An M|Ph|1 queueing system subject to breakdowns with non-homogeneous interarrival, service, life and repair times
Eenige, van, M.J.A.; van der Wal, J.

Published: 01/01/1994

Document Version
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author’s version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 15. Dec. 2018
Memorandum COSOR 94-13

An $M|Ph|1$ Queueing System Subject to Breakdowns with Non-Homegeneous Interarrival, Service, Life and Repair Times

M.J.A. van Eenige
J. van der Wal

Eindhoven, March 1994
The Netherlands
An M|Ph|1 Queueing System Subject to Breakdowns with Non-Homogeneous Interarrival, Service, Life and Repair Times

M.J.A. van Eenige and J. van der Wal
Eindhoven University of Technology
Department of Mathematics and Computing Science
P.O. Box 513
5600 MB Eindhoven
The Netherlands

Abstract

In this paper, an M|Ph|1 queueing system with server breakdowns is studied for both the finite and infinite buffer case. Jobs arrive at the server according to a non-homogeneous Poisson process with a non-homogeneous service time distribution which is of phase type. The life-times of the server are non-homogeneously exponentially distributed and the repair times of the server are non-homogeneously phase type distributed. The non-homogeneity of these distributions is due to the dependence of their parameters on the number of jobs in the system as well as the interdependence of these parameters. For these models, an explicit recursive matrix representation of the stationary queue length distribution is obtained.
1 Introduction

In this paper, we will consider a single server queue with server breakdowns. At this server, jobs arrive according to a non-homogeneous Poisson process with a non-homogeneous service time distribution which is of phase type (as defined in Neuts [8]). The life-times of the server are non-homogeneously exponentially distributed. When a server breakdown occurs, the server will be repaired and in the mean time, all jobs wait. The repair times of the server have a non-homogeneous phase type distribution. The non-homogeneity of these distributions will be due to the dependence of their parameters on the number of jobs in the system and to the interdependence of these parameters. More precisely, the arrival rate depends on the number of jobs at the server, on the phase of the service time and on the phase of the repair time. The service rates depend on the number of jobs in the system. Finally, both the failure rates and repair rates of the server depend on the number of jobs in the system and on the phase of the service time.

Think of the production to order of some part or product which requires a number of different operations such as sawing, boring, milling and polishing. Due to wear and breakage, every now and then a tool has to be repaired or replaced. In another time scale one may think of projects. It may happen that the execution of a project is delayed because some specialist or special equipment is not available or delayed by weather conditions. If the backlog grows, newly arriving projects may have to be rejected or put out to contract. It may be possible to speed up the production process for instance by using more expensive tools or by hiring additional capacity.

Queueing systems with non-stationary arrival and service rates have been studied by numerous authors. For example, Neuts (cf. [6] and [7]), Van Hoorn and Seelen (cf. [4]), Sengupta (cf. [11]), Bhat (cf. [1]), and Van Eenige, Resing and Van der Wal (cf. [2]) discuss queueing systems with arrival and/or service rates depending on the state of a Markovian environment (see Neuts [8] for a definition). This Markovian environment represents, for instance, periodic or aperiodic up and down periods of the server, priority rules and rush-hour behaviour. Further, Schellhaas (cf. [10]), Shanthikumar and Sumita (cf. [12]), and Kijima and Makimoto (cf. [5]), for instance, study queueing systems with arrival and/or service rates depending on the number of jobs in the system. This paper combines these two kinds of queueing systems and allows more general dependences of the arrival and service rates of jobs with failure and repair rates of the server depending on the state of the system.

The goal of this paper is to analyse the queue length process for these queueing systems with a finite as well as an infinite buffer. By generalizing the result of Ramaswami and Latouche (cf. [9]), we can give an explicit recursive matrix representation of the stationary queue length distribution.

The outline of the paper is as follows. In Section 2, we will describe the queueing system in detail. The queue length process for both the finite and infinite buffer case will be analysed in Section 3. Finally, in Section 4 we will give a summary and some final remarks.

