of one of the parameters, and to identify the limit as this parameter goes to infinity.

Our results generalize those of [6], which concern the case $R = 0$ (the *Erlang loss model*).

2 Representations for β

It is well known that the supremum in (2) is in fact a maximum, and that $-\beta$ equals the largest nonzero eigenvalue of the $(N + R + 1) \times (N + R + 1)$ matrix

$$Q \equiv \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & \cdots & 0 & 0 & 0 \\ 0 & 2\mu & -(\lambda+2\mu) & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & \cdots & N\mu & -(\lambda+N\mu) & \lambda & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & N\mu & -(\lambda+N\mu) & \lambda \\ 0 & 0 & 0 & \cdots & 0 & N\mu & -N\mu \end{pmatrix},$$

the q-matrix of $\mathcal{X}$. From Karlin and McGregor [13] we know that the nonzero eigenvalues of $-Q$ can be identified with the (distinct and positive) zeros of

$$S(x) = \frac{1}{x}\left\{(x - N\mu)P_{N+R}(x) + N\mu P_{N+R-1}(x)\right\},$$

where the P_n are polynomials satisfying the recurrence relation

$$\begin{aligned} &P_{-1}(x) = 0, \quad P_0(x) = 1, \\ &\lambda P_{n+1}(x) = (\lambda + n\mu - x)P_n(x) - n\mu P_{n-1}(x), \quad 0 \le n \le N, \\ &\lambda P_{n+1}(x) = (\lambda + N\mu - x)P_n(x) - N\mu P_{n-1}(x), \quad n > N. \end{aligned} \tag{3}$$

So β is the smallest zero of $S(x)$. Note that $P_n(0) = 1$ for all $n \ge 0$, so that $S(x)$ is a polynomial of degree $N + R$, which, by (3), can be represented as

$$S(x) = \frac{\lambda}{x}\left(P_{N+R}(x) - P_{N+R+1}(x)\right). \tag{4}$$

Karlin and McGregor [12, Section 4] have shown that

$$P_n(\mu x) = c_n(x), \quad n \le N,$$

and, for $n > 0$,

$$P_{N+n}(\mu x) = \left(\frac{N}{a}\right)^{n/2}\left(c_N(x)U_n(\xi(x)) - \left(\frac{N}{a}\right)^{1/2} c_{N-1}(x)U_{n-1}(\xi(x))\right), \tag{5}$$

where

$$\xi(x) \equiv \xi(x, a, N) = \frac{1}{2}\frac{N + a - x}{\sqrt{aN}}, \tag{6}$$

the c_n are *Charlier polynomials*, given by

$$c_n(x) \equiv c_n(x, a) = \sum_{k=0}^{n}(-1)^k \binom{n}{k}\binom{x}{k}\frac{k!}{a^k}, \quad n \ge 0, \tag{7}$$

and the U_n are *Chebysev polynomials of the second kind*, defined by

$$U_n(\xi) = \frac{z^{n+1} - z^{-(n+1)}}{z - z^{-1}}, \quad \xi = \frac{1}{2}(z + z^{-1}), \quad n \ge 0. \tag{8}$$

We note for future use that the zeros of $U_n(\xi)$ are real and in the interval $(-1, 1)$ (see, for example, Chihara [3]). Moreover, we have $z \notin \mathbb{R}$ if and only if $|\xi| < 1$, in which case $|z| = 1$ and

$$U_n(\xi) = \frac{\sin(n+1)\phi}{\sin\phi}, \tag{9}$$

with $\xi = \cos\phi$ and $0 \le \phi \le \pi$.

The results (5)-(8) may be substituted in (4), but a more convenient expression for $S(x)$ is obtained by employing the relation

$$2\xi U_n(\xi) = U_{n-1}(\xi) + U_{n+1}(\xi), \quad n \ge 0, \tag{10}$$

(see, for example, [3, p. 25]), and the relations

$$c_n(x) - c_{n-1}(x) = -\frac{x}{a} c_{n-1}(x-1), \quad n > 0, \tag{11}$$

and

$$c_n(x) - c_n(x-1) = -\frac{n}{a} c_{n-1}(x-1), \quad n > 0 \tag{12}$$

(see, for example, Jagerman [11]). Namely, writing

$$v_n(x) \equiv v_n(x, a, N) = \left(\frac{a}{N}\right)^{n/2} U_n(\xi(x)), \tag{13}$$

we find that the v_n satisfy the recurrence relation

$$\begin{aligned} &v_{-1}(x) = 0, \quad v_0(x) = 1, \\ &(N + a - x)v_n(x) = a v_{n-1}(x) + N v_{n+1}(x), \quad n > 0, \end{aligned} \tag{14}$$

while

$$\left(\frac{a}{N}\right)^n P_{N+n}(\mu x) = c_N(x) v_n(x) - c_{N-1}(x) v_{n-1}(x), \quad n \ge 0. \tag{15}$$

