G9090

UNIVERSITY LIBRARY
Cochin - 682 022
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

STOCHASTIC MODELLING AND ANALYSIS

QUEUES WITH RETRIAL/SELF-GENERATION OF PRIORITIES/POSTPONEMENT OF WORK AND SOME RELATED RELIABILITY PROBLEMS

THESIS SUBMITTED TO THE

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

UNDER THE FACULTY OF SCIENCE

BY

VISWANATH. C. NARAYANAN

DEPARTMENT OF MATHEMATICS

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

COCHIN - 682 022, INDIA

SEPTEMBER 2005

CERTIFICATE

This is to certify that the thesis entitled ***Queues with Retrial/Self-Generation of Priorities/Postponement of work and some related Reliability Problems*** *is a bona fide record of the research work carried out by* Mr. Viswanath C. Narayanan *under my supervision in the Department of Mathematics, Cochin University of Science and Technology. The results embodied in this thesis have not been included in any other thesis submitted previously for the award of any degree or diploma.*

September 24, 2005

Dr. A. Krishnamoorthy
(Supervising Guide)
Professor, Department of Mathematics
Cochin University of Science & Technology
Cochin-682 022

Contents

Chapter 1. **Introduction** 8

Chapter 2. **Idle time utilisation through service to customers in a retrial queue maintaining high system reliability** 27

2.1. The mathematical model 29

2.2. Stationary state distribution of the system 31

2.3. Specification of the embedded Markov chain 31

2.4. Stability condition 33

2.5. Stationary distribution of the embedded Markov chain 36

2.6. Stationary distribution of the system at arbitrary time 36

2.7. Performance characteristics 38

2.8. Particular case 39

2.9. System performance measures 49

2.10. Numerical illustration 51

Chapter 3. **Maximization of reliability of a k-out-of-n system with repair by a facility attending external customers in a retrial queue** 55

3.1. Modelling and analysis 57

3.2. System stability 64

3.3. Steady state distribution 66

3.4. Performance measures 67

3.5. Numerical illustration 70

Chapter 4. **Reliability of a k-out-of-n system with repair by a service station attending a queue with postponed work** 79

4.1. Mathematical modelling 80

4.2. Stability condition 84

4.3. Stationary distribution 87

4.4. A cost function and numerical illustrations 96

4.5. Comparison of Models in chapters 2, 3 and 4 102

Chapter 5. **On a queueing system with self generation of priorities** 104

5.1. Mathematical modelling and analysis 106

5.2. Ergodicity 108

5.3. Steady state distribution 109

5.4. Some particular cases 111

5.5. System performance measures 117

Chapter 6. **The impact of self-generation of priorities on multi-server queues with finite capacity** 121

6.1. The finite capacity MAP/PH,PH/$c/c + N$ queue with self-generation of priorities 122

6.2. Basic results of the system 124

6.3. Performance evaluation 130

6.4. Effect of the self-generation of priorities 138

Chapter 7. **Retrial queues with self generation of priority of orbital customers** 144

7.1. Mathematical modelling 144

7.2. Steady state distribution 148

7.3. System performance measures 150

7.4. Numerical illustration 151

Bibliography 156

CHAPTER 1

Introduction

A queue is formed when customers arriving at a service station are met with a busy server and decides to wait for receiving service. To model a queueing system mathematically, we require the arrival pattern, service time distribution, the number of servers, the capacity of the service station and the service discipline. These quantities varies according to the practical situation we want to model mathematically.

Applications of Queueing theory in areas like Computer networking, ATM facilities, Telecommunications and to many other numerous situations made people study Queueings models extensively and it has become an ever expanding branch of applied probability.

Methods for analysing queueing models : A queueing model is often analysed by using a continuous (or discrete) time Markov Chain whose description and analysis depends on the queueing model under consideration. For example, in the case of $M|M|1$ queue, the collection $\{N(t) : t \geq 0\}$ where $N(t)$ denotes the number of customers in the system at time t, is a continuous time Markov Chain whose analysis gives us informations about the queueing model such as the distribution of the number of customers in the system at arbitrary time t, its limiting distributions (when it exists) the waiting time distribution, busy period etc. Below we briefly sketch some of the methods applied for studying a queueing model and we do this by considering the simple $M|M|1$ queueing system.

Let λ, μ denote the arrival and service rates respectively and $N(t)$, the number of customers present in the system at time t. We also assume that $N(0) = i$. Let

$$P_n(t) = P\{N(t) = n\}.$$

Then, since $\{N(t) : t \geq 0\}$ is a Markov Process, we can write

$P_n(t+\Delta t) = P_n(t)(1-(\lambda+\mu)\Delta t) + P_{n-1}(t)\lambda\Delta t + P_{n+1}(t)\mu\Delta t + o(\Delta t)$ for $n \geq 1$ and

$P_0(t+\Delta t) = P_0(t)(1-\lambda\Delta t) + P_1(t)\mu\Delta t + o(\Delta t)$

By subtracting $P_n(t)$ from both sides, dividing throughout by Δt, and then taking limit as $\Delta t \to 0$, we get the **differential-difference** equations:

$$\frac{d}{dt}P_n(t) = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) \quad \text{for} \quad n \geq 1,$$

$$\text{and } \frac{d}{dt}P_0(t) = -\lambda P_0(t) + \mu P_1(t) \tag{1.1}$$

These equations are called the **forward Kolmogorov equations.**

To solve (1.1) the **method of generating functions** is used as follows:

We define $P(z,t) = \sum_{n=0}^{\infty} P_n(t)z^n$, ($z$ complex). Then using (1.1) we arrive at the equations

$$\frac{\partial}{\partial t}P(z,t) = \frac{1-z}{z}\{(\mu-\lambda z)P(z,t) - \mu P_0(t)\} \tag{1.2}$$

and

$$P(z,0) = z^i \tag{1.3}$$

where $\frac{\partial}{\partial t}P(z,t) = \sum_{n=0}^{\infty} p_n'(t)z^n$

Now to sole (1.2) we define the **Laplace transforms** with respect to time t of $P(z,t)$ and $P_i(t)$ as

$$\mathcal{L}\{P(z,t)\} = \bar{P}(z,s) = \int_0^{\infty} e^{-st}P(z,t)dt$$

$$\mathcal{L}\{P_i(t)\} = \bar{P}_i(s) = \int_0^{\infty} e^{-st}P_i(t)dt$$

and then from (1.2) we get

$$\bar{P}(z,s) = \frac{z^{i+1} - \mu(1-z)\bar{P}_0(s)}{(\lambda+\mu+s)z - \mu - \lambda z^2} \tag{1.4}$$

Evaluating $\bar{P}_0(s)$ we get the Laplace transform $\bar{P}(z,s)$ and then inverting it, we get $P(z,t)$. Now for finding the $P_n(t)$s we have to find the coefficient of z^n in the power series expansion of $P(z,t)$. But the inversion of the Laplace transform becomes almost impossible as the complexity of the queueing model increases which makes the above method unattractive from an application point of view.

From (1.1) we derive the stationary equations by putting $\frac{d}{dt}P_n(t) = 0$, as $t \to \infty$:

$$\begin{aligned} 0 &= -(\lambda+\mu)p_n + \lambda p_{n-1} + \mu p_{n+1} \quad (n \geq 1) \\ 0 &= -\lambda p_0 + \mu p_1 \end{aligned} \tag{1.5}$$

A solution $\{P_n\}$ to the above infinite system of equations which satisfies $\sum_{n=0}^{\infty} p_n = 1$ exists if, and only if, $\rho = \frac{\lambda}{\mu} < 1$. To find such a solution (when it exists) one can use the **iterative method** which gives

$p_1 = \rho p_0$

$p_n = \rho^n p_0$ for $n \geq 2$.

Now to find p_0 we use the relation $\sum_{n=0}^{\infty} p_n = 1$, which gives $p_0 = 1 - \rho$. Thus we get $p_n = (1-\rho)\rho^n$ for $n \geq 0$.

For finding p_ns we can also use the **method of generating functions** as follows.

We define

$P(z) = \sum_{n=0}^{\infty} p_n z^n$ (z complex)

then from (1.5) we have $P(z) = \frac{1-\rho}{1-z\rho}$ $(\rho < 1)$,

which implies

$$P(z) = \sum_{n=0}^{\infty} (1-\rho)\rho^n z^n$$

so that the coefficient p_n of z^n, is given by

$$p_n = (1 - \rho)\rho^n \text{ for } n \geq 0.$$

Here we note that each equation in (1.5) contains atmost three $p_n s$; which helped us to apply the above methods successfully. But as the number of p_ns which are interrelated through an equation increases (which often occurs when we use non exponential inter-arrival or service time distributions to model queueing problems) the direct application of the above methods becomes difficult and we seek the help of Matrix Analytic Methods. Before we discuss this method in some detail we shall mention some more methods applied by Queueing Theorists.

In the case of an $M|G|1$ queue where the service time distribution is arbitrary, one cannot get a Markov Chain by considering simply the random variable $N(t)$ which denotes the number of customers present in the system. Following are some methods applied in such a situation.

(a) Method of embedded Markov chain In this method we keep noting the value of the random variable $N(t)$ at certain epochs $\{t_n\}$ so that the collection $\{N(t_n)\}$ becomes a discrete time Markov Chain. For the $M|G|1$ queue, we achieve this by taking t_n as the epoch of n^{th} departure from the system and $N(t_n)$ as the number of customers left behind by the departing customer. Now the Markov Chain $\{N(t_n) : n \geq 1\}$ can be used to study the $M|G|1$ queueing system.

(b) Method of supplementary variables In this method to get a Markov Process, we keep track of some additional information together with the random variable $N(t)$. For $M|G|1$ queue the elapsed service time 'x' at time t of the unit undergoing service at time t serves as this additional information. In otherwords the collection $\{(N(t), x) : t \geq 0, x \geq 0\}$ is a Markov Process which can be used to study the $M|G|1$ queue.

Matrix analytic methods : Even though Queueing systems such as $M|M|1$, $M|M|\infty, G|G|1$ etc. are well studied and are well tractable, using the methods of generating functions and Laplace transform methods, the numerical tractability of Queueing systems through these methods becomes complicated when we assume non exponential interarrival or service time distributions which we mentioned in the above paragraphs. But the introduction of Matrix Analytic Methods in solving Queueing problems by Neuts and others, reduced this problem of numerical intractability considerably and increased the implementation of Queueings Models to analyse practical situations taking non exponential interarrival and service time distributions (for example Phase type) which are more suitable for practical applications. The modelling tools such as Phase type distributions, Markovian Arrival Processes, Batch Markovian Arrival Processes, Markovian Service Processes etc. are well suited for Matrix Analytic Methods.

Below we give a brief description of Matrix Analytic Methods applied for solving quasi-birth-and-death processes.

Level independent quasi-birth-and-death processes : A level independent quasi-birth-and-death process is a Markov process with state space E= $\{(0,j) : 1 \leq j \leq n\} \cup \{(i,j) : i \geq 1, 1 \leq j \leq m\}$ and with infinitesimal generator Q given by

$$Q = \begin{bmatrix} B_1 & B_0 & 0 & 0 & \cdots \\ B_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The generator Q is obtained in the above form by partitioning the state space E into the set of levels $\{\underline{0}, \underline{1}, \underline{2}, \ldots\}$ where $\underline{0} = \{(0,j) : 1 \leq j \leq n\}$, $\underline{i} = \{(i,j) : 1 \leq j \leq m\}$ for $i \geq 1$. The vector $\underline{i}$ is called i^{th} level. B_1 is a square matrix of order $n \times n$ and denotes transition rates from states of level 0 to the states of level 0 itself. B_0 is a matrix of order

$n \times m$ and denotes transition rates from level 0 to level 1. The $m \times n$ matrix B_2 denotes transition rates from level 1 to level 0. A_2, A_1, A_0 are square matrices of order m and denotes transition rates from level i to levels $i-1, i, i+1$ respectively. Assuming that Q is irreducible, we have the following theorem (see Neuts [**44**]).

THEOREM 1.1. *The process Q is positive recurrent if and only if, the minimal non negative solution R to the matrix quadratic equation*

$$R^2 A_2 + R A_1 + A_0 = 0 \tag{1.6}$$

has spectral radius less than 1 and the finite system of equations

$$x_0 B_1 + x_1 B_2 = 0, \qquad x_0 B_0 + x_1(A_1 + R A_2) = 0 \tag{1.7}$$

$$x_0 \mathrm{e} + x_1 (I - R)^{-1} \mathrm{e} = 1$$

has a unique positive solution for x_0, and, x_1.

If the matrix $A \doteq A_0 + A_1 + A_2$ is irreducible, then $sp(R) < 1$ if and only if, $\pi A_0 e < \pi A_2 e$, where π is the stationary probability vector of the generator matrix A.

The stationary probability vector $x = (x_0, x_1, x_2, \ldots)$ of Q is given by

$$x_i = x_1 R^{i-1} \text{ for } i \geq 1 \tag{1.8}$$

To find the minimal solution of (1.6) one can use the iterative formulas (see Neuts [**44**]):

$$R_n = -A_0 (A_1 + R_{n-1} A_2)^{-1} \text{ for } n \geq 1 \tag{1.9}$$

with an initial value R_0, which converges to R if $sp(R) < 1$. An accuracy check for R is given by the equation $R A_2 \mathrm{e} = A_0 \mathrm{e}$. Also the above relation (1.9) shows that if any row of A_0 is a row consisting of zeroes only, then the corresponding row of $R_n, n \geq 1$, has zeros only so that the corresponding row of R also consists of zeros only. So if our A_0 matrix has a special structure, it can be exploited in the evaluation of the R matrix.

Another method to find R is to use the relation

$$R = A_0(-A_1 - A_0G)^{-1} \tag{1.10}$$

where the matrix G is the minimal nonnegative solution of the matrix quadratic equation

$$A_2 + A_1G + A_0G^2 = 0 \tag{1.11}$$

The matrix G will be stochastic of $sp(R) < 1$. When $sp(R) < 1$, the **Logarithmic Reduction Algorithm** due to Ramaswamy (see Latouche and Ramaswamy [**41**]), which is quadratically convergent, can be used to calculate the G matrix and hence the R matrix using relation (1.10). When G is stochastic, from (1.11) we obtain the relation

$$G = (-A_1 - A_0G)^{-1}A_2 \tag{1.12}$$

which shows that if any column of the A_2 matrix is zero then the corresponding column of the G matrix is also zero. Therefore if the A_2 matrix has a special structure, it can be exploited in the calculation of the G matrix. Also one can efficiently use (Block) Gauss-Seidel iteration method to evaluate the G matrix, particularly if the matrix A_2 has a special structure.

For further details on Matrix Analytic Methods for Level independent QBD's we refer to Neuts [**44**], Latouche and Ramaswami [**41**].

Level dependent quasi-birth-and-death processes : A Level dependent quasi-birth-and-death process is a Markov process with state space $E = \{(i, j) : i \geq 0, 1 \leq j \leq n_i\}$ and with infinitesimal generator Q given by

$$Q = \begin{bmatrix} A_{10} & A_{00} & 0 & 0 & \cdots & \cdots \\ A_{21} & A_{11} & A_{01} & 0 & \cdots & \cdots \\ 0 & A_{22} & A_{12} & A_{02} & \cdots & \cdots \\ \cdots & \cdots & & & \ddots & \\ \cdots & \cdots & & & & \ddots \end{bmatrix}$$

The state space is partitioned into levels $\underline{i} = \{(i,j) : 1 \leq j \leq n_i\}$ and transitions take place only to the adjacent levels. However, here the transition rates may depend on the level i and therefore the spatial homogeneity of the associated process is lost. All $A_{1i}s$ are square matrices; but, since different levels may contain different number of phases, the A_{2i} matrices and A_{0i} matrices are in general rectangular. Assuming that the QBD'is irreducible we have the following theorem.

THEOREM 1.2. *When the QBD is positive recurrent, its steady state distribution $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ satisfies the relation*

$$\pi_n = \pi_{n-1} R_n \text{ for } n \geq 1 \tag{1.13}$$

where the matrices R_n are the minimal nonnegative solutions of the system of equations

$$R_n R_{n+1} A_{2,n+1} + R_n A_{1n} + A_{0,n-1} = 0, \text{ for } n \geq 1. \tag{1.14}$$

Regarding the positive recurrence of the above QBD we have the following theorem.

THEOREM 1.3. *The QBD is positive recurrent if, and only if, the system of equations*

$$\pi_0 = \pi_0(A_{10} + R_1 A_{21}) \tag{1.15}$$

$$\pi_0 \sum_{n\geq 1} \{ (\prod_{1\leq k\leq n} R_k) e \} = 1 \tag{1.16}$$

has a positive solution for π_0.

To calculate the matrices R_n and the infinite sum in (1.16), *different truncation procedures such as the one by Bright and Taylor* [13] *(which can be applied in all cases) and in the case of retrial queues, Neuts-Rao Truncation (see* [45]*) etc. can be applied.*

For further details on Matrix Analytic Methods used in Stochastic Processes we refer to Neuts [44], Latouche and Ramaswami [41]. An excellent bibliographical survey on Matrix-Analytic Methods is provided in Gómez-Corral [29].