2 The Model

Consider a single server queue with server breakdowns. At this server, jobs arrive according to a non-homogeneous Poisson process (the rate of this arrival process will be introduced later in this section).
The service of a job consists of a (random) sequence of operations. These operations belong to a finite set of operations of size $M$ and the operations of this set are numbered $1, 2, \ldots, M$. For each job, the sequence of operations is generated by a finite Markov chain with $M+1$ states and non-stationary transition probabilities. The states $1, 2, \ldots, M$ are transient and state $M+1$ is absorbing. When the server is up, the number of jobs in the system is $l$ and the chain is in state $m$, $m = 1, 2, \ldots, M$, the server is executing operation $m$ at the exponential rate $\mu(m) > 0$. If the server is down, the execution rate of each operation is obviously zero. When the execution of operation $m$ is completed, the Markov chain makes a transition to state $m'$ with probability $p_l^m(m, m') = 0$ (with $\sum_{m'}=1 p_l^m(m, m') = 1$), where $l$ denotes the number of jobs in the system just before the completion of the execution. Furthermore, when during the execution of operation $m$ a job arrives, the Markov chain makes a transition to state $m'$ with probability $\tilde{p}_l^m(m, m') \geq 0^1$ (with $\sum_{m'}=1 \tilde{p}_l^m(m, m') = 1$), where $l$ denotes the number of jobs in the system just after the arrival. If $m' = 1, 2, \ldots, M$, the server starts the execution of operation $m'$. If $m' = M+1$, the service of the job is completed (and this job leaves the system) and the service of the next job starts with operation $m$, $m = 1, 2, \ldots, M$, with probability $\alpha_l(m) \geq 0$ (with $\sum_{m}=1 \alpha_l(m) = 1$), where $l$ denotes the (remaining) number of jobs in the system.

The failure rate of the server depends on the number of jobs $l$ in the system and on the operation $m$ which is being executed. This rate is denoted by $\theta_{l,m}$ if $l \geq 1$ and by $\theta_{0,0}$ if the system is empty. Thus, $\theta_{0,0} > 0$ implies that a server breakdown may occur while the system is empty.

When a server breakdown occurs, the service of jobs is stopped immediately and the repair of the server is started. Like the service of a job, a repair of the server consists of a (random) sequence of operations. These operations belong to a finite set of size $N$ and are numbered $1, 2, \ldots, N$. A finite Markov chain with $N+1$ states and non-stationary transition probabilities generates the sequence of repair operations. The states $1, 2, \ldots, N$ are transient and state $N+1$ is absorbing. When the number of jobs in the system is $l$, the phase of the service upon failure was $m$ and the chain is in state $n$, the repair facility is executing operation $n$, $n = 1, 2, \ldots, N$, at the rate $\sigma_{l,m}(n) > 0$. When the execution of operation $n$ is completed or when a job arrives during the execution of operation $n$, the next repair phase will be $n'$ with probability $p_{l,m}^n(n, n') \geq 0$ and $\tilde{p}_{l,m}^n(n, n') \geq 0^1$, respectively, (with $\sum_{n'}=1 p_{l,m}^n(n, n') = \sum_{n'}=1 \tilde{p}_{l,m}^n(n, n') = 1$), where $l$ denotes the number of jobs in the system just before the completion of operation $n$ and just after the arrival, respectively. If $n' = N+1$, the repair is completed. Otherwise, the repair facility starts with the execution of the repair operation $n'$. The repair of the server is started with operation $n$, $n = 1, 2, \ldots, N$, with probability $\beta_{l,m}(n) \geq 0$ (with $\sum_{n}=1 \beta_{l,m}(n) = 1$), where $l$ and $m$ are as before. Finally, when the system is empty, the initial probabilities, the execution rates and the two kinds of transition probabilities are denoted by $\beta_{0,0}(n), \sigma_{0,0}(n)$ and, $p_{0,0}^n(n, n')$ and $\tilde{p}_{0,0}^n(n, n')$, respectively.