Next setting $T(x) = (a/N)^R S(\mu x)$, it follows with (4) that

$$xT(x) = (a v_R(x) - N v_{R+1}(x)) c_N(x) - (a v_{R-1}(x) - N v_R(x)) c_{N-1}(x). \tag{16}$$

By (11) we may replace $c_{N-1}(x)$ by $c_N(x) + \frac{x}{a} c_{N-1}(x-1)$. Rearranging and employing (14) subsequently gives us

$$T(x) = \left(c_N(x) + \frac{N}{a} c_{N-1}(x-1)\right) v_R(x) - c_{N-1}(x-1) v_{R-1}(x),$$

which, by (12), reduces to

$$T(x) = c_N(x-1) v_R(x) - c_{N-1}(x-1) v_{R-1}(x). \tag{17}$$

Thus we have obtained the following characterization of β.

Theorem 1. The rate of convergence β of the $M/M/N/N+R$ service system equals μ times the smallest root of the polynomial $T(x)$ of (17), where v_n is given by (13), (6) and (8).

By way of illustration we will look at two special cases. First suppose $N = 1$. Then the representation of Theorem 1 leads to an explicit result. Namely, since $ac_1(x-1) = 1 + a - x$, (14) and (13) imply

$$T(x) = a^{(R-1)/2} U_{R+1}(\xi(x)).$$

It follows from the properties of Chebysev polynomials mentioned below (8) that $\cos(n\pi/(R+2))$, $n = 1, 2 \ldots, R+1$, are the zeros of $U_{R+1}(\xi)$, so (6) implies that $1 + a - 2\sqrt{a}\cos(n\pi/(R+2))$, $n = 1, 2, \ldots, R+1$, are the zeros of $T(x)$. Hence,

$$\beta(\lambda, 1, R) = \lambda + \mu - 2\sqrt{\lambda\mu}\cos(\pi/(R+2)), \tag{18}$$

which is a known result (cf. Takács [18, p. 13] or Kijima [17, p. 203]).

Secondly, let $R = 0$. Then we have $T(x) = c_N(x-1)$, so that

$$\beta(\lambda, N, 0) = \mu + \mu\xi_{N,1}, \tag{19}$$

where $\xi_{N,1}$ denotes the smallest zero of the Charlier polynomial $c_N(x)$. An explicit expression for $\xi_{N,1}$, and hence for $\beta(\lambda, N, 0)$, is available only for small values of N. In particular, it is easy to see that

$$\beta(\lambda, 1, 0) = \lambda + \mu \tag{20}$$

and

$$\beta(\lambda, 2, 0) = \lambda + \frac{3}{2}\mu - \frac{1}{2}\sqrt{\mu^2 + 4\lambda\mu}. \tag{21}$$

See [6] for representations and bounds for $\beta(\lambda, N, 0)$ when $N > 2$.

The fact that β is the smallest zero of the polynomial $S(x)$ can be embedded in a somewhat different context, yielding additional information. Namely, defining the polynomials

$$Q_n(x) = \frac{(-\lambda)^{n+1}}{x}(P_{n+1}(x) - P_n(x)), \quad n \geq 0, \tag{22}$$

we see from (4), that $Q_{N+R}(x) = (-\lambda)^{N+R} S(x)$. Moreover, in view of (3) the polynomials Q_n are easily seen to satisfy the recurrence relations

$$\begin{aligned} &Q_0(x) = 1, \quad Q_1(x) = x - \lambda - \mu, \\ &Q_n(x) = (x - \lambda - n\mu)Q_{n-1}(x) - (n-1)\lambda\mu Q_{n-2}(x), \quad 1 < n \leq N, \\ &Q_n(x) = (x - \lambda - N\mu)Q_{n-1}(x) - N\lambda\mu Q_{n-2}(x), \quad n > N. \end{aligned} \tag{23}$$

It then follows by *Favard's theorem* that the Q_n constitute a sequence of orthogonal polynomials. (See Chihara [3] for this and subsequent basic results on orthogonal polynomials.) Hence, $Q_n(x)$ has n real and simple zeros $x_{n1} < x_{n2} < \cdots < x_{nn}$. So, since β is the smallest zero of the polynomial $S(x)$, we obtain our second representation.

Theorem 2. The rate of convergence β of the $M/M/N/N+R$ service system equals $x_{N+R,1}$, the smallest zero of the polynomial $Q_{N+R}(x)$ defined by (23).

We note that the polynomial Q_n, $n > N$, can be interpreted as the characteristic polynomial of the $n \times n$ matrix $-A_n$, with

$$A_n \equiv \begin{pmatrix} -(\lambda+\mu) & \mu & 0 & \cdots & 0 & 0 & 0 \\ \lambda & -(\lambda+2\mu) & 2\mu & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \lambda & -(\lambda+N\mu) & N\mu & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & -(\lambda+N\mu) & N\mu \\ 0 & 0 & 0 & \cdots & 0 & \lambda & -(\lambda+N\mu) \end{pmatrix}, \tag{24}$$

so that the zeros of $Q_n(x)$ are the eigenvalues of $-A_n$. This can be seen by setting $Q_n(x) = \det(A_n + xI)$, expanding the determinant by its last row, and noting that the Q_n satisfy the recurrence relation (23). (See [14] and [7] for other approaches towards this identification.) It follows in particular that β is the smallest eigenvalue of the matrix $-A_{N+R}$, a result to which we will have reference in the next section.