Modelling tools

Continuous-time phase type distribution (PH distribution)

To describe a continuous-time Phase Type distribution we consider a continous time Markov Chain with states $\{1, 2, \ldots, m+1\}$ and infinitesimal generator

$$Q = \begin{bmatrix} T & T^0 \\ 0 & 0 \end{bmatrix}$$

where the $m \times m$ matrix $T = (T_{i,j})\ i, j = 1, \ldots, m$ has the property that $T_{ij} < 0$ for $1 \leq i \leq m$, and $T_{i,j} \geq 0$ for $i \neq j$. Also $T\mathrm{e} + T^0 = 0$. The initial probability vector of Q is given by (α, α_{m+1}) where α_{m+1} is a scalar and $\alpha\mathrm{e} + \alpha_{m+1} = 1$. To make all the states $1, 2 \ldots, m$ transient to ensure absorption to the state $m+1$ a certain event, starting from any initial state, we assume that the matrix T is non singular.

DEFINITION 1.1. *A random variable X is said to have phase type distribution with representation (α, T) of order m if and only if X represents the time until absorption in a finite state(with $m+1$ states) Markov process described above.*
If the random variable X has a PH distribution with representation (α, T) of order m then

(1) The distribution function of X is given by
$F(x) = P(X \leq x) = 1 - \alpha \exp(Tx)\mathrm{e}.$

(2) *The distribution $F(\cdot)$ has a jump of magnitude α_{m+1} at $x = 0$ and the probability density function $f(x)$ on $(0, \infty)$ is given by*

$f(x) = \alpha \exp(Tx)T^0$

(3) *The Laplace-Stieltjes transform $f^*(s)$ of X is given by*

$f^*(s) = \alpha_{m+1} + \alpha(sI - T)^{-1}T^0$, *for $Re(s) \geq 0$*

(4) *The moments about origin are given by*

$E(X^i) = \mu_i = (-1)^i i!(\alpha T^{-1}\mathrm{e})$, *for $i \geq 0$*

The class of continuous time Phase type distribution contains a lot of important distributions such as exponential, Erlang, etc.

Discrete-time phase type distribution : To define a discrete time PH distribution, we proceed as in the continuous case but here we take a discrete time Markov Chain with states $\{1, 2, \ldots, m+1\}$ and transition probability matrix P given by

$$P = \begin{bmatrix} T & T^0 \\ 0 & 1 \end{bmatrix}$$

where T is a square matrix of order m and $T\mathrm{e} + T^0 = \mathrm{e}$. Similar to the continuous case, the necessary and sufficient condition for eventual absorption into the absorbing state is that the matrix $I - T$ is nonsingular. The initial probability vector of the Markov Chain is (α, α_{m+1}) where $\alpha\mathrm{e} + \alpha_{m+1} = 1$. If the random variable X denotes the number of steps for absorption in a Markov Chain described as above, the probability distribution $\{p_k = P(X = k)\}_{k \geq 1}$ is given by $p_0 = \alpha_{m+1}$, and $p_k = \alpha T^{k-1}T^0$, for $k \geq 1$

The random variable X is then said to have a discrete-time Phase type distribution with representation (α, T) of order m.

The i^{th} factorial moment of X is given by

$\mu_i' = i!\alpha T^{i-1}(I - T)^{-i}\mathrm{e}$, for $i \geq 1$.

Some **useful properties** of Phase type distributions are the following.

(a) finite convolutions of continuous PH-distributions is again a PH-distribution.

(b) a finite convex mixture of PH-distribution is again a PH-distribution

(c) an infinite mixture, $G(\cdot) = \sum_{k=0}^{\infty} p_k F^{(k)}(\cdot)$ where $\{p_k\}$ is a discrete PH-distribution and $F^{(k)}(\cdot)$ is the k-fold convolution of a continous PH-distribution $F(\cdot)$, is again a PH-distribution.

(d) The class of continuous PH-distributions is dense in the class of all continuous distributions with support on the non negative real line.

PH-renewal processes : A renewal process whose inter-renewal times have a PH-distribution is called a PH-Renewal process.

To construct a PH-Renewal process we consider a continuous time Markov Chain with states $\{1, 2, \ldots, m+1\}$ having infinitesimal generator

$$Q = \begin{bmatrix} T & T^0 \\ 0 & 0 \end{bmatrix}$$

The $m \times m$ matrix T is taken to be nonsingular so that absorption to the state $m + 1$ occurs with probability 1 from any initial state. Let $(\alpha, 0)$ where 0 is a scalar, be the initial probability vector. When absorption occurs in the above chain we assume that an arrival to the system has occurred and the process immediately starts anew in one of the states $\{1, 2, \ldots, m\}$ using the probability vector α. Continuation of this procedure gives us a non terminating arrival process and is called PH-renewal process.

The class of PH-renewal processes include Poisson process, Compound Poisson Process etc.

Continuous time PH distributions and PH-Renewal processes can be used to model service time distributions and arrival processes respectively in Queueing Models.

In the case of Queueing systems which are modelled using a finite continuous time Markov Chain, the random variables associated with the queueing process such as the waiting time of a customer, time between two successive departures, a busy period etc. are often seen to follow a PH-distribution so that the distributions of these random variables as well as their expected values can be efficiently calculated using the properties of PH-distributions.

For more details and properties of PH-type distributions we refer to Neuts [44], Latouche and Ramaswami [41], Chakravarthy [14].

Batch Markovian Arrival Process (BMAP) :

To get a Batch Markovian Arrival Process we consider a two dimensional Markov Process $X(t) = \{(N(t), J(t)) : t \geq 0\}$ on the state space $\{(i, j) : i \geq 0, 1 \leq j \leq m\}$ with infinitesimal generator given by

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \cdots \\ 0 & D_0 & D_1 & D_2 & \cdots \\ 0 & 0 & D_0 & D_1 & \ddots \\ \vdots & \vdots & & \ddots & \ddots \end{bmatrix}$$

where D_k $k \geq 0$, are $m \times m$ matrices; D_0 has negative diagonal elements and nonnegative off-diagonal elements; D_k for $k \geq 1$ are nonnegative and the matrix D given by $D = \sum_{k=0}^{\infty} D_k$ is an irreducible infinitesimal generator of a continuous time Markov chain. We assume that $D \neq D_0$. The variable $N(t)$ denotes the number of arrivals in $(0, t]$, and the variable $J(t)$ denotes phase of the arrival process. The transition from a state (i, j) to a state $(i + k, l)$ where $k \geq 1, 1 \leq j, l \leq m$ with transition rates governed by the matrix D_k, correspond to the arrival of a batch of size k, while a transition from a state (i, j) to a state (i, l), $1 \leq j,\ l \leq m$; $j \neq l$, with transition rates governed by the matrix D_0, correspond to no arrival. Thus the matrix D_0 governs transitions that correspond to no arrival and the matrix D_k governs transitions corresponding to a batch arrival of size k, $k \geq 1$. We assume that the matrix D_0 is a stable matrix (see Bellman [8]) which makes it non singular and which in turn ensures that the sojourn time in the set of states $\{(i, j) : 1 \leq j \leq m\}$ is finite with probability 1 for all i. This ensures that the arrival process $X(t)$ never terminates.

Let π be the stationary probability vector of the Markov process with generator D. The fundamental arrival rate for the arrival process is then given by

$$\delta = \pi(\sum_{k=1}^{\infty} kD_k)\mathbf{e}.$$

For more details on BMAPs we refer to Lucantoni [42].

Markovian arrival process :

A Markovian Arrival Process(MAP) is a particular case of BMAP where maximum possible batch size is 1, that is, we make $D_k = 0$, for $k \geq 2$, so that here $D = D_0 + D_1$. A construction of MAP with representation matrices (D_0, D_1) of order m is as follows: Consider a Markov process with state space $\{1, 2, \ldots, m, m+1\}$ with infinitesimal generator

$$Q = \begin{bmatrix} D_0 & \mathbf{d} \\ 0 & 0 \end{bmatrix}$$

where D_0 is an $m \times m$ matrix, $D_0\mathbf{e} + \mathbf{d} = 0$ and $m+1$ is an absorbing state. Since by assumption D_0 is a stable nonsingular matrix, absorption occurs with probability 1 from any initial state. As in the construction of PH-renewal process, when absorption occurs we assume that an arrival has occurred and we immediately restart the process using an initial probability vector. But different from PH-renewal process here this initial probability vector depends also on the state from which absorption occurred and this brings dependence between interarrival times. Let $(\alpha_i, 0)$, where α_i is an m-dimensional row vector with $\alpha_i\mathbf{e} = 1$, be the probability vector which we use to restart the process after absorption has occurred from the state i and define the $m \times m$ matrix D_1 by $(D_1)_{i,j} = (\mathbf{d})_i(\alpha_i)_j$ $1 \leq i, j \leq m$. Now the matrix $D = D_0 + D_1$ will be the generator matrix of a Markov process $\{Y(t) : t \geq 0\}$ on the state space $\{1, 2, \ldots, m\}$. Let $N(t)$ denotes the number of arrivals in $(0, t]$. Then the 2-dimensional Markov Process $\{(N(t), Y(t)) : t \geq 0\}$ with state space $\{(i, j) : i \geq 0, 1 \leq j \leq m\}$ is the arrival process which we constructed above and is called Markovian Arrival Process. The infinitesimal generator of the process is given by

$$Q = \begin{bmatrix} D_0 & D_1 & 0 & 0 & \cdots \\ 0 & D_0 & D_1 & 0 & \cdots \\ 0 & 0 & D_0 & D_1 & \cdots \\ & & & \ddots & \ddots \end{bmatrix}$$

For more details on MAPs refer to Lucantoni **[42]**, Chakravarthy **[14]**.

Markovian Service Process (MSP) : By defining Markovian service process we wish to bring correlation between two successive service times. We shall construct an MSP in the same way as we constructed a MAP that is by taking a Markov process with state space $\{1, 2, \ldots, m, m+1\}$ and with infinitesimal generator

$$Q = \begin{bmatrix} D_0 & \mathrm{d} \\ 0 & 0 \end{bmatrix}$$

her D_0 is an $m \times m$ matrix, $D_0\mathrm{e} + \mathrm{d} = 0$ and $m+1$ is an absorbing state. The matrix D_0 is assumed to be a stable matrix so that absorption is certain from any initial state i. Here an absorption is considered as a service completion and if the service is to be restarted immediately we do this by restarting the above Markov process otherwise we freeze the process until the beginning of the next service and then restart it. In both cases we restart the process using a probability vector $(\alpha_i, 0)$, where α_i is an m-dimensional row vector and $\alpha_i\mathrm{e} = 1$, if the absorption has occurred from state i. This dependence of the initial probability vector on the state from which absorption has occurred makes two service times dependant random variables.

Literature survey pertaining to the thesis :

For a detailed discussion on retrial queues one may refer to the monograph by Falin and Templeton citeft and for more recent developments the papers by Artalejo **[2, 1]**. An information theoretic approach to the analysis of $M|G|1$ retrial queues is provided in Artalejo [3]. Retrial queues in discrete time has been extensively analyzed by Nobel, see for example [46].

Due to recent applications in health care systems [11, 56, 60] and in queues with impatient customers arising in telecommunication networks [5, 4, 61, 62] and inventory systems with perishable goods [30, 47], there has been renewed interest in prioritization of units in queueing models.

A large number of probabilistic models possessing variety of priorities have been discussed. Ordinarily, most chapters in textbooks [31, 33, 55] and papers [25, 28, 39, 49] on priority queues treat with exogenous priority rules; i.e., the decision of selecting the next unit for service may depend only upon the knowledge of the priority class to which the unit belongs. Nevertheless, in many situations, the exogenous disciplines might not be true. For example, in several medical procedures, patients are treated according to the urgency of their conditions, in such a way that all patients are homogeneous in their initial condition and change while waiting for treatment. Thus a key management issue of a medical service is to prioritize patients to reduce the suffering and risk faced by them in queue by implementing a dynamic priority rule even if they have initial homogeneous conditions. See for example [33, Chapter 7], and [55, Chapter 3], for a review on the methods and models related to endogenous priority disciplines and their applications.

A paper by Wang [60] discuses patient queue models with self-generation of priorities, though he does not mention this terminology explicitly, where all time variables are assumed to be exponentially distributed. To be concrete, Wang incorporates the condition and its changes over the time for a patient in queue, and stresses that it is important to study queueing models in health care systems with more general distributional assumptions on the service times and the arrival pattern. However self-generation of priorities of customers in queues have been introduced by A. Krishnamoorthy, Viswanath. C. Narayanan & T. G. Deepak (2002, unpublished paper).

Self-generation of priorities by units in queue may be thought of as a consequence of their impatient behaviour (see **[61**, Section 2]). Classical queueing theory on impatient units **[5, 4, 51, 53, 54]** usually concerns with models in which units wait for service for a (random or fixed) limited time only and leave the system forever if service has not begun within that time. For the special case of exponentially distributed services, queueing models with impatient units have been studied by Barrer **[6]**, [7] and later by Gnedenko and Kovalenko **[27]** who corrected an error in Barrer's reasoning which, however, does not invalidate his results. For the case of deterministic service times a closely related model was studied by Hokstad [32] and Swensen **[52]**. Other related works can be seen in **[19, 35, 24, 48, 50]** and references therein. See the survey of perishable inventory theory by Nahmias **[43]** for further details on how upper limits on the waiting time indicate maximal times the goods can be stored before their quality degrades.

A k-out-of-n system is characterized by the fact that the system operates as long as there are atleast k operational components. A k-out-of-n system can further be classified as follows:

The system is called 'COLD' if the operational components do not fail while the system is in down state. It is called 'HOT' if operational components continue to deteriorate at the same rate while the system is down as when it is up. The system is called 'WARM' if the deterioration rate while the system is up differs from that when it is down. An extensive study of k-out-of-n systems can be seen in Krishnamoorthy et al **[38]**, Chakravarthy, Krishnamoorthy & Ushakumari **[15]**. Krishnamoorthy and Ushakumari [37] is the first work to introduce retrial into reliability. In that paper they assume the failed components of the k-out-of-n system to proceed to a repair facility which when found busy, these components are sent to an orbit. They studied the system in the three cases, namely, COLD, WARM, and HOT. Ushakumari and Krishnamoorthy **[58]** generalize the above mentioned work to the case of arbitrarily distributed service time and derive several system performance measures. Bocharov et al **[10]** discusses a retrial queueing system with a finite waiting space, Poisson arrival of customers and arbitrarily distributed service time. Customers in the waiting space have priority over customers in the system. Choi and Chang [17] provide a survey of single server queues with priority calls. One may refer to Choi and Chang [16] for results on multi-server queues with two types of arrivals.

Postponement of work is a common phenomena. This may be to attend a more important job than the one being processed at present or for a break or due to lack of quorum (in case of bulk service, or when N-policy for service is applied) and so on. Queueing systems with postponed work is investigated in Deepak, Joshua and Krishnamoorthy [20].

Author's contribution : Chapter 2 discusses Reliability of a 'k-out-of-n system' where where the server also attends external customers when there are no failed components (main customers), under a retrial policy, which can be explained as follows: The external customers arrive according to a BMAP and the components fail at an exponential rate. If an arriving batch of external customers finds a free server one among them gets into service and others (if any) move to an orbit of infinite capacity. If an arriving batch of external customers sees a busy server, the whole batch moves in to the orbit. Service times of main and external customers follow arbitrary distributions. The stability condition and the steady state distribution are obtained. We also consider a particular case of the above problem by assuming that external arrivals are according to a MAP and also that the service times of both the main and external customers follow a PH-distribution. The numerical results obtained shows that this service to external customers decreases the idle time of the server without affecting the system reliability considerably.

Chapter 3 is an extension of the problem in chapter 2. Here also we consider a k-out-of-n system where the server provides service to external customers. The components fail at an exponential rate and the external customers arrive according a MAP. External customers who finds the server busy, joins a pool of finite capacity M, if the pool is not full; otherwise he joins an orbit of infinite capacity with probability γ or leaves the system with probability $1 - \gamma$. The orbital customers retry for service at an exponential rate θ. A retrying customer is accommodated in the pool if the pool is not full otherwise he rejoins the orbit with probability $\delta(< 1)$ and with probability $1 - \delta$ he leaves the system forever. The service to the failed components is according to an N-policy; that is the service to the components starts once all failed components are repaired, only if N failed components accumulate. In the mean time the server attends external customers in the pool. When N failed components accumulate, no more pooled customer is taken for service but the ongoing service of the external customer if there is any, is not pre-empted. The service times of both types of customers are independent and follow different PH distributions. This system is stable irrespective of the parameter values. The steady state distribution is calculated using Bright and Taylor method. Based on this some system performance measures are calculated and numerical illustrations provided.

Chapter 4 discusses reliability of 'k-out-of-n-system' where the server also attends external customers. In contrast to the assumptions in chapters 2 and 3 here instead of an orbit we assume that the external customers join a queue in a pool of infinite capacity with probability 1 if there are $< M$ failed components or with probability γ if there are M or more failed components. To reduce the impatience of a queueing customer in the pool, immediately after a service completion the server attends a pooled customer (if there is any) with probability p if there are $< L$ failed components and with probability 1 selects a pooled customer for the next service if there is any, provided the number of failed components is zero. The stationary distribution is obtained under the stability condition. A number of performance characteristics are derived. A cost function in terms of L, M, γ and p is constructed and its behaviour investigated numerically.