The arrival rate depends on the number of jobs $l$ in the system and on the operation $m$ of the service which is being executed. Moreover, when the server is down, this rate also depends on the operation $n$ of the repair which is being executed. We denote the arrival rate by $\lambda_{l,m,0}$ if the server is up and by $\lambda_{l,m,n}$ if the server is down. When the system is empty, these rates are

---

1 In most cases, one will have $p_l^m(k, k') = 1$ and $\tilde{p}_l^m(k, k') = 1$ for $k' = k$ and zero otherwise. However, in order to describe, for instance, the simultaneous execution of several service operations (repair operations) when the number of jobs in the system exceeds a certain number, one can have $p_l^m(k, k') > 0$ (or $\tilde{p}_l^m(k, k') > 0$) for some $k' \neq k$, where $k'$ denotes a newly defined service operation (repair operation) representing the simultaneous execution of certain operations.
denoted by $\lambda_{0,0,o}$ and $\lambda_{0,0,n}$, respectively.

Jobs are served in the order of their arrival. When a server breakdown occurs during the
describe of a job, the service of the interrupted job is resumed after the repair with operation $m'$
with probability $\hat{p}_l^t(m, m')$, where $l$ denotes the number of jobs in the system (just before the
completion of the repair) and $m$ the service operation which was being executed upon failure.

We now summarize the parameters introduced above. Let $l$ denote the number of jobs in the
system, $m$ the phase of the service which is being executed (was being executed upon failure)
and $n$ the phase of the repair which is being executed. Then, the parameters are defined as

$$
\begin{align*}
\lambda_{l,m,n} &= \text{the arrival rate of jobs}, \\
\mu_l(m) &= \text{the execution rate of the service phase } m, \\
p_l^s(m, m') &= \text{the transition probability from service phase } m \text{ to service phase } m', \\
\hat{p}_l^s(m, m') &= \text{the transition probability from service phase } m \text{ to service phase } m', \\
\sigma_l(m) &= \text{the probability that a job arrives during the execution of phase } m, \\
\theta_{l,m} &= \text{the failure rate of the server}, \\
\sigma_{l,m}(n) &= \text{the execution rate of the repair phase } n, \\
p_{l,m}^r(n, n') &= \text{the transition probability from repair phase } n \text{ to repair phase } n', \\
\hat{p}_{l,m}^r(n, n') &= \text{the transition probability from repair phase } n \text{ to repair phase } n', \\
\beta_{l,m}(n) &= \text{the probability that the repair starts with phase } n,
\end{align*}
$$

where $m$ and $n$ are set zero when the system is empty and the server is up, respectively.

Finally, we make the following natural assumption

\textbf{Assumption 1} For each $m$, $m'$, $n$ and $n'$ ($m, m' = 1, 2, \ldots, M$, $n = 0, 1, 2, \ldots, N$ and $n' = 1, 2, \ldots, N$) the following limits exist

$$
\begin{align*}
\lim_{l \to \infty} \lambda_{l,m,n} &= \lambda_{m,n}, \\
\lim_{l \to \infty} \mu_l(m) &= \mu(m), \\
\lim_{l \to \infty} p_l^s(m, m') &= p^s(m, m'), \\
\lim_{l \to \infty} \hat{p}_l^s(m, m') &= \hat{p}^s(m, m'), \\
\lim_{l \to \infty} \sigma_{l,m}(n) &= \sigma_m(n), \\
\lim_{l \to \infty} \beta_{l,m}(n) &= \beta_m(n), \\
\lim_{l \to \infty} p_{l,m}^r(n, n') &= p_m^r(n, n') \quad n \geq 0, \\
\lim_{l \to \infty} \hat{p}_{l,m}^r(n, n') &= \hat{p}_m^r(n, n') \quad n \geq 0.
\end{align*}
$$

3 The Queue Length Process

In this section, we will analyse the queue length process for the finite as well as the infinite buffer
case. Since the former can be analysed along the same lines, we will give a detailed analysis of
the latter and briefly discuss the former.
3.1 The Infinite Buffer Case