Since $\{Q_n\}$ constitutes an orthogonal polynomial sequence, the zeros of $Q_n(x)$ and $Q_{n+1}(x)$ separate each other, that is,

$$x_{n+1,i} < x_{ni} < x_{n+1,i+1}, \quad i = 1, 2, \ldots, n, \ n \geq 1.$$

It follows that x_{n1} is a stricly decreasing sequence as n increases, and, as a consequence, $\beta(\lambda, N, R)$ is strictly decreasing in R for fixed N. Note that N, unlike R, appears as a parameter in the recurrence relation (23), so the preceding does not imply that $\beta(\lambda, N, R)$ decreases in N for fixed R.

Remark. The fact that $\beta(\lambda, N, R)$ is strictly decreasing in R for fixed N is also implied by Chen [1, Proposition 3.4] and Granovsky and Zeifman [9, Corollary 3], who use different arguments to prove their results (in the more general setting of finite birth-death processes). □

Characterizations of β of an entirely different nature are obtained by applying a result of Zeifman's [19] (see also [7, Theorem 7]) on birth-death processes to the pertinent setting.

Theorem 3. The rate of convergence β of the $M/M/N/N+R$ service system satisfies

$$\max_{\boldsymbol{x}>\mathbf{0}} \left\{ \min_{1\leq j\leq N+R} \alpha_j(\boldsymbol{x}) \right\} = \beta = \min_{\boldsymbol{x}>\mathbf{0}} \left\{ \max_{1\leq j\leq N+R} \alpha_j(\boldsymbol{x}) \right\},$$

where $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_{N+R-1})$, and

$$\alpha_j(\boldsymbol{x}) = \begin{cases} \lambda(1 - x_j^{-1}) + \mu(j - (j-1)x_{j-1}) & 1 \leq j \leq N \\ \lambda(1 - x_j^{-1}) + \mu N(1 - x_{j-1}) & N < j \leq N+R, \end{cases} \tag{25}$$

with $x_0 = x_{N+R}^{-1} = 0$.

Here $\mathbf{0}$ denotes a vector of zeros, and inequality for vectors indicates elementwise inequality. It follows in particular that for *any* vector $\boldsymbol{x} > \mathbf{0}$

$$\min_{1\leq j\leq N+R} \alpha_j(\boldsymbol{x}) \leq \beta \leq \max_{1\leq j\leq N+R} \alpha_j(\boldsymbol{x}). \tag{26}$$

For example, assuming $N > 1$ we can choose $x_j = 1$ for $1 \leq j < N$, and, if $R > 0$, $x_{N+R-1} = 1 - \frac{1}{N}(1 - \frac{\lambda}{\mu})$, and, if $R > 1$, $x_j = 1 - \frac{1}{N}$ for $N \leq j < N+R-1$.

It then follows that

$$\alpha_i(\boldsymbol{x}) = \begin{cases} \mu, & 1 \le i < N \\ \mu - \dfrac{\lambda}{N-1}, & N \le i < N+R-1 \\ \mu + \dfrac{\lambda(\lambda-\mu)}{\lambda+(N-1)\mu}, & i = N+R-1 \text{ and } R > 0 \\ \mu + \lambda \mathbb{I}_{\{R=0\}}, & i = N+R, \end{cases}$$

where $\mathbb{I}_A$ denotes the indicator function of the event A. Hence, for $N > 1$ we obtain the bounds

$$\mu \le \beta(\lambda, N, 0) \le \mu + \lambda, \tag{27}$$

$$\lambda \le \mu \implies \mu - \frac{\lambda(\mu-\lambda)}{\lambda+(N-1)\mu} \le \beta(\lambda, N, 1) \le \mu, \tag{28}$$

$$\lambda > \mu \implies \mu \le \beta(\lambda, N, 1) \le \mu + \frac{\lambda(\lambda-\mu)}{\lambda+(N-1)\mu}, \tag{29}$$

while for $N > 1$ and $R > 1$ we have

$$\lambda \le \mu \implies \mu - \frac{\lambda}{N-1} \le \beta(\lambda, N, R) \le \mu, \tag{30}$$

$$\lambda > \mu \implies \mu - \frac{\lambda}{N-1} \le \beta(\lambda, N, R) \le \mu + \frac{\lambda(\lambda-\mu)}{\lambda+(N-1)\mu}. \tag{31}$$

Further representations for β may be obtained by symmetrizing the matrix Q – or the matrix A_{N+R} – by means of a similarity transformation, and applying the *Courant-Fischer Theorem* for symmetric matrices. (This approach has been elaborated in [6] in the case $R = 0$.) Since we shall not use the resulting expressions in what follows, we will not spell them out.