Chapter 5 studies a multi-server infinite capacity Queueing system where each customer arrives as ordinary but can generate into a priority customer while waiting in the queue. We call this phenomenon as 'self generation of priorities'. This phenomenon is often observed in clinics. We assume that the customer who has generated into priority is given service immediately, if there is at least one server who is not currently busy with a priority generated customer; otherwise the priority customer leaves the system for immediate service elsewhere. Arrival process is poisson and service times of each server is exponential.

The priority generation is also at an exponential rate. This system is stable irrespective of the parameter values. Stationary distribution is obtained using Bright and Taylor method. Some performance characteristics are derived and numerical illustrations provided.

Chapter 6 is on a finite capacity multi-server queueing system with self-generation of priority of customers. As in Chapter 5 the priority generated customer is either taken for service immediately if there is at least one server who is not busy with a priority generated customer; else he leaves the system for getting immediate service. The arrival of customers is according to a MAP and the service time of each server is assumed to follow a PH-distribution. Assumptions of finiteness of system capacity increases the numerical tractability and it is also close to the practical situation where the system capacity is often found to be finite. We give formulas for numerical computation for a variety of performance measures, including the blocking probability, the departure process, and the stationary distributions of the system state at pre-arrival epochs, at post-departure epochs and at epochs at which arriving units are lost. Some numerical illustrations are also provided.

Chapter 7 is on a single server infinite capacity retrial Queue where the customer in the orbit can generate into priority and leave the system if the server is already busy with a priority generated customer; else he is taken for service immediately. Arrival process is according to a MAP and service process is MSP. This system is stable irrespective of the system parameters. The steady state distribution is obtained using Neuts-Rao Truncation method where in order to choose the truncation level we use a dominating process suggested by Bright and Taylor which saves a lot of computational effort. Certain system characteristics are derived and numerical illustrations provided.

CHAPTER 2

Idle time utilisation through service to customers in a retrial queue maintaining high system reliability*

In this chapter, we discuss the reliability of a k-out-of-n system subject to repair of failed components by a server in a retrial queue. We assume that the k-out-of-n system is COLD. A k- out-of-n system is characterised by the fact that the system operates as long as there are at least k operational components. The system is COLD in the sense that operational components do not fail while the system is in down state (number of failed components at that instant is $n-k+1$). Using the same analysis as employed in this chapter, one can study the WARM and HOT systems also (a k-out-of- n system is called HOT system if operational components continue to deteriorate at the same rate while the system is down as when it is up. The system is WARM if the deterioration rate while the system is up differs from that when it is down). A repair facility, consisting of a single server, repairs the failed components one at a time. The life-times of components are independent and exponentially distributed random variables with parameter λ/i when i components are operational. Thus on an average λ failures take place in unit time when the system operates with i components. The failed components are sent to the repair facility and are repaired one at a time. The waiting space has capacity to accommodate a maximum of $n-k+1$ units in addition to the unit undergoing service. Service times of main customers (components of the k-out-of-n system) are *iid rvs* with distribution function B_1.

---

* The material in this chapter was published under the title *Reliability of a k-out-of-n system through retrial queues* in Transactions of XXIV-th International Seminar on Stability Problems for Stochastic Models, Transport & Communication Institute, Riga, Jurmala, Latvia, September, 10–17, 2004, Ed. A. Andronov, P. Bocharov & V. Korolev, pp. 232–245.

In addition to repairing failed components of the system, the repair facility provides service to external customers. However these customers are entertained only when the server is idle (no component of the main system is in repair nor even waiting). These customers are not allowed to use the waiting space at the repair facility. So when external customers arrive for service (arrival process is BMAP) when the server is busy serving a component of the system or an external customer, they are directed to an orbit and try their luck after a random length of time, exponentially distributed with parameter α_i when there are i customers in orbit.

We stress the fact that at the instant when an external customer undergoes service if a component of the system fails the latter's repair starts only on completion of service of the external customer. That is, external customers are provided non-preemptive service. The service times of external customers are *iid rvs* with distribution function B_2. Since external arrivals form a BMAP, either all in an arriving batch will proceed to an orbit on encountering a busy server; else one among the customers in the batch proceeds for service and the rest are directed to the orbit if the server is idle at that arrival epoch.

The objective of this chapter is to maximise the system reliability. Simultaneously we try to utilize the server idle time. k-out-of-n system is investigated extensively (see Krishnamoorthy et al [38] and references therein). Krishnamoorthy and Ushakumari [37] is the first work to introduce retrial into reliability. In that paper they assume the failed components of the k-out-of-n system to proceed to a repair facility, which when found busy, these components are sent to an orbit. They studied the system in the three cases, namely, COLD, WARM and HOT. Ushakumari and Krishnamoorthy [58] generalize the above mentioned work to the case of arbitrarily distributed service time and derive several system performance measures. Bocharov et al [10] discusses a retrial queueing system with a finite waiting space, Poisson arrival of customers and arbitrarily distributed service time. Customers in the waiting space have priority over customers from orbit. However their model differs from our present work in that in the former, orbital customers, at the time of retrial, can join the buffer if it is found to be not full. They obtain the stationary distribution of the primary queue size (number in the waiting space), a recurrent algorithm for the factorial moments of the number of retrial customers and an expression for the expected number of customers in the system. The model discussed here differs from Bocharov et al described above in that in this chapter priority is given to failed components of the k-out-of-n system which alone can be accommodated in the waiting space. Further there is only one service of primary customers in Bocharov et al model whereas the one discussed here has two distinct services-components of k-out-of-n system and internal customers. Choi and Chang **[17]** provides a survey of single server queues with priority calls. One may also refer to Choi and Chang **[16]** for results on multiserver queues with two types of arrivals.

This chapter is arranged as follows. In section 2.1 we provide the mathematical modelling of the system under study. In section 2.2 through 2.5 we investigate the stationary distribution of the embedded Markov chain. In 2.6 distribution of the system state at arbitrary epochs is provided. System performance characteristics are provided in section 2.7. In section 2.8 a particular case of the problem discussed in section 2.1 is analysed in depth. Section 2.9 provides some performance measures of this particular case and in section 2.10 a numerical illustration is given.

2.1. The mathematical model

The system has a single server. The server serves the main customers (components of the k-out-of-n system) and external customers according to distribution functions B_1 and B_2, respectively. Because of the assumption we made about the life times of components of the k-out-of-n system, the main customers arrival (see previous section) has exponentially distributed interarrival times of rate λ. The arrival of external customers is according to a BMAP defined by the matrix generating function

$$D(z) = \sum_{m=0}^{\infty} D_m z^m, \ |z| < 1.$$

This arrival process is governed by the continuous time Markov chain $\{\nu_t, t \geq 0\}$, having state space $\{0, 1, \ldots, W\}$. The sequence of matrices $\{D_k\}$ provide the transition rates from state i to state j in the Markov chain and the consequent arrival of a batch of customers of size k, $k = 0, 1, 2, \ldots$.

The steady state distribution of the process ν_t, $t \geq 0$, is defined by the row vector $\vec{\theta}$ that satisfies equations $\vec{\theta}D(1) = \vec{0}$, $\vec{\theta}\mathrm{e} = 1$. The fundamental rate of the BMAP is $\delta = \vec{\theta}D'(1)\mathrm{e}$. Here and in the sequel $\vec{\theta}$ is a row vector of corresponding dimension, e is a column vector consisting of 1's. See Lucantoni [42] and Chakravarthy [14] for more details about the BMAP. The external customer can access the server only if the server is idle. Otherwise the customer moves to the orbit and tries his luck later. The interretrial times are exponentially distributed with parameter α_i when i customers are present in the orbit, $i > 0$, $\alpha_0 = 0$. The service times of external customers is a random variable characterised by the distribution function $B_2(t)$. Let $b_1^{(r)} = \int_0^\infty t\, dB_r(t)$ the average service time under the service time distribution $B_r(t)$, $r = \overline{1, 2}$.

From the given description, it is clear that the main customers have a priority with respect to the external customers. External customers have a chance to get a service only in case the server is idle which is possible only if there is no main customer in the repair facility at the time of commencement of the service of the former. We assume that the priority is non-preemptive— arrival of a main customer does not interrupt the service of the external customer, if any, in the system.

Our aim is to calculate the main performance characteristics of the model.

2.2. Stationary state distribution of the system

Let j_t be the number of main customers in the queue at the epoch t, $0 \le j_t \le n-k+1$ and i_t be the number of customers in the orbit at t, $i_t \ge 0$.

$$\tau_t = \begin{cases} 0, & \text{if the server is idle at epoch } t, \\ 1, & \text{if a main customer is getting processed at epoch } t, \\ 2, & \text{if an external customer is served at the epoch } t, t \ge 0, \end{cases}$$

ν_t be the state of the BMAP at the epoch t, $\nu_t = 0, \dots, W$.

Consider the process

$$\xi_t = (i_t, j_t, \tau_t, \nu_t), t \ge 0.$$

Unfortunately the process $\{\xi_t, t \ge 0,\}$ is non-Markovian. So, to investigate this process consider first the embedded chain at the service completion epochs, *ie.*, the Markov chain $\{\zeta_n, n \ge 1\}$, that is defined as:

$$\zeta_n = \{i_{t_n+0}, j_{t_n+0}, \nu_{t_n}\}, n \ge 1,$$

where t_n is the nth service completion epoch.

2.3. Specification of the embedded Markov chain

It can be verified that the process ζ_n, $n \ge 1$, actually is a Markov chain. Denote its one-step transition probabilities as

$$P\{(i,j,\nu) \to (l,j',\nu')\} = P\{i_{t_{n+1}+0} = l, j_{t_{n+1}+0} = j', \nu_{t_{n+1}} = \nu' | \\ i_{t_n+0} = i, j_{t_n+0} = j, \nu_{t_n} = \nu\}.$$

Enumerating the states of the chain $\{\zeta_n, n \geq 1\}$ in the lexicographic order, we form transition matrices

$$P_{(i,j),(l,j')} = \|P\{(i,j,\nu) \to (l,j',\nu')\}\|_{\nu,\nu'=\overline{0,W}}$$

and the block matrices $P_{i,l} = \|P_{(i,j),(l,j')}\|_{j,j'=\overline{0,n-k+1}}$.

LEMMA 2.1. *The transition probability matrices $P_{(i,j),(l,j')}$ are calculated as follows.*

$$P_{(i,j),(l,j')} = \Omega^{(1)}(l-i, j'-j+1), \quad i \geq 0,\ l \geq i,\ 0 < j \leq j'+1, j' \leq n-k \quad (2.1)$$

$$P_{(i,j),(l,n-k+1)} = \hat{\Omega}^{(1)}(l-i, n-k+2-j), l \geq i \geq 0, 1 \leq j \leq n-k+1, \quad (2.2)$$

$$P_{(i,0),(l,j')} = \Psi(i,l,j') = R_i\alpha_i I\Omega^{(2)}(l-i+1, j')$$
$$+ R_i \sum_{m=1}^{l-i+1} D_m \Omega^{(2)}(l-i-m+1, j) + R_i \lambda I \Omega^{(1)}(l-i, j'),$$
$$i \geq 0, l \geq \max\{0, l-i\}, j' = \overline{0, n-k}. \quad (2.3)$$

For $j' = n-k+1$, formula (2.3) *is valid if we provide symbols Ψ, $\Omega^{(r)}$ with a hat. Here*

$$R_i = (-D_0 + \alpha_i I + \lambda I)^{-1}, \quad (2.4)$$

$$\Omega^{(r)}(m,\tau) = \int_0^\infty P(m,t)\frac{(\lambda t)^\tau}{\tau!}e^{-\lambda t}dB_r(t), \quad (2.5)$$

$$\hat{\Omega}^{(r)}(m,\tau) = \sum_{l=\tau}^{\infty} \Omega^{(r)}(m,\tau), \quad (2.6)$$

the matrices $P(m,t)$ are defined by the series: $\sum_{m=0}^{\infty} P(m,t)z^m = e^{D(z)t}$.

This Lemma follows from the following reasonings. The matrix $\Omega^{(r)}(m,\tau)$ defines probability of τ main customers and m external customers arrival (with the corresponding transitions of the chain ν_t, $t \geq 0$) during time having distribution function $B_r(t)$, $r = 1, 2$. The matrices $\alpha_i R_i$, $R_i D_m$, λR_i define transitions of the process ν_t, $t \geq 0$, during the idle period of the server that is terminated by the retrial from the orbit, arrival of external

customers in batch of size m and arrival of main customer, respectively.
From (2.1)–(2.3) we see that transition matrix $P_{i,l}$ has the following structure:

$$P_{i,l} = \begin{pmatrix} \Psi(i,l,0) & \Psi(i,l,1) & \cdots & \Psi(i,l,n-k) & \hat{\Psi}(i,l,n-k+1) \\ \Omega^{(1)}(l-i,0) & \Omega^{(1)}(l-i,1) & \cdots & \Omega^{(1)}(l-i,n-k) & \hat{\Omega}^{(1)}(l-i,n-k+1) \\ 0 & \Omega^{(1)}(l-i,0) & \cdots & \Omega^{(1)}(l-i,n-k-1) & \hat{\Omega}^{(1)}(l-i,n-k) \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \cdots & \Omega^{(1)}(l-i,0) & \hat{\Omega}^{(1)}(l-i,1) \end{pmatrix} \tag{2.7}$$

Thus we calculated the one-step transition probabilities of the Markov chain ζ_n, $n \geq 1$.

2.4. Stability condition

To investigate the Markov chain ζ_n, $n \geq 1$, we should make some assumptions about the limiting behaviour of the total intensities of retrials α_i, $i > 0$. We distinguish two cases : $\lim_{i\to\infty} \alpha_i = \infty$ and $\lim_{i\to\infty} \alpha_i = \gamma < +\infty$. The first case includes the classical strategy of retrials when $\alpha_i = i\alpha$ and the second one includes the constant retrial rate $(\alpha_i = \theta, i \geq 1)$. In case $\lim\limits_{i\to\infty} \alpha_i$ does not exist, we can not speak definitely about the limiting behaviour of the queueing system. So we restrict ourselves only to the first cases described above.

In the case $\lim_{i\to\infty} \alpha_i = \gamma$, we see that the matrices $\Psi(i,l,\nu)$, $\hat{\Psi}(i,l,n-k+1)$ depend on l and i only via the difference $l - i$. In this case, for $i > 0$ we have

$$Y(z) = \sum_{l=i-1}^{\infty} P_{i,l} z^{l-i+1} =$$

$$\begin{pmatrix} \Psi(z,0) & \Psi(z,1) & & \Psi(z,n-k) & \hat{\Psi}(z,n-k+1) \\ zY^{(1)}(z,0) & zY^{(1)}(z,1) & \cdots & zY^{(1)}(z,n-k) & z\hat{Y}^{(1)}(z,n-k+1) \\ 0 & zY^{(1)}(z,o) & \cdots & zY^{(1)}(z,n-k-1) & z\hat{Y}^{(1)}(z,n-k) \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & zY^{(1)}(z,0) & z\hat{Y}^{(1)}(z,1) \end{pmatrix} \tag{2.8}$$

where

$$\Psi(z,\nu) = R(\gamma Y^{(2)}(z,\nu) + (D(z) - D_0)Y^{(2)}(z,\nu) + \lambda z Y^{(1)}(z,\nu))$$

$$\hat{\Psi}(z,\nu) = R(\gamma \hat{Y}^{(2)}(z,\nu) + (D(z) - D_0)\hat{Y}^{(2)}(z,\nu) + \lambda z \hat{Y}^{(1)}(z,\nu)),$$

$$R = (-D_0 + \gamma I + \lambda I)^{-1},$$

$$Y^{(r)}(z,\nu) = \int_0^\infty e^{D(z)t}\frac{(\lambda t)^\nu}{\nu!}e^{-\lambda t}dB_r(t),$$

$$\hat{Y}^{(r)}(z,\nu) = \sum_{l=\nu}^{\infty} Y^{(r)}(z,l), r = 1,2.$$

Stability condition for this case is given by the following.

THEOREM 2.1. *The stationary distribution of the Markov chain ζ_n, $n \geq 1$, exists if, and only if, the inequality*

$$\mathrm{x}Y'(1)\mathrm{e} < 1 \tag{2.9}$$

holds where x *is the row vector which is the unique solution to the system:*

$$\mathrm{x}Y(1) = \mathrm{x}, \mathrm{xe} = 1. \tag{2.10}$$

Proof follows from (Klimenok [34]).

Consider now the case $\lim_{i\to\infty} \alpha_i = \infty$.