Let \( (X_t, Y_t, Z_t) \) denote the number of jobs in the system, the phase of the service and the phase of the repair, respectively, at time \( t \). When the system is empty we set \( Y_t := 0 \) and when the server is up we set \( Z_t := 0 \). Then, the process \( \{(X_t, Y_t, Z_t)\}_{t \geq 0} \) is a continuous time irreducible Markov chain with state space

\[
S = \{(0, 0, n) | n = 0, 1, \ldots, N \} \cup \{(l, m, n) | l = 1, 2, \ldots, M, n = 0, 1, \ldots, N \}.
\]

The state space of this Markov chain is ordered lexicographically. Moreover, we define for each \( l, l = 0, 1, 2 \ldots \), level \( l \) as the set \( S_l \) of all states for which the number of jobs in the system is equal to \( l \), i.e.,

\[
S_0 = \{(0, 0, n) | n = 0, 1, \ldots, N \}
\]

and for \( l = 1, 2, \ldots \)

\[
S_l = \{(l, m, n) | m = 1, 2, \ldots, M, n = 0, 1, \ldots, N \}.
\]

Note that we have here implicitly assumed that for each \( l \) the number of service phases as well as the number of repair phases is fixed and independent of \( l \) (by some trivial adjustments, the results of this subsection can be extended to the case for which the number of service phases and the number of repair phases depend on \( l \)). When partitioning the state space into these levels, the infinitesimal generator \( Q \) of the Markov chain can be written into the following block tri-diagonal form

\[
Q = \begin{pmatrix}
B_0 & B_1 & 0 & 0 & 0 & \cdots \\
B_2 & A_{1,1} & A_{0,1} & 0 & 0 & \cdots \\
0 & A_{2,2} & A_{1,2} & A_{0,2} & 0 & \cdots \\
0 & 0 & A_{2,3} & A_{1,3} & A_{0,3} & \cdots \\
& & & \ddots & \ddots & \ddots \\
& & & & \ddots & \ddots & \ddots
\end{pmatrix}
\]

For all \( l \), the \( M(N+1) \times (N+1) \) matrices \( A_{2,l} \), \( A_{1,l} \) and \( A_{0,l} \) represent transitions from level \( l \) to level \( l-1 \), to itself and to level \( l+1 \), respectively. The \( (N+1) \times (N+1) \) matrix \( B_0 \), the \( (N+1) \times M(N+1) \) matrix \( B_1 \) and the \( M(N+1) \times (N+1) \) matrix \( B_2 \) represent transitions from level 0 to itself, from level 0 to level 1 and from level 1 to level 0, respectively.

Since the elements of the submatrices of the infinitesimal generator \( Q \) can be easily obtained, we do not give them all explicitly. However, since we will use the special structure of the matrices \( A_{2,l} \), \( l = 2, 3, \ldots \), we give these matrices here explicitly. Define for \( m, m' = 1, 2, \ldots, M \) the \( (N+1) \times (N+1) \) matrices \( T_{m,m'}^l \) by

\[
T_{m,m'}^l = \begin{pmatrix}
0 \quad 0 \quad 0 \quad \cdots \\
\alpha_{l-1}(m') \mu_l(m) p_l^i(m, M+1) \quad 0 \quad \cdots \\
0 \quad 0 \quad \cdots \\
\vdots \quad \vdots \quad \vdots \\
0 \quad 0 \quad \cdots 
\end{pmatrix},
\]

Then, the matrix \( A_{2,l} \) is given by

\[
A_{2,l} = \begin{pmatrix}
T_{1,1}^l & T_{1,2}^l & \cdots & T_{1,M}^l \\
T_{2,1}^l & T_{2,2}^l & \cdots & T_{2,M}^l \\
\vdots & \vdots & \ddots & \vdots \\
T_{M,1}^l & T_{M,2}^l & \cdots & T_{M,M}^l
\end{pmatrix}, \quad (1)
\]
for all $l$.\footnote{Note that we do not allow for the occurrence of several events at the same time epoch. By some straightforward adjustments, however, the analysis can be extended to the case of multiple events occurring at the same time epoch. For example, removing a finished job from a machine may cause a failure. Hence, the start of the service of the next job (or the departure of the job) and a machine breakdown occur at the same time epoch.}