3 Behaviour of β as a function of λ

The representation of β as the smallest eigenvalue of the matrix $-A_{N+R}$ defined by (24), readily implies the limits

$$\lim_{\lambda\to 0} \beta(\lambda, N, R) = \mu \quad \text{and} \quad \lim_{\lambda\to\infty} \frac{\beta(\lambda, N, R)}{\lambda} = \lim_{\mu\to 0} \beta(1, N, R) = 1, \tag{32}$$

since the eigenvalues of a matrix are continuous functions of the matrix elements. Kijima [15, Theorem 1] has shown that $\beta(\lambda, N, R)$ is a strictly increasing function of λ when $R = 0$, but his method of proof (which hinges on the

observation that Perron-Frobenius theory may be applied to the matrix $A_n + rI$ for r sufficiently large) breaks down when $R > 0$. Indeed, from the explicit expression (18) we note that $\beta(\lambda, 1, 1)$ decreases for λ sufficiently small. However, we can prove monotonicity of $\beta(\lambda, N, R)$ as a function of λ for λ sufficiently large. In the proof of this result we shall use the concept of a (finite) *chain sequence*, which is a numerical sequence $\{a_k\}_{k=1}^K$ for which there exists a sequence $\{g_k\}_{k=0}^K$ – a *parameter sequence* for $\{a_k\}$ – such that

$$(i) \quad 0 \le g_0 < 1, \quad 0 < g_k < 1, \ k = 1, 2, \ldots, K-1, \quad 0 < g_K \le 1,$$

$$(ii) \quad a_k = (1 - g_{k-1})g_k, \quad k = 1, 2, \ldots, K.$$

Theorem 4. For constant $N \ge 1$ and $R \ge 0$ the function $\beta(\lambda, N, R)$ is strictly increasing in λ for $\lambda \ge N\mu$.

Proof. By using the representation for β of Theorem 2 and applying Theorem 3.3 in Ismail and Muldoon [10] to the orthogonal polynomial sequence $\{Q_n\}$, we conclude that all zeros of $Q_{N+R}(x)$, and hence $\beta = x_{N+R,1}$ in particular, are strictly increasing functions of λ if the sequence $\{a_k\}_{k=1}^{N+R}$, where

$$a_k = \frac{\min\{k, N\}\mu}{4\lambda},$$

is a chain sequence. If $\lambda \ge N\mu$ then $a_k \le \frac{1}{4}$. Since the constant sequence $\{\frac{1}{4}\}$ is a chain sequence, while, by [3, Theorem III.5.7], a sequence of positive numbers is itself a chain sequence if it is dominated by a chain sequence, the result follows. □

Remarks. *(i)* The monotonicity of $\beta(\lambda, N, 0)$ as a function of λ for all $\lambda > 0$, as well as the results (32) for the special case $R = 0$, are also given in [20, Corollary 29].

(ii) The lower bound for λ in Theorem 4 can be slightly improved (that is, decreased) by noting that the sequence whose kth element is $\frac{1}{4} + \frac{1}{16k(k+1)}$ is a chain sequence (see [3, p. 98]). □

4 Behaviour of β as a function of N and R

In this section we are interested in the behaviour of β as a function of N and R. In Section 2 we have noted already that $\beta(\lambda, N, R)$ is strictly decreasing

in R for fixed N. To characterize $\lim_{R\to\infty}\beta(\lambda,N,R)$ we recall another result from the theory of orthogonal polynomials (see [3]). Namely, the smallest zeros x_{n1} of the polynomials Q_n converge, as $n\to\infty$, to a real number x_1, which is the first point in the support of the Borel measure with respect to which the polynomials Q_n are orthogonal. (The orthogonalizing measure is unique since the parameters in the recurrence relation are bounded.) So, by Theorem 2, $\lim_{R\to\infty}\beta(\lambda,N,R)=x_1$. Moreover, from [5, Section 2.4] we see that the sequence $\{Q_n\}$ is actually the *dual* of the sequence $\{P_n\}$ defined by (3), while the latter is the sequence of orthogonal polynomials corresponding to the (birth-death) process $\mathcal{X}^\infty$ of the number of customers in the system $M/M/N/\infty$. From [5, Theorem 3.3] we therefore conclude that $\lim_{R\to\infty}\beta(\lambda,N,R)$ equals $\beta(\lambda,N,\infty)$, the rate of convergence to stationarity of the process $\mathcal{X}^\infty$.