Let $\rho_\tau = \delta b_i^{(\tau)}$, $\tau = 1,2$, and y_0 be the probability of the idle state for the $M|G|1|n-k+1$ system with the stationary Poisson arrival process with intensity λ and the service time distribution $B_1(t)$ if the system was not idle at the previous service completion epoch and the service time distribution $B_2(t)$ in the opposite case. The problem of calculating the value y_0 can be solved trivially and we consider it to be known (stable procedure for its calculation directly follows from (Dudin, Klimenok, Tsarenkov [22]))

THEOREM 2.2. *The stationary distribution of the Markov chain ζ_n, $n \geq 1$, exists if, and only if, the inequality*

$$y_0(1 + \rho_1 - \rho_2) > \rho_1 \quad \textit{holds.} \tag{2.11}$$

PROOF. It can be verified that the Markov chain ζ_n, $n \geq 1$, which has transition probabilities (2.7), belongs to the class of asymptotically quasitoeplitz Markov chains (see Dudin, Klimenok [**21**]). Stability condition for such chains is known in terms of the matrix generating function

$$\tilde{Y}(z) = \lim_{i\to\infty} \sum_{m=-1}^{\infty} P_{i,i+m} z^{m+1}.$$

It is defined by formulas (2.9), (2.10) where the matrix generating functions $Y(z)$ is replaced by the function $\tilde{Y}(z)$. It is easy to see that the function $\tilde{Y}(z)$ is defined by the formula (2.8) where the symbol Ψ is replaced by the symbols $Y^{(2)}$. By means of substitution, it can be verified that the vector $\mathbf{x}$, which is the solution of the system $\mathbf{x}\tilde{Y}(1) = \mathbf{x}$, $\mathbf{x}\mathbf{e} = 1$, has the form:

$$\mathbf{x} = \mathbf{y} \otimes \vec{\theta}, \tag{2.12}$$

where $\mathbf{y}$ is the vector of stationary probabilities of the queueing system $M|G|1|n-k+1$ defined above and $\otimes$ stands for Kronecker product of matrices.

Inequality (2.9) is reduced to the inequality

$$y_0\rho_2 + \sum_{l=1}^{n-k+1} y_l\rho_1 - y_0 < 0 \tag{2.13}$$

if we take into account that $\vec{\theta}\sum_{m=o}^{\infty}\left(Y^{(r)}(z,m)\right)'\Big|_{z=1} e = \rho_r$. Now inequality (2.11) follows from (2.13) and the normalisation condition $\sum_{l=0}^{n-k+1} y_l = 1$.

This completes the proof of Theorem 2.2. □

REMARK 2.1. *Condition* (2.11) *is well tractable. When the number of customers in the orbit is large, the value* $(1 - y_0)\rho_1 + y_0\rho_2$ *is the average number of external customers* *arriving into the system during the arbitrary service time. Average number of the external customers leaving the system after the arbitrary service completion epoch is equal to* $y_0(y_0 = 1 \cdot y_0 + 0 \cdot (1 - y_0))$. *The intuitive stability condition* $y_0 > (1 - y_0)\rho_1 + y_0\rho_2$ *is equivalent to* (2.11).

Assume that condition (2.9) or (2.11) (depending on the case considered) is fulfilled.

2.5. Stationary distribution of the embedded Markov chain

Define the steady state probabilities of the Markov chain ζ_n, $n \geq 1$, as

$$\pi(i, j, \nu) = \lim_{n\to\infty} P\{i_{t_n+0} = i, j_{t_n+0} = j, \nu_{t_r} = \nu\}$$

and form vectors

$$\vec{\pi}(i, j) = (\pi(i, j, 0), \dots, \pi(i, j, W)),$$

$$\vec{\pi}(i) = (\vec{\pi}(i, 0), \dots, \vec{\pi}(i, n - k + 1)), \quad i \geq 0.$$

Stable procedure for calculating the vectors $\vec{\pi}(i)$, $i \geq 0$, presented in (Breuer, Dudin, Klimenok, [12]) is applicable to our model. So, the problem of calculation of the stationary probabilities of the embedded Markov chain can be considered as being solved.

2.6. Stationary distribution of the system at arbitrary time

We assume that the service times are not negligible and have a finite mean. It implies that under the fulfilment of stability conditions (2.9) or (2.11) for the embedded Markov chain ζ_n, $n \geq 1$, the stationary state distribution of the process ξ_t, $t \geq 0$, exists as well. Write

$$p(i, j, r, \nu) = \lim_{t\to\infty} P\{i_t = i, j_t = j, r_t = r, \nu_t = \nu\},$$

$$i \geq 0, r = 0, 1, 2, \nu = \overline{0, W}, 0 \leq j \leq n - k + 1.$$

THEOREM 2.3. *The vectors* $\vec{p}(i,j,r) = (p(i,j,r,0),\dots,p(i,j,r,W))$ *are calculated as follows:*

$$\vec{p}(i,0,0) = \tau^{-1}\vec{\pi}(i,0)R_i, i \geq 0,$$

$$\vec{p}(i,j,1) = \tau^{-1}\Big[\sum_{l=0}^{i}\sum_{m=1}^{j+1}\vec{\Pi}(l,m)\bar{\Omega}^{(1)}(i-l,j-m+1)$$
$$+\sum_{l=0}^{i}\vec{\Pi}(l,0)R_l\lambda\bar{\Omega}^{(1)}(i-l,j)\Big], i \geq 0, j=\overline{0,n-k},$$

$$\vec{p}(i,n-k+1,1) = \tau^{-1}\Big[\sum_{l=0}^{i}\sum_{m=1}^{n-k+1}\vec{\Pi}(l,m)\hat{\bar{\Omega}}(i-l,n-k+2-m)$$
$$+\sum_{l=0}^{i}\vec{\Pi}(l,0)R_l\lambda\hat{\bar{\Omega}}^{(1)}(i-l,n-k+1)\Big], i \geq 0$$

$$\vec{p}(i,j,2) = \tau^{-1}\Big[\sum_{l=0}^{i+1}\vec{\Pi}(l,0)R_l(\alpha_l I\bar{\Omega}^{(2)}(i-l+l,j)$$
$$+\sum_{m=1}^{i-l+1}D_m\bar{\Omega}^{(2)}(i-l-m+1,j)\Big], \; i \geq 0, j=\overline{0,n-k},$$

$$\vec{p}(i,n-k+1,2) = \tau^{-1}\Big[\sum_{l=0}^{i+1}\vec{\Pi}(l,0)R_l(\alpha_l I\hat{\bar{\Omega}}^{(2)}(i-l+1,n-k+1)$$
$$+\sum_{m=1}^{i-l+1}D_m\hat{\bar{\Omega}}^{(2)}(i-l-m+1,n-k+1))\Big], i \geq 0,$$

where

$$\tau = b_1^{(1)} + \sum_{i=0}^{\infty} \vec{\Pi}(i,0)R_i((-D_0 + \alpha_i I)\mathrm{e}(b_1^{(2)} - b_1^{(1)}) + \mathrm{e}),$$

$$\bar{\Omega}^{(r)}(m,\tau) = \int_0^\infty P(m,t)\frac{(\lambda t)^\tau}{\tau!}e^{-\lambda t}(1 - B_r(t))\,dt,$$

$$\hat{\bar{\Omega}}^{(r)}(m,\tau) = \sum_{l=\tau}^{\infty} \bar{\Omega}^{(r)}(m,l).$$

Proof of this theorem follows from the theory of Markov renewal processes (see Cinlar [18]). The value τ is the mean inter-departure time in the system.

2.7. Performance characteristics

(1) Probability of the system be empty is

$$\tau^{-1}\vec{\Pi}(0,0)(-D_0 + \lambda I)^{-1}e;$$

(2) The proportion of times during which the server is idle is

$$\tau^{-1}\sum_{i=0}^{\infty} \vec{\Pi}(i,0)R_i e;$$

(3) The proportion of time when the main customers are processed is

$$\sum_{i=0}^{\infty}\sum_{j=0}^{n-k+1} \vec{p}(i,j,1)e;$$

(4) The fraction of time during which the external customers are processed is

$$\sum_{i=0}^{\infty}\sum_{j=0}^{n-k+1} \vec{p}(i,j,2)e;$$

(5) Probability to have j main customers in the buffer is

$$\sum_{i=0}^{\infty}(\vec{p}(i,j,1)+\vec{p}(i,j,2))e;$$

(6) Probability that arbitrary external customer reaches the server without visiting the orbit is

$$\frac{1}{\delta}\sum_{i=0}^{\infty}\vec{p}(i,0,0)\sum_{l=1}^{\infty}D_l e;$$

(7) Mean number of external customers in the orbit is

$$\sum_{i=1}^{\infty}i(\vec{p}(i,0,0)+\sum_{j=0}^{n-k+1}(\vec{p}(i,j,1)+\vec{p}(i,j,2)))e.$$

2.8. Particular case

Here we assume that the arrival of external customers is according to a MAP with representation (D_0, D_1) of order m. The k-out-of-n system is assumed to be **COLD** and the lifetimes of components are assumed to be exponentially distributed with parameter $\frac{\lambda}{i}$ when i components are operational. An external arrival seeing a busy server moves to an orbit of infinite capacity. The inter-retrial times are assumed to be exponentially distributed with parameter $i\theta$, when there are i customers in the orbit. The service times of both type of customers follow a PH-distribution with representation (β, S) of order m_1. The average arrival rate of external customers is defined as $\delta = \vec{\theta}D_1\mathrm{e}$, where the vector $\vec{\theta}$ satisfies $\vec{\theta}(D_0 + D_1) = 0, \vec{\theta}\mathrm{e} = 1$. The average service rate is defined as $\mu = \frac{1}{-(\beta S^{-1}\mathrm{e})}$.

Let $N(t)$ denotes the number of customers in the orbit at time t,

$$I(t) = \begin{cases} 0 & \text{if the server is idle at time } t, \\ 1 & \text{if the server is busy with a failed component at time } t, \\ 2 & \text{if the server is busy with an external customer at time } t, \end{cases}$$

$M(t)$ = number of failed components in the system including the one getting service, if any, at time t;

$J_1(t)$ = phase of the arrival process at time t;

$J_2(t)$ = phase of the service process at time t.

Then $X(t) = (N(t), I(t), M(t), J_1(t), J_2(t)),\ t \geq 0$ forms a continuous time Markov chain with state space

$$\{(i,0,0,j_1)|i \geq 0, 1 \leq j_1 \leq m\}$$
$$\cup\{(i,1,j_3,j_1,j_2)|i \geq 0, 1 \leq j_3 \leq n-k+1, 1 \leq j_1 \leq m, 1 \leq j_2 \leq m_1\}$$
$$\cup\{(i,2,j_3,j_1,j_2)|i \geq 0, 0 \leq j_3 \leq n-k+1, 1 \leq j_1 \leq m, 1 \leq j_2 \leq m_1\}$$

and infinitesimal generator

$$Q = \begin{bmatrix} A_{10} & A_0 & 0 & 0 & \cdots \\ A_{21} & A_{11} & A_0 & 0 & \cdots \\ 0 & A_{22} & A_{12} & A_0 & \cdots \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

with $A_{10} = \begin{bmatrix} A_{10}^{(1)} & A_{10}^{(2)} \\ A_{10}^{(3)} & A_{10}^{(4)} \end{bmatrix}$

where

$$A_{10}^{(1)} = \begin{bmatrix} D_0 - \lambda I_m & \lambda I_m \otimes \beta & 0 & 0 \ldots \\ I_m \otimes S^0 & D_0 \oplus S - \lambda I_{mm_1} & \lambda I_{mm_1} & 0 \ldots \\ 0 & I_m \otimes (S^0\beta) & D_0 \oplus S - \lambda I_{mm_1} & \lambda I_{mm_1} \cdots \\ & \ddots & \ddots & \\ & & I_m \otimes (S^0\beta) & D_0 \oplus S \end{bmatrix}$$

$$A_{10}^{(2)} = \begin{bmatrix} D_1 \otimes \beta & 0 \\ 0 & 0 \end{bmatrix}, \quad A_{10}^{(3)} = \begin{bmatrix} I_m \otimes S^0 & 0 \\ 0 & I_{n-k+1} \otimes (I_m \otimes (S^0\beta)) \end{bmatrix}$$

$$A_{10}^{(4)} = \begin{bmatrix} D_0 \oplus S - \lambda I_{mm_1} & \lambda I_{mm_1} & & & \\ 0 & D_0 \oplus S - \lambda I_{mm_1} & \lambda I_{mm_1} & & \\ & & \ddots & & \\ & & & D_0 \oplus S - \lambda I_{mm_1} & \lambda I_{mm_1} \\ & & & 0 & D_0 \oplus S \end{bmatrix}.$$

For $i \geq 1$, $A_{1i} = A_{10} - \tilde{A}_{1i}$, where $\tilde{A}_{1i} = \begin{bmatrix} i\theta I_m & 0 \\ 0 & 0 \end{bmatrix}$

$A_0 = \begin{bmatrix} 0 & 0 \\ 0 & \bar{A}_0 \end{bmatrix}$, with $\bar{A}_0 = I_{2(n-k+1)+1} \otimes (D_1 \otimes I_{m_1})$. For $i \geq 1$,

$$A_{2i} = \begin{bmatrix} 0_{m\times(m+(n-k+1)mm_1)} & i\theta I_m \otimes \beta & 0_{m\times((n-k+1)mm_1)} \\ 0 & 0 & 0 \end{bmatrix}$$

The steady state distribution of the process $\{X(t) : t \geq 0\}$, when it exists, is obtained applying Neuts-Rao Truncation. That is we assume that for some fixed $N \geq 1$, the inter-retrial times are exponentially distributed with parameter $i\theta$ when the number of customers in the orbit is $i < N$ and with parameter $N\theta$ when there are N or more customers in the orbit. This assumption transforms the generator matrix Q to Q_N given by

$$Q_N = \begin{bmatrix} A_{10} & A_0 & 0 & 0 & \cdots & & & \\ A_{21} & A_{11} & A_0 & 0 & \cdots & & & \\ 0 & A_{22} & A_{12} & A_0 & \cdots & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & A_{2N-1} & A_{1N-1} & A_0 & & \\ & & & & A_{2N} & A_{1N} & A_0 & \\ & & & & & A_{2N} & A_{1N} & A_0 \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

The continuous time Markov chain with generator matrix Q_N, being a level independent QBD, is positive recurrent if and only if

$$\pi_N A_0 e < \pi_N A_{2N} e \tag{2.14}$$

where π_N is the stationary probability vector of the generator matrix

$$A_N = A_0 + A_{1N} + A_{2N} = \begin{bmatrix} A_N^{(1)} & A_N^{(2)} \\ A_N^{(3)} & A_N^{(4)} \end{bmatrix}$$

with

$$A_N^{(1)} = \begin{bmatrix} D_0 - (\lambda + N\theta)I_m & \lambda I_m \otimes \beta & 0 & 0 \dots \\ I_m \otimes S^0 & H - \lambda I_{mm_1} & \lambda I_{mm_1} & 0 \dots \\ 0 & I_m \otimes (S^0\beta) & H - \lambda I_{mm_1} & \lambda I_{mm_1} \dots \\ & \ddots & \ddots & \\ & & I_m \otimes (S^0\beta) & H \end{bmatrix}$$

where $H = (D_0 + D_1) \oplus S$.

Now,

$$A_N^{(2)} = \begin{bmatrix} (D_1 + N\theta I_m) \otimes \beta & 0 \\ 0 & 0 \end{bmatrix}, A_N^{(3)} = A_{10}^{(3)},$$

$$A_N^{(4)} = \begin{bmatrix} H - \lambda I_{mm_1} & \lambda I_{mm_1} & & & \\ 0 & H - \lambda I_{mm_1} & \lambda I_{mm_1} & & \\ & & \ddots & & \\ & & & H - \lambda I_{mm_1} & \lambda I_{mm_1} \\ & & & 0 & H \end{bmatrix}$$

Now, partitioning π_N as

$$\pi_N = (\pi_N^{0,0}, \pi_N^{1,1}, \pi_N^{1,2}, \dots, \pi_N^{1,n-k+1}, \pi_N^{2,0}, \pi_N^{2,1}, \dots, \pi_N^{2,n-k+1})$$

where each sub-vector $\pi_N^{1,j}$, $1 \leq j \leq n-k+1$ and $\pi_N^{2,j}$, $0 \leq j \leq n-k+1$, contains mm_1 elements and the sub-vector $\pi_N^{0,0}$ contains m elements, the equation $\pi_N A_N = 0$ implies

$$\pi_N^{0,0}(D_0 - (\lambda + N\theta)I_m) + (\pi_N^{1,1} + \pi_N^{2,0})(I_m \otimes S^0) = 0$$

$$\pi_N^{0,0}(\lambda I_m \otimes \beta) + \pi_N^{1,1}(H - \lambda I_{mm_1}) + (\pi_N^{1,2} + \pi_N^{2,1})(I_m \otimes (S^0\beta)) = 0$$

$$\lambda\pi_N^{1,i-1} + \pi_N^{1,i}(H - \lambda I_{mm_1}) + (\pi_N^{1,i+1} + \pi_N^{2,i})(I_m \otimes (S^0\beta)) = 0, \quad 2 \leq i \leq n-k$$

$$\lambda\pi_N^{1,n-k} + \pi_N^{1,n-k+1}H + \pi_N^{2,n-k+1}(I_m \otimes (S^0\beta)) = 0,$$

$$\pi_N^{0,0}((D_1 + N\theta I_m) \otimes \beta) + \pi_N^{2,0}(H - \lambda I_{mm_1}) = 0$$

$$\lambda\pi_N^{2,i-1} + \pi_N^{2,i}(H - \lambda I_{mm_1}) = 0, \quad 1 \leq i \leq n-k$$

$$\lambda\pi_N^{2,n-k} + \pi_N^{2,n-k+1}H = 0$$

These equations give rise to the equations:

$$\pi_N^{0,0} = (\pi_N^{1,1} + \pi_N^{2,0})(I_m \otimes S^0)((\lambda + N\theta)I_m - D_0)^{-1} \tag{2.15}$$

$$\pi_N^{1,1} = (\pi_N^{0,0}(\lambda I_m \otimes \beta) + (\pi_N^{1,2} + \pi_N^{2,1})(I_m \otimes (S^0\beta)))(\lambda I_{mm_1} - H)^{-1} \tag{2.16}$$

$$\pi_N^{1,i} = (\lambda\pi_N^{1,i-1} + (\pi_N^{1,i+1} + \pi_N^{2,i})(I_m \otimes (S^0\beta)))(\lambda I_{mm_1} - H)^{-1} \quad 2 \leq i \leq n-k \tag{2.17}$$

$$\pi_N^{1,n-k+1} = (\lambda\pi_N^{1,n-k} + \pi_N^{2,n-k+1}(I_m \otimes (S^0\beta)))(-H)^{-1} \tag{2.18}$$

$$\pi_N^{2,0} = \pi_N^{0,0}((D_1 + N\theta I_m) \otimes \beta)(\lambda I_{mm_1} - H)^{-1} \tag{2.19}$$

$$\pi_N^{2,i} = \lambda\pi_N^{2,i-1}(\lambda I_{mm_1} - H)^{-1}, \quad 1 \leq i \leq n-k \tag{2.20}$$

$$\pi_N^{2,n-k+1} = \lambda\pi_N^{2,n-k}(-H)^{-1} \tag{2.21}$$

The equations from (2.15) to (2.21) together with the normalising condition

$$\pi_N \mathbf{e} = 1 \tag{2.22}$$

can be solved using Block Gauss-Seidel iteration procedure to obtain the vector π_N.