By Assumption 1, the following limits exist
\begin{equation}
\lim_{l \to \infty} A_{0,l} =: A_0, \quad \lim_{l \to \infty} A_{1,l} =: A_1 \quad \text{and} \quad \lim_{l \to \infty} A_{2,l} =: A_2.
\end{equation}

Let the row vector $y$ denote the solution of
\[ y(A_0 + A_1 + A_2) = 0 \quad \text{and} \quad y \cdot e = 1, \]
where $e$ denotes the column vector with all its elements equal to one (for the existence of this solution we refer to Neuts [8]). Then, we make the following assumption

**Assumption 2** $yA_0e < yA_2e$.

Note that this assumption is the drift condition of Neuts (cf. [8]) of a Markov chain with block tri-diagonal infinitesimal generator with the submatrices $A_0$, $A_1$ and $A_2$ just above, on and just below the diagonal, respectively. For our model, this assumption implies that from a certain level onwards the total rate to the left (i.e., to a lower level) exceeds the total rate to the right (i.e., to a higher level).

We will construct a non-normalized solution of the balance equations of the Markov chain and show that under Assumption 1 (i.e., the limits (2) exist) and Assumption 2 this solution is bounded. Consequently, by a Foster's criterion (see Foster [3]) the stationary distribution of this chain is equal to this solution after normalizing it in the usual way.

Let the row vector $x$ denote the non-normalized solution of the balance equations, i.e., $x$ satisfies
\[ x \cdot Q = 0. \]

Moreover for all $l$, let $x_l$ denote the row vector consisting of the non-normalized solutions of the balance equations which correspond to the states at level $l$. Then, the balance equations (3) read
\begin{align}
x_0 \cdot B_0 + x_1 \cdot B_2 &= 0, \\
x_0 \cdot B_1 + x_1 \cdot A_{1,1} + x_2 \cdot A_{2,2} &= 0, \\
x_{l-1} \cdot A_{0,l-1} + x_l \cdot A_{1,l} + x_{l+1} \cdot A_{2,l+1} &= 0, \quad l = 2, 3, \ldots
\end{align}

It is useful to notice that (1) implies that the matrix $A_{2,l}$ can be written as
\begin{equation}
A_{2,l} = \mu_l \cdot \alpha_{l-1}, \quad l = 2, 3, \ldots
\end{equation}

(compare Ramaswami and Latouche [9]), where $\mu_l$ denotes the $M(N + 1) \times 1$ vector
\[
\mu_l = (\mu_l(1)p_l^1(1, M + 1), 0, \ldots, 0, \mu_l(2)p_l^2(2, M + 1), 0, \ldots, 0, \ldots, \mu_l(M)p_l^M(M, M + 1), 0, \ldots, 0)^T
\]
and \( \alpha_{l-1} \) the \( 1 \times M(N + 1) \) vector

\[
\alpha_{l-1} = (\alpha_{l-1}(1), 0, \ldots, 0, \alpha_{l-1}(2), 0, \ldots, 0, \ldots, \alpha_{l-1}(M), 0, \ldots, 0).
\]

From the special structure of the matrices \( A_{2,l} \), we can make the following important observation. Given a transition from a state at level \( l \geq 2 \) to a state at level \( l - 1 \) (i.e., a service completion), the probability distribution of the state at level \( l - 1 \) is equal to \( \alpha_{l-1} \). Further, since transitions from a level to itself or to its neighbouring levels are the only ones possible, the solution \( \mathbf{x} \) satisfies

\[
x_{l-1}A_{0,l-1}e = x_lA_{2,l}e, \quad l = 2, 3, \ldots.
\]

Then, from (7) and

\[
x_{l-1}A_{0,l-1}e = x_lA_{2,l}e = x_l\mu_l\alpha_{l-1}e = x_l\mu_l,
\]

we have

\[
x_lA_{2,l} = x_l\mu_l\alpha_{l-1} = x_{l-1}A_{0,l-1}e\alpha_{l-1}, \quad l = 2, 3, \ldots \tag{8}
\]

**Lemma 1** The non-normalized solution of the balance matrix equations (5) and (6) is given by

\[
x_l = x_{l-1}S_l = x_0\prod_{j=1}^{l-1}S_j, \quad l = 1, 2, \ldots,
\]

with \( S_l = A_{0,l-1}(-\{A_{1,l} + A_{0,l}e\alpha_l\}^{-1}) \), where \( A_{0,0} := B_1 \).