We point out that here (and in what follows) the rate of convergence of an irreducible birth-death process taking values in the *countably infinite* state space $\{0,1,\dots\}$, is defined as

$$\beta=\sup\{b>0:\ |p_j(t)-\pi_j|=\mathcal{O}(e^{-bt})\ \text{ as }\ t\to\infty\ \text{ for all } j \text{ and } \boldsymbol{p}(0)\},\quad (33)$$

where $\pi_j=\lim_{t\to\infty}p_j(t)$ (and hence $\pi_j=0$ if the process is not ergodic). Chen [2] has shown that the definitions (2) and (33) are actually equivalent if the birth-death process is ergodic, but (2) is obviously unsuitable as a definition of decay rate if the process is transient or null recurrent, that is, in the setting of $\mathcal{X}^\infty$, if $\lambda\geq\mu N$.

Summarizing we can state the following theorem.

Theorem 5. For constant $\lambda>0$ and $N\geq 1$ the function $\beta(\lambda,N,R)$ is strictly decreasing in R, and converges to $\beta(\lambda,N,\infty)$ as $R\to\infty$.

Remark. It has been observed in [9] in the setting of *ergodic* birth-death processes that the convergence of $\beta(\lambda,N,R)$ to $\beta(\lambda,N,\infty)$ as $R\to\infty$ may also be established by an appeal to the *Trotter-Kurtz Theorem* on the convergence of strongly continuous semigroups and their generators. □

No explicit expression for $\beta(\lambda,N,\infty)$ exists, but information on how to obtain its value can be found in [4, Chapter 6], a summary of which is given in [5,

Section 4, Example 2] (see also Kijima [16, Example 3.1]). Specifically, it is shown in [4] that

$$\beta(\lambda, N, \infty) \le \left(\sqrt{\lambda} - \sqrt{\mu N}\right)^2,$$

while for every N there exists a real number $\rho_N^*,\ 0 \le \rho_N^* < 1$, such that

$$\beta(\lambda, N, \infty) = \left(\sqrt{\lambda} - \sqrt{\mu N}\right)^2 \iff \rho \equiv \frac{\lambda}{\mu N} \ge \rho_N^*. \tag{34}$$

If $\rho < \rho_N^*$ then $\beta(\lambda, N, \infty)$ equals μ times the second smallest root of the equation

$$\frac{c_N(x)}{c_{N-1}(x)} = \frac{1}{2a}\left(N + a - x - \sqrt{(N + a - x)^2 - 4aN}\right) \tag{35}$$

(recall that $a \equiv \lambda/\mu$), the smallest root of this equation being 0.

For example, it is shown in [12] that

$$\beta(\lambda, 2, \infty) = \lambda + \frac{1}{2}\mu + \frac{1}{2}\sqrt{\mu^2 - 4\lambda\mu} \tag{36}$$

if $\rho < \rho_2^* = \frac{1}{9}$. Some more values of ρ_N^* are listed in [4]; specifically, we have $\rho_1^* = 0, \rho_2^* = 1/9,\ \rho_3^* = 2(4+\sqrt{7})/63 \approx 0.211,\ \rho_4^* \approx 0.284$, and $\rho_5^* \approx 0.340$. Moreover, it has recently been established by Gamarnik and Goldberg [8, Corollary 1] that

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho_N^*) = B^*, \tag{37}$$

where $B^* \approx 1.8572$ is the solution of an equation involving parabolic cylinder functions.

Remark. Choosing (16) as a starting point we observe that $xT(x) = 0$ if and only if

$$\frac{c_N(x)}{c_{N-1}(x)} = \frac{av_{R-1}(x) - Nv_R(x)}{av_R(x) - Nv_{R+1}(x)}.$$

Letting $R \to \infty$ in the latter equation (and assuming $\mu x < (\sqrt{\lambda} - \sqrt{\mu N})^2$), results after some algebra in the equation (35), whose second smallest root equals $\beta(\lambda, N, \infty)/\mu$ in the case $\rho < \rho_N^*$. An alternative characterization of $\beta(\lambda, N, \infty)$ in the case $\rho < \rho_N^*$ may be obtained by choosing (17) rather than (16) as a starting point, that is, letting $R \to \infty$ in

$$\frac{c_N(x-1)}{c_{N-1}(x-1)} = \frac{v_{R-1}(x)}{v_R(x)}.$$

By *Hurwitz' Theorem* $\beta(\lambda, N, \infty)/\mu$ must be the smallest root of the resulting equation

$$\frac{c_N(x-1)}{c_{N-1}(x-1)} = \frac{1}{2a}\left(N + a - x - \sqrt{(N+a-x)^2 - 4aN}\right).$$

Apparently, if $\rho < \rho_N^*$ then $x = \beta(\lambda, N, \infty)/\mu$ solves the equation

$$\frac{c_N(x-1)}{c_{N-1}(x-1)} = \frac{c_N(x)}{c_{N-1}(x)}.$$ □

Next assuming R to be constant we wish to obtain information on the behaviour of $\beta(\lambda, N, R)$ as N increases. This, however, appears to be a more complicated problem, since N, unlike R, features as a parameter in the recurrence relation (23). However, we do have the following result, which is formulated in a setting that encompasses the case of a constant arrival rate.