Now inequality (2.14) becomes

$$(\sum_{i=1}^{n-k+1} \pi_N^{1,i} + \sum_{i=0}^{n-k+1} \pi_N^{2,i})((D_1\mathbf{e}_m) \otimes \mathbf{e}_{m_1}) < N\theta\pi_N^{0,0}\mathbf{e}_m \tag{2.23}$$

which is the stability condition for the Markov chain Q_N.

Now let R_N be the minimal nonnegative solution of the matrix quadratic equation

$$R^2 A_{2N} + RA_{1N} + A_0 = 0$$

The spectral radius of R_N is less than 1 if and only if inequality (2.23) is satisfied. Let η_N be the spectral radius of R_N for $N \geq 1$. Now the truncation level N is selected in such a way that inequality (2.23) is satisfied and that $|\eta_N - \eta_{N+1}| < \epsilon$, for some fixed real number $\epsilon > 0$

After selecting the truncation level N, we approximate the steady state distribution x of Q by the steady state distribution x_N of Q_N, which when partitioned according to the levels as

$$x_N = (x_N(0), x_N(1), \ldots, x_N(N), x_N(N+1), \ldots)$$

satisfies the equations:

$$x_N(0)A_{10} + x_N(1)A_{21} = 0 \tag{2.24}$$

$$x_N(i-1)A_0 + x_N(i)A_{1i} + x_N(i+1)A_{2\,i+1} = 0, 1 \leq i \leq N-2 \tag{2.25}$$

$$x_N(N-2)A_0 + x_N(N-1)A_{1\,N-1} + x_N(N)A_{2\,N} = 0, \tag{2.26}$$

$$x_N(i-1)A_0 + x_N(i)A_{1N} + x_N(i+1)A_{2\,N} = 0, i \geq N \tag{2.27}$$

together with the normalising condition

$$x_N\mathbf{e} = 1. \tag{2.28}$$

Then by the property of level independent QBDs, we can write

$$x_N(i) = x_N(N-1)R_N^{i-N+1} \quad i \geq N \tag{2.29}$$

Substituting $X_N(N) = x_N(N-1)R_N$ in (2.26) we get

$$x_N(N-2)A_0 + x_N(N-1)(A_{1\,N-1} + R_N A_{2N}) = 0$$

which implies

$$x_N(N-1) = x_N(N-2)(A_0(-(A_{1\,N-1} + R_N A_{2\,N}))^{-1})$$

Defining

$$W_{N-1} = A_0(-(A_{1\,N-1} + R_N A_{2N}))^{-1}$$

we get,

$$x_N(N-1) = x_N(N-2)W_{N-1} \tag{2.30}$$

Now the equation

$$x_N(N-3)A_0 + x_N(N-2)A_{1\,N-2} + x_N(N-1)A_{2N} = 0$$

becomes

$$x_N(N-3)A_0 + x_N(N-2)A_{1\,N-2} + x_N(N-2)W_{N-1}A_{2\,N-1} = 0$$

which gives

$$x_N(N-2) = x_N(N-3)W_{N-2} \tag{2.31}$$

where

$$W_{N-2} = A_0(-(A_{1\,N-2} + W_{N-1}A_{2\,N-1})^{-1})$$

Thus defining W_is as

$$W_i = A_0(-(A_{1i} + W_{i+1}A_{2\,i+1})^{-1}) \quad \text{for } i = N-2, N-3, \ldots, 1 \tag{2.32}$$

with

$$W_{N-1} = A_0(-(A_{1\,N-1} + R_N A_{2\,N})^{-1}) \tag{2.33}$$

we get

$$x_N(i) = x_N(i-1)W_i, \quad 1 \le i \le N-1 \tag{2.34}$$

Substituting $x_N(1) = x_N(0)W_1$ in (2.24), we get

$$x_N(0)(A_{10} + W_1 A_{21}) = 0 \tag{2.35}$$

Also equation (2.28) becomes

$$x_N(0)(I + \sum_{i=1}^{N-2}(\prod_{j=1}^{i} W_j) + (\prod_{j=1}^{N-1} W_j)(I - R_N)^{-1})\mathrm{e} = 1 \tag{2.36}$$

Now equations (2.35) and (2.36) can be solved for $x_N(0)$ and hence we can obtain the stationary distribution x_N.

2.8.1. Computation of the matrix R_N. To compute the matrix R_N we first compute G_N which is the minimal non-negative solution of the matrix quadratic equation.

$$A_0 G_N^2 + A_{1N} G_N + A_{2N} = 0 \tag{2.37}$$

Now the special structure of the matrix A_{2N} shows that G_N will have the form

$$G_N = \begin{bmatrix} 0 & 0 & \cdots & G_{00}^{(N)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ldots & G_{11}^{(N)} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & G_{12}^{(N)} & 0 & 0 & \ldots & 0 \\ & & & \vdots & & & & \\ 0 & 0 & \ldots & G_{1,n-k+1}^{(N)} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & G_{20}^{(N)} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & G_{21}^{(N)} & 0 & 0 & \ldots & 0 \\ & & & \vdots & & & & \\ 0 & 0 & \ldots & G_{2,n-k+1}^{(N)} & 0 & 0 & \ldots & 0 \end{bmatrix}$$

$$G_N^2 = \begin{bmatrix} 0 & 0 & \cdots & G_{00}^{(N)} G_{20}^{(N)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ldots & G_{11}^{(N)} G_{20}^{(N)} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & G_{12}^{(N)} G_{20}^{(N)} & 0 & 0 & \ldots & 0 \\ & & & \vdots & & & & \\ 0 & 0 & \ldots & G_{1,n-k+1}^{(N)} G_{20}^{(N)} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & (G_{20}^{(N)})^2 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & G_{21}^{(N)} G_{20}^{(N)} & 0 & 0 & \ldots & 0 \\ & & & \vdots & & & & \\ 0 & 0 & \ldots & G_{2,n-k+1}^{(N)} G_{20}^{(N)} & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Now equation (2.37) reduces to a system of equations

$$G_{00}^{(N)} = ((\lambda + N\theta)I_m - D_0)^{-1}[(\lambda I_m \otimes \beta)G_{11}^{(N)} + [D_1 \otimes \beta]G_{20}^{(N)} + N\theta I_m \otimes \beta]$$

$$G_{11}^{(N)} = [\lambda I_{mm_1} - (D_0 \oplus S)]^{-1}[(D_1 \otimes I_{m_1})G_{11}^{(N)}G_{20}^{(N)} + (I_m \otimes S^0)G_{00}^{(N)} + \lambda I_{mm_1}G_{12}^{(N)}]$$

$$G_{1j}^{(N)} = [\lambda I_{mm_1} - (D_0 \oplus S)]^{-1}\left[(D_1 \otimes I_{m_1})G_{1j}^{(N)}G_{20}^{(N)} + [(I_m \otimes (S^0\beta)]G_{1j-1}^{(N)} + \lambda I_{mm_1}G_{1j+1}^{(N)}\right], 2 \le j \le n-k$$

$$G_{1n-k+1}^{(N)} = -(D_0 \oplus S)^{-1}\left[(D_1 \otimes I_{m_1})G_{1n-k+1}^{(N)}G_{20}^{(N)} + [(I_m \otimes (S^0\beta)]G_{1n-k}^{(N)}\right]$$

$$G_{2,0}^{(N)} = [\lambda I_{mm_1} - (D_0 \oplus S)]^{-1}\left[(D_1 \otimes I_{m_1})(G_{20}^{(N)})^2 + (I_m \otimes S^0)G_{00}^{(N)} + \lambda I_{mm_1}G_{21}^{(N)}\right]$$

$$G_{2,j}^{(N)} = [\lambda I_{mm_1} - (D_0 \oplus S)]^{-1}\left\{(D_1 \otimes I_{m_1})G_{2j}^{(N)}G_{20}^{(N)} + ((I_m \otimes (S^0\beta))G_{1j}^{(N)} + \lambda I_{mm_1}G_{2j+1}^{(N)}\right\} \quad 1 \le j \le n-k$$

$$G_{2,n-k+1}^{(N)} = -(D_0 \oplus S)^{-1}\left[(D_1 \otimes I_{m_1})G_{2,n-k+1}^{(N)}G_{20}^{(N)} + [(I_m \otimes (S^0\beta)]G_{1n-k+1}^{(N)}\right]$$

Now we can use Block Gauss-Seidel iterative procedure to evaluate G_N, and then R_N can be evaluated using the formula

$$R_N = A_0(-A_{1N} - A_0G_N)^{-1}$$

2.9. System performance measures

(1) Mean number of customers in the orbit:

$$\mathcal{N}_{\text{orbit}} = (\sum_{i=1}^{\infty} i x_N(i))\mathbf{e} = p_1 \mathbf{e}$$

where

$$p_1 = x_N(0)(\sum_{i=1}^{N-1} i(\prod_{j=1}^{i} W_j)) + N x_N(0)(\prod_{j=1}^{N-1} W_j) R_N (I - R_N)^{-1}$$
$$+ x_N(0)(\prod_{j=1}^{N-1} W_j) R_N^2 (I - R_N)^{-2}$$

(2) The overall rate of retrials

$$\theta_1^* = \theta \mathcal{N}_{\text{orbit}}$$

(3) The successful rate of retrials

$$\theta_2^* = (\sum_{i=1}^{\infty} i x_N(i))\mathbf{e}_\theta = p_1 \mathbf{e}_\theta$$

where $\mathbf{e}_\theta$ is a column vector whose first m entries are 1s and all other entries are 0s.

(4) The fraction of successful rate of retrials

$$\theta_3^* = \frac{\theta_2^*}{\theta_1^*}$$

(5) The probability that the server is busy

$$P_{\text{busy}} = (\sum_{i=0}^{\infty} x_N(i))\mathbf{e}_b = p_2 \mathbf{e}_b$$

where

$$p_2 = x_N(0)(I + \sum_{i=1}^{N-2}(\prod_{j=1}^{i} W_j) + (\prod_{j=1}^{N-1} W_j)(I - R_N)^{-1})$$

(6) Probability that the system is down = Probability that the number of failed components equal to $n - k + 1$

$$P_{\text{down}} = (\sum_{i=0}^{\infty} x_N(i))\mathbf{e}_c = p_2\mathbf{e}_c$$

where $\mathbf{e}_c$ is a column vector given by

$$\mathbf{e}_c = \begin{bmatrix} 0_{(m+(n-k)mm_1)\times 1} \\ \mathbf{e}_{mm_1} \\ 0_{((n-k+1)mm_1)\times 1} \\ \mathbf{e}_{mm_1} \end{bmatrix}$$

(7) Expected number of failed components in the system,

$$\mathcal{N}_{\text{comp}} = (\sum_{i=0}^{\infty} x_N(i))\mathbf{e}_d = p_2\mathbf{e}_d$$

where $\mathbf{e}_d$ is a column vector given by

$$\mathbf{e}_d = \begin{bmatrix} 0_{m\times 1} \\ \mathbf{e}_f \\ 0_{mm_1\times 1} \\ \mathbf{e}_f \end{bmatrix} \quad \text{with } \mathbf{e}_f = \begin{bmatrix} 1\mathbf{e}_{mm_1} \\ 2\mathbf{e}_{mm_1} \\ \vdots \\ (n-k+1)\mathbf{e}_{mm_1} \end{bmatrix}$$

(8) $P_{bcom} = \left(\sum_{i=0}^{\infty} x_N(i)\right) \mathbf{e}_g = P_2 \mathbf{e}_g$
where the column vector $\mathbf{e}_g$ is given by

$$\mathbf{e}_g = \begin{bmatrix} 0_{m \times 1} \\ \mathbf{e}_{(n-k+1)mm_1} \\ 0 \end{bmatrix}$$

(9) $P_{bext} = \left(\sum_{i=0}^{\infty} x_N(i)\right) \mathbf{e}_h = P_2 \mathbf{e}_h$
where the column vector $\mathbf{e}_h$ is given by

$$\mathbf{e}_h = \begin{bmatrix} 0 \\ \mathbf{e}_{(n-k+2)mm_1} \end{bmatrix}$$

2.10. Numerical illustration

2.10.1. Effect of variation of component failure rate λ. $n = 18$, $k = 6$, $D_0 = \begin{bmatrix} -5.5 & 3.5 \\ 1.0 & -3.5 \end{bmatrix}$, $D_1 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.5 \end{bmatrix}$, $S = \begin{bmatrix} -7.5 & 2.0 \\ 2.1 & -7.7 \end{bmatrix}$,
$S^0 = \begin{bmatrix} 5.5 \\ 5.6 \end{bmatrix}$, $\beta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$,
average service rate=5.549, average arrival rate= 2.346,
correlation between two interarrival times= -0.00029

Table 1 shows that as λ increases the mean number of customers in the orbit, the mean number of failed components, and the fraction of time the server is busy with a failed component increases. But th increase of λ has not much effect on the fraction of time the server is busy with an external customer. Table 1 also shows that the increase in the retrial rate has a considerable effect only on the mean number of customers and in the fraction of successful rate of retrials.

TABLE 1

	λ	0.1	0.5	1.0	1.5	2.0	2.5
	$\mathcal{N}_{\text{orbit}}$	0.7019	0.9749	1.4858	2.3553	4.065	8.5052
	$\mathcal{N}_{\text{comp}}$	0.0185	0.1028	0.2366	0.4128	0.6494	0.9757
$\theta = 5.0$	$\mathcal{P}_{\text{bcom}}$	0.0180	0.0901	0.1802	0.2703	0.3604	0.4505
	$\mathcal{P}_{\text{bext}}$	0.4228	0.4228	0.4228	0.4228	0.4228	0.4228
	θ_3^*	0.2928	0.2457	0.1898	0.1378	0.0903	0.0482
	$\mathcal{N}_{\text{orbit}}$	0.5182	0.7291	1.1307	1.8266	3.2185	6.8929
	$\mathcal{N}_{\text{comp}}$	0.0185	0.1028	0.2366	0.4128	0.6494	0.9757
$\theta = 10.0$	$\mathcal{P}_{\text{bcom}}$	0.0180	0.0901	0.1802	0.2703	0.3604	0.4505
	$\mathcal{P}_{\text{bext}}$	0.4228	0.4228	0.4228	0.4228	0.4228	0.4228
	θ_3^*	0.1982	0.1642	0.1247	0.0888	0.0577	0.0298

2.10.2. Effects of correlation. $n = 18, k = 6, D_0 = \begin{bmatrix} -4.05 & 1.55 \\ 3.5 & -5.5 \end{bmatrix}$

$D_1 = \begin{bmatrix} 2.05 & 0.45 \\ 1.0 & 1.0 \end{bmatrix}$ $S = \begin{bmatrix} -7.5 & 2.0 \\ 2.1 & -7.7 \end{bmatrix}$ $S_0 = \begin{bmatrix} 5.5 & 5.6 \\ 2.1 & -7.7 \end{bmatrix}$ $\beta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$, average arrival rate = 2.346, correlation between two interretrial times =0.00029. The other parameters are same as in the case of Table 1

TABLE 2

	λ	0.1	0.5	1.0	1.5	2.0	2.5
	$\mathcal{N}_{\text{orbit}}$	0.7121	0.9865	1.4998	2.3731	4.09	8.5452
	$\mathcal{N}_{\text{comp}}$	0.0185	0.1028	0.2366	0.4128	0.6494	0.9757
$\theta = 5.0$	$\mathcal{P}_{\text{bcom}}$	0.0180	0.0901	0.1802	0.2703	0.3604	0.4505
	$\mathcal{P}_{\text{bext}}$	0.4228	0.4228	0.4228	0.4228	0.4228	0.4228
	θ_3^*	0.2911	0.2444	0.1898	0.1371	0.0899	0.048
	$\mathcal{N}_{\text{orbit}}$	0.5267	0.739	1.1431	1.8429	3.242	6.929
	$\mathcal{N}_{\text{comp}}$	0.0185	0.1028	0.2366	0.4128	0.6494	0.9757
$\theta = 10.0$	$\mathcal{P}_{\text{bcom}}$	0.0180	0.0901	0.1802	0.2703	0.3604	0.4505
	$\mathcal{P}_{\text{bext}}$	0.4228	0.4228	0.4228	0.4228	0.4228	0.4228
	θ_3^*	0.1968	0.1631	0.1239	0.0883	0.0567	0.0296

TABLE 3

	θ	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$	$\lambda = 2.5$
$\mathcal{P}_{down}$	0	$< 10^{-13}$	$< 10^{-13}$	$.174 \times 10^{-9}$	$.3 \times 10^{-7}$	$.111 \times 10^{-5}$	$.173\times^{-4}$
	5.0	$< 10^{-13}$	$.1 \times 10^{-12}$	$.533 \times 10^{-9}$	$.749 \times 10^{-7}$	$.238 \times 10^{-5}$	$.334 \times 10^{-4}$
$\mathcal{P}_{busy}$	0	0.018	0.0901	0.1802	0.2703	0.3604	0.4505
	5.0	0.4408	0.5129	0.603	0.6931	0.7832	0.8733

TABLE 4. $\text{ID}_{cost} = C_{11} * \mathcal{P}_{down} - C_{12} * \mathcal{P}_{busy}$

			$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$	$\lambda = 2.5$
ID_{cost}	$C_{11} = 100$	$\theta = 0$	-0.1800	-0.9010	-1.8020	-2.7030	-3.6039	-4.5033
	$C_{12} = 10$	$\theta = 5.0$	-4.4080	-5.1290	-6.0300	-6.9310	-7.8318	-8.7297
	$C_{11} = 1000$	$\theta = 0$	-0.1800	-0.9010	-1.8020	-2.7030	-3.6029	-4.4877
	$C_{12} = 10$	$\theta = 5.0$	-4.4080	-5.1290	-6.0300	-6.9309	-7.8296	-8.6996
	$C_{11} = 10000$	$\theta = 0$	-0.1800	-0.9010	-1.8020	-2.7027	-3.5929	-4.3320
	$C_{12} = 10$	$\theta = 5.0$	-4.4080	-5.1290	-6.0300	-6.9303	-7.8082	-8.3990

Tables 1 and 2 shows that when correlation between arrivals becomes positive, there is an increase in the mean number of customers in the orbit and there is a decrease in the rate of successful retrials. The other parameters are same as in the case of Table 1.