**Proof.** Substituting (8) into the matrix equations (5) and (6) yields

\[
x_0B_1 + x_1(A_{1,1} + A_{0,1}e\alpha_1) = 0 \tag{9}
\]

and

\[
x_{l-1}A_{0,l-1} + x_l(A_{1,l} + A_{0,l}e\alpha_l) = 0, \quad l = 2, 3, \ldots,
\]

respectively.

Since for all \( l \) the matrix \( A_{1,l} + A_{0,l}e\alpha_l \) can be regarded as a transient infinitesimal generator it is non-singular (see, e.g., Lemma 2.2.1 in Neuts [8]) and hence, the proof is complete. \( \square \)

The following lemma states that (under Assumption 1 and 2) this solution is bounded.

**Lemma 2** Under Assumption 1 (i.e., the limits (2) exist) and Assumption 2, the solution given in Lemma 1 of the balance equations (5) and (6) is bounded (i.e., \( \Sigma_{l=1}^{\infty} x_l e < \infty \)).

**Proof.** See the Appendix. \( \square \)

By applying a Foster's criterion (see Foster [3]), we conclude from Lemma 2 that the Markov chain contains a stationary distribution which we will denote by \( \pi \). Furthermore, for \( l = 1, 2, \ldots \), the row vector \( \pi_l \) containing the stationary probabilities of the states at level \( l \) is given by Lemma 1, and the vector consisting of the stationary probabilities of the states at level 0 (i.e., \( \pi_0 \)) is obtained by solving the boundary condition (4) and the normalization (i.e., \( \pi e = 1 \)). These results are summarized in the following theorem
Theorem 1 Under Assumption 1 and Assumption 2, the Markov chain with infinitesimal generator $Q$ contains a stationary distribution $\pi$, which is given by

$$\pi_l = \pi_0 \prod_{j=1}^l A_{0,j-1}(- (A_{1,j} + A_{0,j} e \alpha_j)^{-1}), \quad l = 1, 2, \ldots,$$

where the vector $\pi_0$ is obtained by substituting (9) (after solving for $\pi_1$) into the balance equation (4), and solving this rewritten balance equation and the normalization

$$\pi_0 (I + \sum_{l=1}^{\infty} \prod_{j=1}^l A_{0,j-1}(- (A_{1,j} + A_{0,j} e \alpha_j)^{-1})) e = 1,$$

with $A_{0,0} := B_1$.

3.2 The Finite Buffer Case

The analysis of the queue length process for the finite buffer case is quite similar to the one for the infinite buffer case. Unless stated otherwise, we will use the same notations and definitions as in the previous subsection.

The size of the buffer is equal to $L$ (including a possible job in service). Now, consider the process $\{(X_t, Y_t, Z_t)\}_{t \geq 0}$. Then, this process is a continuous time Markov chain with finite state space

$$S^F = \{(0,0,n)|n=0,1,\ldots,N\} \cup \{(l,m,n)|l=1,\ldots,L, m=1,\ldots,M, n=0,1,\ldots,N\}.$$ 

Since this chain is finite and irreducible, it is positive recurrent and hence, it contains a stationary distribution, which will be denoted by the (finite) row vector $\pi$.