Theorem 6. If $R \geq 0$ is constant and $\lambda \equiv \lambda(N) = o(\sqrt{N})$ as $N \to \infty$, then $\lim_{N\to\infty} \beta(\lambda, N, R) = \mu$.

This result follows immediately from the bounds (28)-(31) when $R > 0$, and is implied by [6, Theorem 7] when $R = 0$. We note that the limit μ is, in fact, the rate of convergence of the $M/M/\infty$ service system (see, for example, [12]).

Remark. Since the bounds (30) and (31) are independent of R, we also have $\lim_{N\to\infty} \beta(\lambda, N, \infty) = \mu$ if $\lambda = o(\sqrt{N})$ as $N \to \infty$. □

Monotonicity of $\beta(\lambda, N, 0)$ as a function of N has been established in [20, Corollary 28] by using the representation (3). An alternative argument employs the fact that Charlier polynomials are orthogonal with respect to a measure consisting of point masses at the points $0, 1, \ldots$, so that $\xi_{N,1}$, the smallest zero of the Charlier polynomial $c_N(x)$, decreases to 0 as $N \to \infty$. Hence, by (19), $\beta(\lambda, N, 0)$ *decreases* to μ as $N \to \infty$. Since $\lim_{R\to\infty} \beta(\lambda, 1, R) = (\sqrt{\lambda} - \sqrt{\mu})^2$ is smaller than $\lim_{N\to\infty} \beta(\lambda, N, R) = \mu$ if $\lambda < 4\mu$, the function $\beta(\lambda, N, R)$ will not be decreasing in N in general.

5 Asymptotics for β if $\lambda = \rho\mu N$ with $\rho \neq 1$

In this section we are mainly interested in the limiting behaviour of $\beta(\lambda, N, R)$ as $N \to \infty$ assuming that $\lambda \equiv \lambda(N) = \rho\mu N$ for some constant traffic intensity

$\rho \neq 1$, while the number of waiting positions R is arbitrary but fixed. However, we start off in a more general setting by observing the following.

Lemma 7. Let $c < \mu$ and $R \geq 0$ be constants and suppose $\lambda \equiv \lambda(N) \leq cN$ for N sufficiently large. Then

$$\mu - c\mathbb{I}_{\{R>1\}} \leq \lim_{N\to\infty} \inf \beta(\lambda, N, R) \leq \lim_{N\to\infty} \sup \beta(\lambda, N, R) \leq \mu.$$

Proof. We have $\beta(\lambda, N, R) \leq \beta(\lambda, N, 0)$ by Theorem 5, while $\beta(\lambda, N, 0) \to \mu$ as $N \to \infty$ under the condition imposed on λ, by [6, Theorem 7]. This proves the upper bound. The lower bounds in (28)-(31) imply the lower bound. □

However, we can do better in the special case $\lambda = \rho\mu N$, with $\rho < 1$. Namely, by applying Theorem 1, we see that $\beta(\lambda, N, R)$ can be represented as $\mu N x^*$, where x^* is the smallest root of the equation

$$\frac{c_N(Nx-1)}{c_{N-1}(Nx-1)} = \frac{1}{\sqrt{\rho}} \frac{U_{R-1}(\xi(Nx))}{U_R(\xi(Nx))},$$

which, in view of (6) and (8), reduces to the equation

$$\frac{c_N(Nx-1)}{c_{N-1}(Nx-1)} = H_R(x) \equiv \frac{1}{\sqrt{\rho}} \left(\frac{z^R - z^{-R}}{z^{R+1} - z^{-(R+1)}} \right), \tag{38}$$

where z is such that

$$z + z^{-1} = 2\xi(Nx) = \frac{1+\rho-x}{\sqrt{\rho}}. \tag{39}$$

As noted before, we have $z \notin \mathbb{R}$ if and only if $|\xi(Nx)| < 1$, that is, $(1-\sqrt{\rho})^2 < x < (1+\sqrt{\rho})^2$. In view of (9) $H_R(x)$ can then be represented as

$$H_R(x) = \frac{1}{\sqrt{\rho}} \frac{\sin R\phi}{\sin(R+1)\phi},$$

with ϕ such that $0 \leq \phi \leq \pi$, and

$$\cos\phi = \frac{1+\rho-x}{2\sqrt{\rho}}.$$

Observe that $H_R(x)$ is a positive, continuous function in the interval $0 \leq x < (1-\sqrt{\rho})^2$. Moreover,

$$H_R(0) = \frac{\rho^{-(R+1)/2} - \rho^{(R-1)/2}}{\rho^{-(R+1)/2} - \rho^{(R+1)/2}} < 1, \tag{40}$$

and

$$H_R((1-\sqrt{\rho})^2) = \frac{1}{\sqrt{\rho}} \frac{R}{R+1} \geq 0. \tag{41}$$