The above table shows that there is only negligible increase in the system down probability if we provide service to external customers in a k-out-of-n system as described earlier in this chapter, but there has been a considerable increase in the fraction of time the server is found busy. To make these statements more clear, we consider a cost function

$$\text{ID}_{\text{cost}} = C_{11} \cdot P_{\text{down}} - C_{12} \cdot P_{\text{busy}}$$

where C_{11} is the cost per unit time due to the system becoming down and C_{12} is the profit per unit time obtained by making the server busy.

From table 4 we note that even when C_{11} is 1000 times bigger than C_{12} and the component failure rate $\lambda = 2.5$, the function ID_{cost} as a lesser value when $\theta = 5.0$ than when $\theta = 0$ which shows that our goal of ideal time utilization is achieved, atleast numerically.

CHAPTER 3

Maximization of reliability of a k-out-of-n system with repair by a facility attending external customers in a retrial queue*

In this chapter, we study a k-out-of-n system with single server who provides service to external customers also as described in the following paragraphs.

The system consists of two parts:(i) a main queue consisting of customers (failed components of the k-out-of-n system) and (ii) a pool (of finite capacity M) of external customers together with an orbit for external customers who find the pool full. An external customer who finds the pool full on arrival joins the orbit with probability γ and with probability $1 - \gamma$ leave the system forever. An orbital customer, who finds the pool full, at an epoch of repeated attempt, returns to orbit with probability δ (< 1) and with probability 1 $-\delta$ leaves the system forever.

The arrival process : Arrival of main customers have interoccurence time exponentially distributed with parameter λ_i when the number of operational components of the k-out-of-n system is i. By taking $\lambda_i = \frac{\lambda}{i}$ we notice that the cumulative failure rate is a constant λ. We assume that the k-out-of-n system is COLD (components fail only when system is operational). The case of WARM and HOT system can be studied on the same lines (see Krishnamoorthy and Ushakumari [37]). External customers arrive according to a Markovian Arrival Process (MAP) with representation (D_0, D_1) where D_0 and D_1 are assumed to be matrices of order m. Fundamental arrival rate $\lambda_g = -\pi D_0 e$

* This chapter was published in the Proceedings of the V-th International Workshop on Retrial Queues, Korea, September, 2004, Ed. B. D. Choi, pp. 31–38

The service process : Service to the failed components of the main system is governed by the N-policy. That is each epoch the system starts with all components operational (*ie.*, all n components are in operation), the server starts attending one by one the customers from the pool (if there is any). The moment the number of failed components of the main system reaches N, no more customer from the pool is taken for service until there is no components of the main system waiting for repair. However service of the external customer, if there is any, will not be disrupted even when N components accumulate in the main queue (that is the external customer in service will not get pre-empted on realization of the event that N components of the main system failed and got accumulated; instead the moment the service of the present external customer is completed, the server is switched to the service of main customers).

Service time of main customers follow PH distribution or order n_1 and representation (α, S_1) and that of external customers have PH distribution of order n_2 with representation (β, S_2);

S_1^0 and S_2^0 are such that $S_i\mathbf{e} + s_i^0 = 0$, $i = 1, 2$ where $\mathbf{e}$ is column vector of ones. The two service times are independent of each other and also independent of the failure of components of the main system as well as the arrival of external customers.

Objective : To utilize server idle time without affecting the system reliability.

Krishnamoorthy and Ushakumari [**37**] deals with the study of the reliability of a k-out-of-n system with repairs by server in a retrial queue. They do not give any priority to the failed components of the main system nor do they investigate any control policy. Krishnamoorthy, Ushakumari and Lakshmi [**38**] introduced the repair of failed components of a k-out-of-n system under the N-policy. For further details one may refer to the paper and references therein as well as Ushakumari and Krishnamoorthy [**59**] Bocharov *et al* [**10**] examine an $M/G/1/r$ retrial queue with priority of primary customers. They obtain the stationary distribution of the primary queue size, an algorithm for the factorial moments of the number of retrial customers and an expression for the expected number of customers in the system. Nevertheless, we wish to emphasise that their paper does not distinguish between the priority and ordinary customers. This is distinctly done in this chapter (our priority customers are the failed components of the k-out-of-n system):

This chapter differs from chapter 2 mainly by the fact that here together with the orbit, we also consider an intermediate pool of finite capacity to which external customers join after seeing a busy server on arrival or after a successful retrial from the orbit. We expect that this intermediate pool from which an external customer can be selected for service, whenever the server becomes idle, will help us to decrease the server idle time.

The steady state distribution is derived in this chapter. Note that the non-persistence of orbital customers together with the fact that an external customer, finding the pool full, may not join the pool ensures that even under very heavy traffic the system can attain stability. Several performance measures are obtained.
One can refer Deepak, Joshua, and Krishnamoorthy [20] for a detailed analysis of queues with pooled customers (postponed work).

3.1. Modelling and analysis

The following notations are used in the equal:
$N_1(t)$ = # orbital customers at time t
$N_2(t)$ = # customers in the pool (including the one getting service, if any,) at time t.
$N_3(t)$ = # failed components (including the one under repair, if any) at time t

$$N_4(t) = \begin{cases} 0 & \text{if the server is idle} \\ 1 & \text{if the server is busy with repair} \\ & \text{of a failed component of the main system} \\ 2 & \text{if the server is attending an external customer at time } t. \end{cases}$$

$N_5(t)$ = Phase of the arrival process,

$$N_6(t) = \begin{cases} \text{Phase of service of the customer, if any, in service at } t \\ 0, \quad \text{if no service is going on at time } t. \end{cases}$$

It follows that $\{X(t) : t \geq 0\}$ where

$$X(t) = (N_1(t), N_2(t), N_3(t), N_4(t), N_5(t), N_6(t))$$

is a continuous time Markov chain on the state space

$$\begin{aligned} S = & \{(j_1, 0, j_3, 0, j_5, 0) | j_1 \geq 0;\ 0 \leq j_3 \leq N-1;\ 1 \leq j_5 \leq m\} \\ & \cup \{(j_1, j_2, j_3, 1, j_5, j_6) | j_1 \geq 0,\ 0 \leq j_2 \leq M;\ 1 \leq j_3 \leq n-k+1; \\ & \quad 1 \leq j_5 \leq m; 1 \leq j_6 \leq n_1\} \\ & \cup \{(j_1, j_2, j_3, 2, j_5, j_6) | j_1 \geq 0;\ 1 \leq j_2 \leq M; \\ & \quad 0 \leq j_3 \leq n-k+1; 1 \leq j_5 \leq m;\ 1 \leq j_6 \leq n_2\} \end{aligned}$$

Arranging the states lexicographically, and then partitioning the state space into levels $\underline{i}$, where each level i correspond to the collection of states with i customers in the orbit, we get the infinitesimal generator of the above chain as

$$Q = \begin{bmatrix} A_{10} & A_0 & 0 & 0 \ldots \\ A_{21} & A_{11} & A_0 & 0 \ldots \\ 0 & A_{22} & A_{12} & A_0 \ldots \\ \vdots & \vdots & \vdots & \end{bmatrix}.$$

where

$$A_{10} = \begin{bmatrix} W_0 & W_5 & & & & & \\ W_3 & W_1 & W_6 & & & & \\ & W_4 & W_1 & W_6 & & & \\ & & & & & & \\ & & & & W_4 & W_1 & W_6 \\ & & & & & W_4 & W_2 \end{bmatrix}$$

where

$$W_0 = \begin{bmatrix} B_0 & B_8 & & & & & & & & & & \\ B_4 & B_1 & B_9 & & & & & & & & & \\ & B_5 & B_1 & B_9 & & & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & & & \\ & & & B_5 & B_1 & B_9 & & & & & & \\ & & & & B_5 & B_1 & B_{10} & & & & & \\ & & & & & B_6 & B_2 & B_{11} & & & & \\ & & & & & & B_7 & B_2 & B_{11} & & & \\ & & & & & & & & \ddots & & & \\ & & & & & & & & & & B_2 & B_{11} \\ & & & & & & & & & & B_7 & B_3 \end{bmatrix}$$

$$B_0 = D_0 - \lambda I_m,\ B_1 = \begin{bmatrix} D_0 - \lambda I_m & 0 \\ 0 & D_0 \oplus S_1 - \lambda I_{mn_1} \end{bmatrix}$$

$$B_2 = D_0 \oplus S_1 - \lambda I_{mn_1},\ B_3 = D_0 \oplus S_1$$

$$B_4 = \begin{bmatrix} 0 \\ I_m \otimes S_1^0 \end{bmatrix},\ B_5 = \begin{bmatrix} 0 & 0 \\ 0 & I_m \otimes (S_1^0\alpha) \end{bmatrix},\ B_6 = \begin{bmatrix} 0 & I_m \otimes (S_1^0\alpha) \end{bmatrix}$$

$$B_7 = I_m \otimes (S_1^0\alpha),\ B_8 = \begin{bmatrix} \lambda I_m & 0 \end{bmatrix},\ B_9 = \lambda I_{m+mn_1}$$

$$B_{10} = \begin{bmatrix} I_m \otimes (\lambda\alpha) \\ \lambda I_{mn_1} \end{bmatrix},\ B_{11} = \lambda I_{mn_1}$$

$$W_1 = \begin{bmatrix} C_0 & C5 & & & & \\ C_3 & C_1 & C_6 & & & \\ & C_4 & C_1 & & & \\ & & & \ddots & & \\ & & & & C_1 & C_6 \\ & & & & C_4 & C_2 \end{bmatrix}$$

$$C_0 = D_0 \oplus S_2 - \lambda I_{mn_2}$$

$$C_1 = C_2 - \lambda I_{m(n_1+n_2)}$$

$$C_2 = \begin{bmatrix} D_0 \oplus S_1 & 0 \\ 0 & D_0 \oplus S_2 \end{bmatrix}, C_3 = \begin{bmatrix} I_m \otimes (S_1^0 \beta) \\ 0 \end{bmatrix}$$

$$C_4 = \begin{bmatrix} I_m \otimes (S_1^0 \alpha) & 0 \\ 0 & 0 \end{bmatrix}, C_5 = \begin{bmatrix} 0 & \lambda I_{mn_2} \end{bmatrix}, C_6 = \lambda I_{m(n_1+n_2)}$$

$$W_2 = W_1 + \bar{W}_1$$

where, $$\bar{W}_1 = \begin{bmatrix} (1-\gamma)(D_1 \otimes I_{n_2}) & 0 \\ 0 & I_{n-k+1} \otimes \bar{\bar{W}}_1 \end{bmatrix}$$

with $$\bar{\bar{W}}_1 = \begin{bmatrix} (1-\gamma)(D_1 \otimes I_{n_1}) & 0 \\ 0 & (1-\gamma)(D_1 \otimes I_{n_2}) \end{bmatrix}$$

$$W_3 = \begin{bmatrix} W_{30} & 0 & 0 \\ 0 & I_{N-1} \otimes W_{31} & 0 \\ 0 & 0 & I_{n-k-N+2} \otimes W_{32} \end{bmatrix}$$

where

$$W_{30} = I_m \otimes S_2^0, \quad W_{31} = \begin{bmatrix} 0 & 0 \\ I_m \otimes S_2^0 & 0 \end{bmatrix}_{m(n_1+n_2)\times m(n_1+n_2)}$$

$$W_{32} = \begin{bmatrix} 0 \\ I_m \otimes (S_2^o \alpha) \end{bmatrix}_{m(n_1+n_2)\times mn_1}$$

$$W_4 = \begin{bmatrix} E_0 & 0 & 0 \\ 0 & I_{N-1} \otimes E_1 & 0 \\ 0 & 0 & I_{n-k-N+2} \otimes E_2 \end{bmatrix}$$

$$E_0 = I_m \otimes (S_2^0 \beta), E_1 = \begin{bmatrix} 0 & 0 \\ 0 & I_m \otimes (S_2^0 \beta) \end{bmatrix}_{m(n_1+n_2)\times m(n_1+n_2)}.$$

$$E_2 = \begin{bmatrix} 0 & 0 \\ I_m \otimes (S_2^o \alpha) & 0 \end{bmatrix}$$

$$W_5 = \begin{bmatrix} F_0 & 0 & 0 \\ 0 & F_1 & 0 \\ 0 & 0 & F_2 \end{bmatrix}$$

$$F_0 = D_1 \otimes \beta, \quad F_1 = I_{N-1} \otimes F_1', \quad F_1' = \begin{bmatrix} 0 & D_1 \otimes \beta \\ D_1 \otimes I_{n_1} & 0 \end{bmatrix}$$

$$F_2 = I_{n-k+2-N} \otimes F_2', \quad F_2' = \begin{bmatrix} D_1 \otimes I_{n_1} & 0 \end{bmatrix}$$

$$W_6 = \begin{bmatrix} H_0 & 0 \\ 0 & I_{n-k+1} \otimes H_1 \end{bmatrix}$$

$$H_0 = D_1 \otimes I_{n_2}, \quad H_1 = \begin{bmatrix} D_1 \otimes I_{n_1} & 0 \\ 0 & D_1 \otimes I_{n_2} \end{bmatrix}$$

and

$$A_{1i} = A_{10} - \tilde{A}_{1i} \quad \text{for } i \geq 1$$

where

$$\tilde{A}_{1i} = \begin{bmatrix} i\theta I_{L_2} & 0 \\ 0 & i\theta(1-\delta) I_{L_1} \end{bmatrix}.$$

Where

$$L_1 = (n-k+2)mn_2 + (n-k+1)mn_1$$

$$L_2 = Nm + (n-k+1)mn_1 + (M-1)L_1$$

$$A_{2i} = \begin{bmatrix} 0 & Z_i & 0 \\ 0 & 0 & i\theta I_{(M-1)L_1} \\ 0 & 0 & i\theta(1-\delta)I_{L_1} \end{bmatrix}, \quad i \geq 1$$

$$Z_i = \begin{bmatrix} Z_{1i} & 0 & 0 \\ 0 & I_{N-1} \otimes Z_{2i} & 0 \\ 0 & 0 & I_{(n-k-N+2)} \otimes Z_{3i} \end{bmatrix}, \quad Z_{1i} = I_m \otimes (i\theta\beta)$$

$$Z_{2i} = \begin{bmatrix} 0 & I_m \otimes (i\theta\beta) \\ i\theta I_{mn_1} & 0 \end{bmatrix}, \quad Z_{3i} = \begin{bmatrix} i\theta I_{mn_1} & 0 \end{bmatrix}$$

$$A_0 = \begin{bmatrix} 0 & 0 \\ 0 & \bar{A}_0 \end{bmatrix}$$

$$\bar{A}_0 = \begin{bmatrix} (\gamma D_1) \otimes I_{n_2} & 0 \\ 0 & I_{n-k+1} \otimes \bar{A}_0^{(1)} \end{bmatrix}, \bar{A}_0^{(1)} = \begin{bmatrix} (\gamma D_1) \otimes I_{n_1} & 0 \\ 0 & (\gamma D_1) \otimes I_{n_2} \end{bmatrix}$$

3.2. System stability

THEOREM 3.1. *The assumption that after each retrial a customer may leave the system with probability* $1 - \delta$ *makes the system stable irrespective of the parameter values.*

PROOF. To prove the theorem we use a result due to Tweedie [57]. For the model under consideration we consider the following Lyapunov function:

$$\phi(s) = i \text{ if } s \text{ is a state belonging to level } i$$

The mean drift y_s for an s belonging to level $i \geq 1$ is given by

$$
\begin{aligned}
y_s &= \sum_{p \neq s} q_{sp}(\phi(p) - \phi(s)) \\
&= \sum_{s'} q_{ss'}(\phi(s') - \phi(s)) + \sum_{s''} q_{ss''}\Big(\phi(s'') - \phi(s)\Big) \\
&+ \sum_{s'''} q_{ss'''}\Big(\phi(s''') - \phi(s)\Big)
\end{aligned}
$$

where s', s'', s''' varies over the states belonging to levels $i-1, i, i+1$ respectively. Then by definition of ϕ, $\phi(s) = i$, $\phi(s') = i - 1$, $\phi(s'') = i$, $\phi(s''') = i + 1$

So that

$$
y_s = -\sum_{s'} q_{ss'} + \sum_{s'''} q_{ss'''}
$$

$$
y_s = \begin{cases} -i\theta + \sum_{s'''} q_{ss'''}, & \text{if } s \in I_i \\ -i\theta(1-\delta) + \sum_{s'''} q_{ss'''}, & \text{if } s \in \bar{I}_i \end{cases}
$$

where I_i denotes the collection of states in level i which corresponds to $N_2(t) < M$, and $\bar{I}_i$ denotes the collection of states in level i which correspond to $N_2(t) = M$.