When partitioning the state space into the sets $S_l$ (see Section 3.1), the infinitesimal generator $Q^F$ of the Markov chain has the following block tri-diagonal form

$$Q^F = \begin{pmatrix} B_0 & B_1 & 0 & 0 & 0 & \ldots & 0 \\ B_2 & A_{1,1} & A_{0,1} & 0 & 0 & \ldots & 0 \\ 0 & A_{2,2} & A_{1,2} & A_{0,2} & 0 & \ldots & 0 \\ 0 & 0 & A_{2,3} & A_{1,3} & A_{0,3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & A_{2,L-1} & A_{1,L-1} & A_{0,L-1} \\ 0 & \ldots & 0 & A_{2,L} & B_L \end{pmatrix},$$

where the $M(N+1) \times M(N+1)$ matrix $B_L$ represents transitions from level $L$ to itself.

When defining, for all $l$, $\pi_l$ as the row vector containing the stationary probabilities for the states at level $l$, the balance equations of the Markov chain are

$$\pi_0 \cdot B_0 + \pi_1 \cdot B_2 = 0,$$

$$\pi_0 \cdot B_1 + \pi_1 \cdot A_{1,1} + \pi_2 \cdot A_{2,2} = 0,$$

$$\pi_{l-1} \cdot A_{0,l-1} + \pi_l \cdot A_{1,l} + \pi_{l+1} \cdot A_{2,l+1} = 0, \quad l = 2, 3, \ldots, L - 1,$$

$$\pi_{L-1} \cdot A_{0,L-1} + \pi_L \cdot B_L = 0.$$
By a similar argument as for the infinite buffer case, relation (8) holds for \( l = 2, 3, \ldots, L \) and hence from Lemma 1 we have that

\[
\pi_l = \pi_{l-1} A_{0,l-1} \left( -\left( A_{1,l} + A_{0,l} e \alpha_l \right)^{-1} \right), \quad l = 1, 2, \ldots, L - 1,
\]

is a solution of the balance equations (11) and (12). Further, since the matrix \( B_L \) can be regarded as a transient infinitesimal generator it is non-singular (see, e.g., Lemma 2.2.1 in Neuts [8]). Thus, from the matrix equation (13), we have

\[
\pi_L = \pi_{L-1} A_{0,L-1} (-B_L)^{-1}.
\]

Finally, the vector \( \pi_0 \) is obtained by solving the boundary condition (10) and using the normalization. These results are summarized in following theorem

**Theorem 2** The stationary distribution of the Markov chain with infinitesimal generator \( Q^F \) is given by

\[
\pi_l = \pi_0 \prod_{j=1}^{L} A_{0,j-1} \left( -\left( A_{1,j} + A_{0,j} e \alpha_j \right)^{-1} \right), \quad l = 1, 2, \ldots, L,
\]

where the vector \( \pi_0 \) is obtained by substituting (14) for \( l = 1 \) into the balance equation (10), and solving this rewritten balance equation and the normalization

\[
\pi_0 (I + \sum_{i=1}^{L} \prod_{j=1}^{i} A_{0,j-1} \left( -\left( A_{1,j} + A_{0,j} e \alpha_j \right)^{-1} \right)) e = 1.
\]

with \( A_{0,0} := B_1 \) and \( A_{1,L} + A_{0,L} e \alpha_L := B_L \).

### 4 Conclusions

We have considered a single server queue with server breakdowns for both the finite and infinite case. At the server, jobs arrive according to a non-homogeneous Poisson process with a non-homogeneous service time distribution which is of phase type. The life-times and repair times of the server are non-homogeneously exponentially and non-homogeneously phase type distributed, respectively. The special one-level backwards structure (i.e., property (7)) implies that the distribution of the state at level \( l-1 \) which is reached by a one step transition from a state at level \( l \) is independent of this latter state (but may depend on \( l \)). By using this property, we have obtained an explicit recursive matrix representation of the stationary queue length distribution for both queueing systems. For our model, as can be easily verified, the key condition for this property to hold is the (non-homogeneous) exponential distribution of the life-times of the server. In other words, when generalizing the life-time distribution to, for instance, Erlang or phase type distributions, this property is lost and hence the recursive expression for the stationary queue length distribution.