Theorem 8. Let $R \geq 0$ and $\rho < 1$. Then

$$\lim_{N\to\infty} \beta(\rho\mu N, N, R) = \mu.$$

Proof. In view of the preceding lemma it suffices to show, for $R > 1$, that $\beta(\lambda, N, R) \geq \mu$, that is, $x^* > N^{-1}$, for N sufficiently large. We denote, as before, the smallest zero of $c_n(x)$ by $\xi_{n,1}$, and recall from the theory of orthogonal polynomials that $\xi_{n,1}$ is positive and decreasing in n. As a consequence, by choosing N sufficiently large, we have $(1 + \xi_{N-1,1})/N < (1 - \sqrt{\rho})^2$. Moreover, since $H_R(0) < 1$, we also have $H_R(1/N) < 1$ by choosing N sufficiently large. It is shown in [4, p. 50] that the function $c_N(x)/c_{N-1}(x)$ decreases continuously from $+\infty$ to $-\infty$ in the interval $-\infty < x < \xi_{N-1,1}$. Since $c_n(0) = 1$, it follows that $c_N(Nx - 1)/c_{N-1}(Nx - 1)$ decreases continuously from 1 to $-\infty$ in the interval $[1/N, (1 + \xi_{N-1,1})/N)$. So, in view of the behaviour of $H_R(x)$ on this interval, we must have $x^* > N^{-1}$ for N sufficiently large, as required. □

Remark. It is not difficult to see that $H_R(x)$ is increasing in R, and $H_\infty(1/N) > 1$, so that $H_R(1/N) > 1$ for R sufficiently large. It follows that $\beta(\rho\mu N, N, R) < \mu$ for R sufficiently large (and $\rho < 1$). □

To obtain an asymptotic result in the case $\rho > 1$, we note that, by Theorem 5 and (34),

$$\beta(\lambda, N, R) > \beta(\lambda, N, \infty) = \mu N(\sqrt{\rho} - 1)^2, \tag{42}$$

if $\rho \equiv \lambda/(\mu N) \geq \rho_N^*$. This observation enables us to prove the next theorem, which, together with Theorem 8, generalizes [7, Theorem 12] on the Erlang loss system ($R = 0$).

Theorem 9. Let $R \geq 0$ and $\rho > 1$. Then

$$\lim_{N\to\infty} \frac{\beta(\rho\mu N, N, R)}{N} = \mu(\sqrt{\rho} - 1)^2.$$

Proof. By [7, Theorem 12] we know the result to be valid for $R = 0$, so we may assume in what follows that $R \geq 1$. Moreover, by Theorem 5 we have $\beta(\lambda, N, R) \leq \beta(\lambda, N, 0)$, so the result is implied by (42) since $\rho_N^* < 1$. □

The case $\lambda = \mu N$ is apparently a borderline case. In the next section we will study the asymptotic behaviour of $\beta(\lambda, N, R)$ in the more general setting $\lambda \sim \mu N$ as $N \to \infty$.

6 Asymptotics for β if $\lambda \sim \mu N$ as $N \to \infty$

We will first study asymptotics in the case of a constant traffic intensity $\rho = 1$.

Theorem 10. Let $R \geq 0$ be constant. Then, for N sufficiently large,

$$\mu < \beta(\mu N, N, R) \leq 2\mu.$$

Proof. By Theorem 5 we have $\beta(\mu N, N, R) \leq \beta(\mu N, N, 0)$, while the latter equals 2μ by Theorem 1 of Kijima [15] (see also statement (7) in [6]). So it remains to be shown that $\beta(\mu N, N, R) > \mu$, that is, $x^* > N^{-1}$, for N sufficiently large. To this end choose $N > 4$ so large that $N^{-1} \leq 2(1 - \cos \pi/(2R + 2))$. Then, for $0 < x \leq N^{-1}$, the roots of (39) are non-real, and $H_R(x)$ satisfies

$$H_R(x) = \frac{\sin R\phi}{\sin(R+1)\phi},$$

with $\cos\phi = 1 - \frac{1}{2}x > \cos\frac{\pi}{2(R+1)}$, that is, $0 \leq \phi < \frac{\pi}{2(R+1)}$. Hence, $H_R(x) < 1$ if $0 < x \leq N^{-1}$. The proof can be completed by arguments similar to those in the proof of Theorem 8. □

More detailed information on the case $\lambda \sim \mu N$ can be obtained if we assume

$$\lambda \equiv \lambda(N) = \mu N + 2b(\mu N)^d + \mathcal{O}(1) \quad \text{as} \quad N \to \infty, \tag{43}$$

where $0 \leq d < 1$. We discern four cases, in each of which we use the monotonicity of $\beta(\lambda, N, R)$ as a function of R (Theorem 5).
(i) If $b < 0$ then, by [6, Eq. (7)],

$$\beta(\lambda, N, R) \leq \beta(\lambda, N, 0) < 2\mu \quad \text{for } N \text{ sufficiently large.} \tag{44}$$

(ii) If $b = 0$, or $b > 0$ and $d < \frac{1}{2}$, or $0 < b < \sqrt{\mu}$ and $d = \frac{1}{2}$, then, by [6, Theorem 2],

$$\beta(\lambda, N, R) \leq \beta(\lambda, N, 0) < 5\mu \quad \text{for } N \text{ sufficiently large.} \tag{45}$$