We note that $\sum_{s'''} q_{ss'''}$ is bounded by some fixed constant for any s in any level $i \geq 1$. So, let $\sum_{s'''} q_{ss'''} < \kappa$, for some real number $\kappa > 0$, for all states s belonging to level $i \geq 1$. Also since $1 - \delta > 0$, for any $\epsilon > 0$, we can find N' large enough that $y_s < -\epsilon$ for any s belonging to level $i \geq N'$.

Hence by Tweedie's result, the theorem follows. □

3.3. Steady state distribution

Since the process under consideration is an $LDQBD$, to calculate the steady state distribution, we use the methods described in Bright and Taylor [**13**].

By partitioning the steady state vector x as $\mathrm{x} = (x_0, x_1, x_2, \ldots)$ we can write

$$x_k = x_0 \prod_{l=0}^{k-1} R_l \quad \text{for } k \geq 1$$

where the family of matrices $\{R_k, k \geq 0\}$ are minimal non-negative solutions to the system of equations:

$$A_0 + R_k A_{1\,k+1} + R_k[R_{k+1} A_{2\,k+2}] = 0, \; k \geq 0 \tag{3.1}$$

x_0 is calculated by solving

$$x_0[A_{10} + R_0 A_{21}] = 0 \tag{3.2}$$

such that

$$x_0 \mathrm{e} + x_0 \sum_{k=1}^{\infty} [\prod_{l=0}^{k-1} R_l] \mathrm{e} < \infty \tag{3.3}$$

The calculation of the above infinite sums does not seem to be practical, so we approximate x_ks by $x_k(K^*)$s where $(x_k(K^*))_j$, $0 \leq k \leq K^*$, is defined as the stationary probability that $X(t)$ is in the j^{th} state of level k, conditional on $X(t)$ being in level i, $0 \leq i \leq K^*$.

Then $x_k(K^*)$, $0 \leq k \leq K^*$ is given by

$$x_k(K^*) = x_0(K^*) \prod_{l=0}^{k-1} R_l \tag{3.4}$$

where $x_0(K^*)$ satisfies (3.2) and

$$x_0(K^*)\mathrm{e} + x_0(K^*) \Big[\sum_{k=1}^{K^*} [\prod_{l=0}^{k-1} R_l] \Big] \mathrm{e} = 1 \tag{3.5}$$

Here we have that for all $i \geq 1$, and for all k, there exists j such that $[A_{2i}]_{k,j} > 0$. So we can construct a dominating process $\bar{X}(t)$ of $X(t)$ and can use it to find the truncation level

K^* in the same way as in [13], as follows. The dominating process $\bar{X}(t)$ has generator

$$\bar{Q} = \begin{bmatrix} A_{10} & A_0 & 0 & 0 & 0 & \dots \\ 0 & \bar{A}_{11} & \bar{A}_0 & 0 & 0 & \dots \\ 0 & \bar{A}_{22} & \bar{A}_{12} & \bar{A}_0 & 0 & \dots \\ 0 & 0 & \bar{A}_{23} & \bar{A}_{13} & \bar{A}_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

where
$(\bar{A}_0)_{i,j} = \frac{1}{C}[(A_0 e)_{\max}]$, $(\bar{A}_{2k})_{i,j} = \frac{1}{C}\Big((A_{2,k-1})\mathrm{e}\Big)_{\min}$ for $k \geq 2$, $(\bar{A}_{1k})_{ij} = (A_{1k})_{ij}$, $j \neq i$, $k \geq 1$; and $C = Nm + (M+1)(n-k+1)mn_1 + M(n-k+2)mn_2$ is the dimension of a level $i \geq 1$.

3.4. Performance measures

We partition the steady state vector $\mathbf{x}$ as $\mathbf{x} = (x_0, x_1, x_2, \dots)$ where the sub-vectors x_{j_1}s are again partitioned as $x_{j_1} = x(j_1, j_2, j_3, j_4)$ which correspond to $N_i(t) = j_i$, $1 \leq i \leq 4$

(1) Fraction of time the system is down is given by

$$\mathcal{P}_{\text{down}} = \sum_{j_1=0}^{K^*} \sum_{j_2=0}^{M} \sum_{j_4=1}^{2} x(j_1, j_2, n-k+1, j_4)\mathrm{e}.$$

(2) System reliability, defined as the probability that atleast k components are operational, $\mathcal{P}_{\text{rel}}$ is given by

$$\mathcal{P}_{\text{rel}} = 1 - \mathcal{P}_{\text{down}}.$$

(3) Average no. of external units waiting in the pool is given by

$$\mathcal{N}_{\text{pool}} = \sum_{j_2=1}^{M} j_2 (\sum_{j_1=0}^{K^*} \sum_{j_3=1}^{n-k+1} x(j_1, j_2, j_3, 1)\mathrm{e}$$

$$+ \sum_{j_2=2}^{M} (j_2 - 1) \sum_{j_1=0}^{K^*} \sum_{j_3=0}^{n-k+1} x(j_1, j_2, j_3, 2)\mathrm{e}$$

(4) Average no. of external units in the orbit is given by

$$\mathcal{N}_{\text{orbit}} = \sum_{j_1=1}^{K^*} j_1 [x(j_1)\mathrm{e}]$$

(5) Average no. of failed components is given by

$$\mathcal{N}_{\text{faic}} = \sum_{j_3=1}^{n-k+1} j_3 (\sum_{j_1=0}^{K^*} \sum_{j_2=1}^{M} x(j_1, j_2, j_3, 2)\mathrm{e}$$

$$+ \sum_{j_1=0}^{K^*} \sum_{j_2=0}^{M} x(j_1, j_2, j_3, 1)\mathrm{e}) + \sum_{j_3=1}^{N-1} j_3 \sum_{j_1=0}^{K^*} x(j_1, 0, j_3, 0)\mathrm{e}$$

(6) The probability that an external unit, on its arrival joins the queue in the pool is given by

$$\mathcal{P}_{\text{queue}} = \frac{1}{\lambda_g} \Big\{ \sum_{j_1=0}^{K^*} \sum_{j_2=1}^{M-1} \sum_{j_3=1}^{n-k+1} \sum_{j_4=1}^{2} x(j_1, j_2, j_3, j_4)[D_1 \otimes I_{n_{j_4}}]\mathrm{e}$$

$$+ \sum_{j_1=0}^{K^*} \sum_{j_3=1}^{n-k+1} x(j_1, 0, j_3, 1)(D_1 \otimes I_{n_1})\mathrm{e} \Big\}$$

(7) The probability that an external unit, on its arrival gets service directly is given by

$$\mathcal{P}_{\text{ds}} = \frac{1}{\lambda_g} \Big\{ \sum_{j_1=0}^{K^*} \sum_{j_3=0}^{N-1} x(j_1, 0, j_3, 0) D_1 \mathrm{e} \Big\}$$

(8) The probability that an external unit, on its arrival enters orbit is given by

$$\mathcal{P}_{\text{orbit}} = \frac{1}{\lambda_g}\Big\{\sum_{i=0}^{K^*} x(i)A_0\text{e}\Big\}$$

(9) Fraction of time the server is busy with external customers is given by

$$\mathcal{P}_{\text{exbusy}} = \sum_{j_1=0}^{K^*}\sum_{j_2=1}^{M}\sum_{j_3=0}^{n-k+1} x(j_1, j_2, j_3, 2)\text{e}$$

(10) Probability that the server is found idle is given by

$$\mathcal{P}_{\text{idle}} = \sum_{j_1=0}^{K^*}\sum_{j_2=0}^{N-1} x(j_1, 0, j_2, 0)$$

(11) Probability that the server is found busy is given by

$$\mathcal{P}_{\text{busy}} = 1 - \mathcal{P}_{\text{idle}}$$

(12) Expected loss rate of external customers is given by

$$\begin{aligned}\lambda_{\text{loss}} &= \sum_{j_1=0}^{K^*}\sum_{j_2=1}^{n-k+1} x(j_1, M, j_2, 1)(1-\gamma)(D_1 \otimes I_{n_1})\text{e} \\ &+ \sum_{j_1=0}^{K^*}\sum_{j_2=0}^{n-k+1} x(j_1, M, j_2, 2)(1-\gamma)(D_1 \otimes I_{n_2})\text{e} \\ &+ \sum_{j_1=1}^{K^*}\sum_{j_2=1}^{n-k+1} (1-\delta)j_1\theta x(j_1, M, j_2, 1)\text{e} \\ &+ \sum_{j_1=1}^{K^*}\sum_{j_2=0}^{n-k+1} (1-\delta)j_1\theta x(j_1, M, j_2, 2)\text{e}\end{aligned}$$

(13) We construct a cost function as where C_1 is the holding cost per unit time per customer waiting in the pool, C_2 is the loss per unit time due to the system becoming down, C_3 is the loss per unit time due to a customer leaves the system without

taking service, C_4 is the holding cost per unit time per failed component in the system, C_5 is the loss per unit time due to the server becoming idle and C_6 is the profit per unit time due to the server becoming busy with an external customer.

3.5. Numerical illustration

Set $\theta = 15.0$, $\lambda = 1.0$, $\gamma = 0.7$, $\delta = 0.7$, $n = 11$, $k = 4$, $M = 5$, $N = 4$

$$S_1 = \begin{bmatrix} -6.5 & 4.0 \\ 1.5 & -4.5 \end{bmatrix} \quad S_2 = \begin{bmatrix} -5.06 & 2.06 \\ 4.0 & -6.5 \end{bmatrix} \quad S_1^0 = \begin{bmatrix} 2.5 \\ 3.0 \end{bmatrix} \quad S_2^0 = \begin{bmatrix} 3.0 \\ 2.5 \end{bmatrix} \quad \alpha = (0.5, 0.5)$$

$\beta = (0.5, 0.5)$

$C_1 = 10.0$, $C_2 = 1500.0$, $C_3 = 100.0$, $C_4 = 20.0$, $C_5 = 50.0$, $C_6 = 200.0$.

Effect of correlation : The additional parameters for table 1 are the following

$$D_0 = \begin{bmatrix} -5.5 & 3.5 \\ 1.0 & -3.5 \end{bmatrix} \quad D_1 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.5 \end{bmatrix} \qquad \text{(A1)}$$

average arrival rate = 2.34615, correlation =-0.00029

$$D_0 = \begin{bmatrix} -4.05 & 1.55 \\ 3.5 & -5.5 \end{bmatrix} \quad D_1 = \begin{bmatrix} 2.05 & 0.45 \\ 1.0 & 1.0 \end{bmatrix} \qquad \text{(A2)}$$

average arrival rate = 2.34615, correlation =0.00029

$$D_0 = \begin{bmatrix} -6.5 & 4.0 \\ 1.5 & -4.5 \end{bmatrix} \quad D_1 = \begin{bmatrix} 1.5 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \qquad \text{(B1)}$$

average arrival rate = 2.83333, correlation =-0.00042

$$D_0 = \begin{bmatrix} -5.06 & 2.06 \\ 4.0 & -6.5 \end{bmatrix} \quad D_1 = \begin{bmatrix} 2.56 & 0.44 \\ 1.0 & 1.5 \end{bmatrix} \tag{B2}$$

average arrival rate = 2.83333, correlation =0.00042

$$D_0 = \begin{bmatrix} -6.6 & 4.05 \\ 1.55 & -4.6 \end{bmatrix} \quad D_1 = \begin{bmatrix} 1.55 & 1.0 \\ 1.0 & 2.05 \end{bmatrix} \tag{C1}$$

average arrival rate = 2.88224, correlation =-0.00041

$$D_0 = \begin{bmatrix} -5.15 & 2.1 \\ 4.05 & -6.6 \end{bmatrix} \quad D_1 = \begin{bmatrix} 2.6 & 0.45 \\ 1.0 & 1.55 \end{bmatrix} \tag{C2}$$

average arrival rate = 2.88224, correlation =0.00041

In the above correlation is between two inter-arrival times.

TABLE 1

	$\mathcal{P}_{\text{down}}$	$\mathcal{N}_{\text{pool}}$	$\mathcal{N}_{\text{orbit}}$	$\mathcal{N}_{\text{faic}}$	$\mathcal{P}_{\text{exbusy}}$	$\mathcal{P}_{\text{idle}}$	Cost
A1	$.2805 \times 10^{-2}$	3.262	0.1204	2.2281	0.5620	0.0842	37.8228
A2	$.2803 \times 10^{-2}$	3.2572	0.1207	2.2278	0.5612	0.0850	38.1696
B1	$.2923 \times 10^{-2}$	3.6689	0.1822	2.2431	0.5940	0.0522	68.2556
B2	$.2922 \times 10^{-2}$	3.6647	0.1824	2.2429	0.5935	0.0526	68.4537
C1	$.2932 \times 10^{-2}$	3.7031	0.1888	2.2442	0.5964	0.0497	71.6377
C2	$.2931 \times 10^{-2}$	3.6992	0.1890	2.2440	0.5960	0.0502	71.8214

The table 1 shows that as the external arrival rate increases the system down probability increases; but this increase is narrow as compared to the decrease in server idle probability. Also as expected, the expected number in the pool, in the orbit and the expected number of failed components and the fraction of time the server is found busy with an external customer increases as the external arrival rate increases. The table also shows that as the correlation changes from negative to positive, there is a slight increase in cost and in the server idle probability. Also when correlation changes from negative to positive, the expected number of pooled customers and failed components decrease while the expected number in the orbit increases. The increase in probability $\mathcal{P}_{\text{exbusy}}$ being small compared to the increase in other parameters can be thought of as the reason behind increase in cost. But all these changes are narrow as the difference between negative and positive correlation is small.

Effect of component failure rate : Take $\theta = 20.0$, $\gamma = 0.7$, $\delta = 0.7$, $n = 11$, $k = 4$, $M = 5$, $N = 4$.

Arrival process is according to (A1).

TABLE 2. Effect of component failure rate

λ	$\mathcal{P}_{\text{down}}$	$\mathcal{N}_{\text{pool}}$	$\mathcal{N}_{\text{orbit}}$	$\mathcal{N}_{\text{faic}}$	$\mathcal{P}_{\text{exbusy}}$	$\mathcal{P}_{\text{idle}}$	Cost
0.05	$.196 \times 10^{-8}$	2.1163	0.0285	1.5266	0.7513	0.2310	-67.3177
0.1	$.5933 \times 10^{-7}$	2.1765	0.0311	1.5538	0.7432	0.2213	-63.3658
1.0	$.2801 \times 10^{-2}$	3.2399	0.0907	2.2276	0.5607	0.0855	38.4979
2.0	0.04702	4.2095	0.1748	3.5505	0.3029	0.0208	261.502
3.0	0.17207	4.7390	0.2362	5.1091	0.1149	0.0038	580.397

Table 2 shows that when the component failure rate λ increases, the system down probability as well as expected number of failed components increase and the idle time probability of the server decreases, as expected. But note that as λ increases, the fraction of time the server is found busy with an external customer, decreases and as a result the expected pool size increases. Also note that the expected orbit size is small, which shows that the orbital customers are either transfered to the pool (when λ is small) or leaves the system forever (when λ is large). Since the probability $\mathcal{P}_{\text{down}}$ increases and the probability $\mathcal{P}_{\text{exbusy}}$ decreases, as λ increases, the cost also increases.

Effect of N policy level : $\theta = 20.0, \lambda = 2.0, n = 13, k = 4, M = 5$

The other parameters are same as for table 2.