In principle, one can obtain performance measures like the average number of jobs served per unit time (i.e., the throughput) and the mean sojourn time or waiting time (by using Little’s Law) from the stationary queue length distribution. In applications, however, exact results seem only possible for the finite buffer case and for the case that from a certain level onwards the transition rates of the Markov chain are independent of the level. In all other cases, one has to use approximations, e.g., by reducing the infinite buffer case to the finite buffer case with \( L \) sufficiently large or by assuming that from a certain level onwards the transition rates
are independent of the level. Further, since the sojourn time (waiting time) of a job depends on arrivals during its service time (waiting time), the determination of the distribution of the sojourn time (waiting time) is complicated.
Appendix: Proof of Lemma 2

In this Appendix, we will prove Lemma 2 by showing that $yA_0e < yA_2e$ implies that the matrix $\sum_{i=1}^{\infty} \prod_{j} S_j$ is bounded and consequently, that the solution given in Lemma 1 is bounded.

Consider the irreducible Markov chain with the partitioned state space $S = \bigcup_{i=0}^{\infty} S_i$ (see Section 3.1) and infinitesimal generator

$$
\begin{pmatrix}
C_0 & C_1 & 0 & 0 & 0 & \ldots \\
C_2 & A_1 & A_0 & 0 & 0 & \ldots \\
0 & A_2 & A_1 & A_0 & 0 & \ldots \\
0 & 0 & A_2 & A_1 & A_0 & \ldots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{pmatrix},
$$

where the matrices $A_0$, $A_1$ and $A_2$ are as defined in (2), and the matrices $C_0$, $C_1$ and $C_2$ are (the limits of) the matrices $B_0$, $B_1$ and $B_2$, respectively.

From Theorem 1.7.1 in Neuts [8], the assumption

$$
yA_0e < yA_2e
$$

implies that this chain is positive recurrent. Moreover by the same theorem, there exists a matrix $R$ which is the minimal nonnegative solution of the quadratic matrix equation

$$
A_0 + RA_1 + R^2 A_2 = 0,
$$

with spectral radius, denoted by $\rho(R)$, strictly less than one.

Using the special structure of the matrix $A_2$ (cf. (7), i.e., $A_2 = \mu \alpha := \lim_{t \to -\infty} \mu_t \alpha_{t-1}$), we have by Theorem 3 of Ramaswami and Latouche (cf. [9]) and Lemma 1

$$
R = A_0(-A_1 + A_0 e \alpha)^{-1} = \lim_{t \to -\infty} A_{0,t-1}(-A_{1,t} + A_{0,t} e \alpha_t)^{-1} = \lim_{t \to -\infty} S_t.
$$

Let $E$ denote the matrix of appropriate size with all its elements equal to one. Moreover, define the matrix inequality $A < B$ as $a_{i,j} < b_{i,j}$ for each pair $(i,j)$ where $a_{i,j}$ and $b_{i,j}$ denote the entry $(i,j)$ of the matrix $A$ and $B$, respectively. Since the spectral radius is a continuous function in the elements of a matrix, there exists a $\delta > 0$ such that

$$
\rho(R) < \rho(R + \delta E) < 1.
$$

Since

$$
\lim_{t \to -\infty} S_t = R,
$$

there exists a positive integer $L(\delta)$ such that for all $l \geq L(\delta)$

$$
S_l < R + \delta E.
$$
Thus,

$$
\sum_{l=1}^{\infty} \prod_{j=1}^{l} S_j = \sum_{l=1}^{L(\delta)-1} \prod_{j=1}^{l} S_j + \sum_{l=L(\delta)}^{\infty} \prod_{j=1}^{l} S_j
\leq \sum_{l=1}^{L(\delta)-1} \prod_{j=1}^{l} S_j + \sum_{l=L(\delta)}^{\infty} \prod_{j=1}^{l} (R + \delta E)
\leq \sum_{l=1}^{L(\delta)-1} \prod_{j=1}^{l} S_j + (R + \delta E)^{L(\delta)} (I - (R + \delta E))^{-1} < \infty
$$

as the spectral radius of the matrix $R + \delta E$ is strictly less than one. So, $\sum_{l=1}^{\infty} \prod_{j=1}^{l} S_j$ is bounded and hence the solution given in Lemma 1. \qed
References