(*iii*) If $b > 0$ and $d > \frac{1}{2}$, then, by (34),

$$\beta(\lambda, N, R) > \beta(\lambda, N, \infty) = b^2(\mu N)^{2d-1} + o(N^{2d-1}) \to \infty \text{ as } N \to \infty. \quad (46)$$

(*iv*) If $b \geq \sqrt{\mu}$ and $d = \frac{1}{2}$, then, by [6, Theorem 2] again,

$$\beta(\lambda, N, R) > \beta(\lambda, N, \infty) = b^2 + o(1) \text{ as } N \to \infty. \quad (47)$$

The case $d = \frac{1}{2}$ in (43) is particularly interesting since it corresponds precisely to the setting in which $(\sqrt{\lambda} - \sqrt{\mu N})^2$ – the value of $\beta(\lambda, N, \infty)$ if $\rho > \rho_N^*$ – remains bounded as $N \to \infty$. It is not known whether $\beta(\lambda, N, R)$ remains bounded as $N \to \infty$ in case (*iv*).

Remark. It is shown in [6] that if $d = \frac{1}{2}$ and $b > \sqrt{\mu}$ then

$$\beta(\lambda, N, 0) > b^2 + \frac{3}{2}\mu + \frac{1}{2}\sqrt[3]{\mu^2 b^2}.$$

(There is an error in [6, (33)]: a^2 should be replaced (twice) by $a^2\mu$.) □

Acknowledgement

The author thanks an anonymous referee for pointing out an omission in the original manuscript.

References

[1] Chen, M.F. (1996) Estimation of spectral gap for Markov chains. *Acta Math. Sinica (N.S.)* 12: 337–360.

[2] Chen, M.F. (1998) Estimate of exponential convergence rate in total variation by spectral gap. *Acta Math. Sinica (N.S.)* 14: 9–16.

[3] Chihara, T.S. (1978) *An Introduction to Orthogonal Polynomials.* Gordon and Breach, New York.

[4] van Doorn, E.A. (1981) *Stochastic Monotonicity and Queueing Applications of Birth-Death Processes.* Lecture Notes in Statistics 4, Springer-Verlag, New York.

[5] van Doorn, E.A. (1985) Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process. *Adv. Appl. Probab.* 17: 514–530.

[6] van Doorn, E.A., and Zeifman, A.I. (2009) On the speed of convergence to stationarity of the Erlang loss system. *Queueing Syst.* 63: 241–252.

[7] van Doorn, E.A., Zeifman, A.I., and Panfilova, T.L. (2010) Bounds and asymptotics for the rate of convergence of birth-death processes. *Theory Probab. Appl.* 54: 97–113.

[8] Gamarnik, G., and Goldberg, D. (2010) On the rate of convergence to stationarity of the $M/M/N$ queue in the Halfin-Whitt regime. Preprint (`arXiv:1003.2004]`).

[9] Granovsky, B.L., and Zeifman, A.I. (1997) The decay function of nonhomogeneous birth-death processes, with application to mean-field models. *Stochastic Processes Appl.* 72: 105–120.

[10] Ismail, M.E.H., and Muldoon, M.E. (1991) A discrete approach to monotonicity of zeros of orthogonal polynomials. *Trans. Amer. Math. Soc.* 323: 65–78.

[11] Jagerman, D.L. (1974) Some properties of the Erlang loss function. *Bell System Tech. J.* 53: 525–551.

[12] Karlin, S., and McGregor, J.L. (1958) Many server queueing processes with Poisson input and exponential service times. *Pacific J. Math.* 8: 87–118.

[13] Karlin, S., and McGregor, J.L. (1965) Ehrenfest urn models. *J. Appl. Probab.* 2: 352–376.

[14] Keilson, J., and Ramaswamy, R. (1987) The relaxation time for truncated birth-death processes. *Probab. Engrg. Inform. Sci.* 1: 367–381.

[15] Kijima, M. (1990) On the largest negative eigenvalue of the infinitesimal generator associated with $M/M/n/n$ queues. *Oper. Res. Lett.* 9: 59–64.

[16] Kijima, M. (1992) Evaluation of the decay parameter for some specialized birth-death processes. *J. Appl. Probab.* 29: 781–791.

[17] Kijima, M. (1997) *Markov Processes for Stochastic Modelling.* Chapman & Hall, London.

[18] Takács, L. (1962) *Introduction to the Theory of Queues.* Oxford University Press, New York.

[19] Zeifman, A.I. (1995) Upper and lower bounds on the rate of convergence for non-homogeneous birth and death processes. *Stochastic Process. Appl.* 59: 157–173.

[20] Zeifman, A.I., Bening, V.E. and Sokolov, I.A. (2008) *Markov Chains and Models in Continuous Time.* Elex-KM, Moscow (in Russian).