Table 3 shows that the system performance measure which is most affected by the N-

TABLE 3. Effect of N-policy level

N	$\mathcal{P}_{\text{down}}$	$\mathcal{N}_{\text{pool}}$	$\mathcal{N}_{\text{orbit}}$	$\mathcal{N}_{\text{faic}}$	$\mathcal{P}_{\text{exbusy}}$	$\mathcal{P}_{\text{idle}}$	Cost
4	0.02245	4.2521	0.1802	3.8666	0.2866	0.01969	203.559
5	0.02795	4.2249	0.1801	4.2456	0.2869	0.02325	219.258
6	0.03528	4.1968	0.1796	4.6087	0.2882	0.02717	237.002
7	0.04509	4.1658	0.1787	4.9473	0.2910	0.03135	257.358
8	0.05830	4.1300	0.1771	5.2518	0.2959	0.03577	281.200

policy level is the expected number of failed components; which is expected because as N increases, time for the service of failed components to be started, once the system started with all components operational, increases so that during this time more components may fail. For the same reason a pooled customer has a better chance of getting service and as a result $\mathcal{P}_{\text{exbusy}}$ increases, $\mathcal{N}_{\text{pool}}$ and $\mathcal{N}_{\text{orbit}}$ decreases. Also note that the server idle probability is small. The increase in $\mathcal{N}_{\text{faic}}$ might be the reason behind the increase in cost.

Effect of retrial rate θ : Take $\lambda = 1.0, n = 11, k = 4, M = 5, N = 4$

The other parameters are the same as in table 2.

Table 4 shows that as θ increases, expected number in the orbit decreases but the expected pool size also decreases which tells that retrying customers may be leaving the system. Note that the idle probability of the server is very small and the expected pool size is also close to the maximum pool capacity so that retrying customers may choose to leave the system after a failed retrial. Also this can be thought of as the reason behind the decrease in the fraction of time the server is found busy with an external customer and the increase in cost as θ increases.

Effect of pool size M : $\theta = 10.0, \lambda = 1.0$

The other parameters are same as for table 2.

Table 5 shows that as M, the pool size, increases, expected number of pooled customers increases and as a result the expected number of failed components, the system down probability and the fraction of time the server is found busy with and external customer increases. But the expected number in the orbit decreases, which is expected because as M increases more customers can join the pool. As expected, the idle probability of the server decreases as M increases.

Comparison with the case where no external customers are allowed : Below we compare the k-out-of-n-system described in this chapter with a k-out-of-n system where no external customers are allowed.

Case 1: k-out-of-n system where no external customers are allowed,

Case 2: k-out-of-n system described in this chapter

$\theta = 10.0, \lambda = 1.0, \gamma = 0.7, \delta = 0.7, n = 11, k = 4, N = 4$

TABLE 4. Effect of retrial rate

θ	$\mathcal{P}_{\text{down}}$	$\mathcal{N}_{\text{pool}}$	$\mathcal{N}_{\text{orbit}}$	$\mathcal{N}_{\text{faic}}$	$\mathcal{P}_{\text{exbusy}}$	$\mathcal{P}_{\text{idle}}$	cost
5.0	$.2832 \times 10^{-2}$	3.3908	0.3501	2.2315	0.5704	0.07579	33.688
10.0	$.2813 \times 10^{-2}$	3.3008	0.1790	2.2290	0.5644	0.08176	36.612
15.0	$.2805 \times 10^{-2}$	3.2620	0.1204	2.2281	0.5620	0.08415	37.823
20.0	$.2801 \times 10^{-2}$	3.2399	0.0907	2.2276	0.5607	0.08546	38.498
25.0	$.2798 \times 10^{-2}$	3.2255	0.0728	2.2272	0.5598	0.08630	38.932

TABLE 5. Effect of pool size

M	$\mathcal{P}_{\text{down}}$	$\mathcal{N}_{\text{pool}}$	$\mathcal{N}_{\text{orbit}}$	$\mathcal{N}_{\text{faic}}$	$\mathcal{P}_{\text{exbusy}}$	$\mathcal{P}_{\text{idle}}$	cost
3	$.2655 \times 10^{-2}$	1.9658	0.2155	2.2090	0.5084	0.1377	65.402
4	$.2743 \times 10^{-2}$	2.6238	0.1942	2.2201	0.5410	0.1051	55.047
5	$.2813 \times 10^{-2}$	3.3008	0.1790	2.2290	0.5644	0.0818	36.612

$$D_0 = \begin{bmatrix} -5.5 & 3.5 \\ 1.0 & -3.5 \end{bmatrix} \quad D_1 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.5 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} -7.5 & 2.0 \\ 2.1 & -7.7 \end{bmatrix} \quad S_2 = \begin{bmatrix} -5.06 & 2.06 \\ 4.0 & -6.5 \end{bmatrix}$$

$$S_1^0 = \begin{bmatrix} 5.5 \\ 5.6 \end{bmatrix} \quad S_2^0 = \begin{bmatrix} 3.0 \\ 2.5 \end{bmatrix}$$

$$\alpha = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad \beta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$

TABLE 6. Comparison with the k-out-of-n system where no external customers are allowed

			$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$	$\lambda = 2.5$
$M = 1$	$\mathcal{P}_{\text{down}}$	Case 1	$< 10^{-13}$	$.3956 \times 10^{8}$	$.9124 \times 10^{-6}$	$.2081 \times 10^{-4}$	$.1822 \times 10^{-3}$	$.9335 \times 10^{-3}$
		Case 2	$.129 \times 10^{-7}$	$.2379 \times 10^{-4}$	$.4329 \times 10^{-3}$	$.2039 \times 10^{-2}$	$.5728 \times 10^{-2}$	.01237
	$\mathcal{P}_{\text{busy}}$	Case 1	0.0180	0.0901	0.1802	0.2703	0.3603	0.4501
		Case 2	0.5347	0.5836	0.6415	0.6958	0.7458	0.7914
$M = 2$	$\mathcal{P}_{\text{down}}$	Case 1	$< 10^{-13}$	$.3956 \times 10^{8}$	$.9124 \times 10^{-6}$	$.2081 \times 10^{-4}$	$.1822 \times 10^{-3}$	$.9335 \times 10^{-3}$
		Case 2	$.1801 \times 10^{-7}$	$.3289 \times 10^{-4}$	$.5952 \times 10^{-3}$	$.2782 \times 10^{-3}$	$.7689 \times 10^{-2}$	$.1616 \times 10^{-1}$
	$\mathcal{P}_{\text{busy}}$	Case 1	0.0180	0.0901	0.1802	0.2703	0.3603	0.4501
		Case 2	0.7500	0.7941	0.8434	0.8848	0.9179	0.9433

TABLE 7. Variation in ID_{cost}

	ID_{cost}		$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$	$\lambda = 2.5$
$M = 1$	$C_{11} = 100$	Case 1	−0.1800	−0.9010	−1.8019	−2.7009	−3.5848	−4.4077
	$C_{12} = 10$	Case 2	−5.3470	−5.8336	−6.3717	−6.7541	−6.8852	−6.6770
	$C_{11} = 1000$	Case 1	−0.1800	−0.9010	−1.8011	−2.6822	−3.4208	−3.5675
	$C_{12} = 10$	Case 2	−5.3470	−5.8122	−5.9821	−4.9190	−1.7300	4.4560
	$C_{11} = 10000$	Case 1	−0.1800	−0.9010	−1.7929	−2.4949	−1.7810	4.8340
	$C_{12} = 10$	Case 2	−26.7349	−28.9421	−27.7460	−14.4000	19.9900	84.1300
$M = 4$	$C_{11} = 100$	Case 1	−0.1800	−0.9010	−1.8019	−2.7009	−3.5848	−4.4077
	$C_{12} = 10$	Case 2	−7.5000	−7.9377	−8.3745	−8.5698	−8.4101	−7.8170
	$C_{11} = 1000$	Case 1	−0.1800	−0.9010	−1.8011	−2.6822	−3.4208	−3.5675
	$C_{12} = 10$	Case 2	−7.5000	−7.9081	−7.8388	−6.0660	−1.4900	6.7270
	$C_{11} = 10000$	Case 1	−0.1800	−0.9010	−1.7929	−2.4949	−1.7810	4.8340
	$C_{12} = 10$	Case 2	−7.4998	−7.6121	−2.4820	18.9720	67.7110	152.167

Table 6 shows that compared to the increase in the fraction of time the server is found busy, the increase in the system down probability is not high, if we provide service to external customers in a k-out-of-n system as described in this chapter. To make these statements more clear we consider the cost function

$$ID_{\text{cost}} = C_{11} \cdot \mathcal{P}_{\text{down}} - C_{12} \cdot \mathcal{P}_{\text{busy}}$$

where C_{11} is the loss per unit time the system being down and C_{12} is the profit per unit time due to the server being busy.

Table 7 shows that when $M = 1$ and $\lambda \leq 1.5$, ID_{cost} is smaller in case 2 than case 1, even when C_{11} is 1000 times bigger than C_{12}. But when $\lambda = 2.0$ and 2.5, ID_{cost} is larger in case 2 than case 1, when C_{11} is 100 times larger than C_{12}. When $M = 4$ and $\lambda \leq 1.0$, the table shows that ID_{cost} is smaller in case 2 than in case 1, even when C_{11} is 1000 times bigger than C_{12}. But when $\lambda = 2.0$ and 2.5, ID_{cost} is larger in case 2 than case 1, when C_{11} is 100 times larger than C_{12}.

Table 7 proves atleast numerically that we are able to utilize server idle time without much effecting system reliability.

CHAPTER 4

Reliability of a k-out-of-n system with repair by a service station attending a queue with postponed work*

In this chapter the reliability of a repairable k-out-of-n system is studied. Repair times of components follow a phase type distribution. In addition, the service facility offers service to external customers which on arrive according to a **MAP**. An external customer, who sees an idle server on its arrival, is immediately selected for service. Otherwise, the external customer joins the queue in a pool of postponed work of infinite capacity with probability 1 if the number of failed components in the system is $< M$ $(M \leq n - k + 1)$ and if the number of failed components $\geq M$ it joins the pool with probability γ or leaves the system forever. Repair times of components of the system and that of the external customers have independent phase type distributions. At a service completion epoch if the buffer has less than L customers, a pooled customer is taken for service with probability p, $0 < p < 1$ If at a service completion epoch no component of the system is waiting for repair, a pooled customer, if any waiting, is immediately taken for service.

Thus in this chapter also we study the effect of allowing service to external customers in a k-out-of-n system with single server. But different from chapters 2 and 3, here the external customers are never directed to an orbit instead, they join the queue in a pool or leaves the system forever. Also different from chapter 3, the capacity of the pool is assumed to be infinite and we give freedom for an external customer not to join the pool if he wishes. We expect that such a move will help us to utilize the server idle time more effectively.

---

* This chapter was published in the Proceedings of the Asian International Workshop on Advanced Reliability Modelling (AIWARM) 2004, Hiroshima, Japan, Eds. T. Dohi & W. Y. Yun, World Scientific, pp. 293–300

We obtain the system state distribution under the condition of stability. A number of performance characteristics are derived. A cost function involving L, M, γ and p is constructed and its behaviour investigated numerically.

4.1. Mathematical modelling

We consider a k-out-of-n **cold** system in which the components have exponentially distributed lifetimes with parameter $\frac{\lambda}{i}$, when there are i operational components. There is a single server repair facility which gives service to failed components (main customers) and also to external customers. The external customers arrive according to a MAP with representation (D_0, D_1) of order m. Repair times of main and external customers follow PH-distribution with representations (β_1, S_1) of order m_1 and (β_2, S_2) of order m_2, respectively.

Let $Y_1(t)$ be the number of external customers in the system including the one getting service, if any, and $Y_2(t)$ be the number of main customers in the system including the one getting service, if any, at time t. If an external customer, on arrival, finds a busy server and that $Y_2(t) < M$ $(M \leq n-k+1)$, it joins a pool of infinite capacity with probability 1; on the other hand if $Y_2(t) \geq M$ then with probability γ it joins the pool or leaves the system forever.

If $Y_2(t) = 0$ at a service completion epoch then, with probability 1 a pooled customer, if any, gets service. If $0 < Y_2(t) \leq L-1$, $(L \leq M)$, at a service completion epoch, then with probability p a pooled customer, if there is any, is given service. If $Y_2(t) > L-1$ at a service completion epoch, then with probability 1 a main customer gets service. If $Y_1(t) = Y_2(t) = 0$ then an external customer arriving at time t is taken for service.

Define

$$Y_3(t) = \begin{cases} 0 & \text{if a main customer is getting service at time } t \\ 1 & \text{if an external customer is getting service at time} t \end{cases}$$

Let $Y_4(t)$ and $Y_5(t)$ denote the phases of the arrival and service process respectively.

Now $\mathcal{H} = \{(Y_1(t), Y_2(t), Y_3(t), Y_4(t), Y_5(t)) | t \geq 0\}$ forms a continuous time Markov chain which turns out to be a level independent quasi birth and death process with state space $\cup_{i=0}^{\infty} l(i)$ where $l(i)$ denotes the collection of states in level i and are defined as $l(0) = \{\mathbf{0}\} \cup \{(0, j_1, 0, j_2, j_3) : 1 \leq j_1 \leq n-k+1, 1 \leq j_2 \leq m, 1 \leq j_3 \leq m_1\}$ and for $i \geq 1$,

$$l(i) = \{(i, j_1, 0, j_2, j_3) : 1 \leq j_1 \leq n-k+1, 1 \leq j_2 \leq m, 1 \leq j_3 \leq m_1\}$$
$$\cup\{(i, j_1, 1, j_2, j_3) : 0 \leq j_1 \leq n-k+1, 1 \leq j_2 \leq m, 1 \leq j_3 \leq m_2\}$$

where $\{\mathbf{0}\} = \{(0, j) : 1 \leq j \leq m\}$ represents the collection of states corresponding to $Y_1(t) = Y_2(t) = 0$. Let $J_1 = m + (n-k+1)mm_1$ be the dimension of level $l(0)$ and $J_2 = mm_2 + (n-k+1)m(m_1+m_2)$ be the dimension of levels $l(i)$ for $i \geq 1$. Arranging the states lexicographically we get the infinitesimal generator Q of the process $\mathcal{H}$ as

$$Q = \begin{bmatrix} B_0 & B_1 & 0 & 0 & 0 & \cdots \\ B_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{bmatrix}$$

with

$$B_0 = \begin{bmatrix} B_0^{(1)} & B_0^{(5)} & & & & & & \\ B_0^{(7)} & B_0^{(2)} & B_0^{(6)} & & & & & \\ & B_0^{(8)} & B_0^{(2)} & B_0^{(6)} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & B_0^{(8)} & B_0^{(3)} & B_0^{(6)} & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & B_0^{(8)} & B_0^{(3)} & B_0^{(6)} \\ & & & & & & B_0^{(8)} & B_0^{(4)} \end{bmatrix}$$

where

$$B_0^{(1)} = D_0 - \lambda I_m$$

$$B_0^{(2)} = D_0 \oplus S_1 - \lambda I_{mm_1}$$

$$B_0^{(3)} = (D_0 + (1-\gamma)D_1) \oplus S_1 - \lambda I_{mm_1}$$

$$B_0^{(4)} = (D_0 + (1-\gamma)D_1) \oplus S_1$$

$$B_0^{(5)} = I_m \otimes (\lambda \beta_1)$$

$$B_0^{(6)} = \lambda I_{mm_1}$$

$$B_0^{(7)} = I_m \otimes S_1^0$$

$$B_0^{(8)} = I_m \otimes (S_1^0 \beta_1)$$

$$B_1 = \begin{bmatrix} B_1^{(1)} & 0 & 0 \\ 0 & I_{M-1} \otimes B_1^{(2)} & 0 \\ 0 & 0 & I_{n-k-M+2} \otimes B_1^{(3)} \end{bmatrix}$$

where

$$B_1^{(1)} = D_1 \otimes \beta_2 \qquad B_1^{(2)} = \begin{bmatrix} D_1 \otimes I_{m_1} & 0 \end{bmatrix}_{mm_1 \times m(m_1+m_2)} \qquad B_1^{(3)} = \gamma B_1^{(2)}$$

$$B_2 = \begin{bmatrix} B_2^{(1)} & 0 \\ 0 & I_{n-k+1} \otimes B_2^{(2)} \end{bmatrix}$$

where

$$B_2^{(1)} = I_m \otimes S_2^0 \qquad B_2^{(2)} = \begin{bmatrix} 0 \\ I_m \otimes (S_2^0 \beta_1) \end{bmatrix}_{m(m_1+m_2) \times mm_1}$$

$$A_1 = \begin{bmatrix} A_1^{(1)} & A_1^{(5)} & & & & & & & & & \\ A_1^{(7)} & A_1^{(2)} & A_1^{(6)} & & & & & & & & \\ & A_1^{(8)} & A_1^{(2)} & A_1^{(6)} & & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & & \\ & & & A_1^{(8)} & A_1^{(2)} & A_1^{(6)} & & & & & \\ & & & & A_1^{(9)} & A_1^{(2)} & A_1^{(6)} & & & & \\ & & & & & \ddots & \ddots & \ddots & & & \\ & & & & & & A_1^{(9)} & A_1^{(2)} & A_1^{(6)} & & \\ & & & & & & & A_1^{(9)} & A_1^{(3)} & A_1^{(6)} & \\ & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & A_1^{(9)} & A_1^{(3)} & A_1^{(6)} \\ & & & & & & & & & A_1^{(9)} & A_1^{(4)} \end{bmatrix}.$$

where

$$A_1^{(1)} = (D_0 \oplus S_2) - \lambda I_{mm_2} \qquad A_1^{(2)} = \begin{bmatrix} D_0 \oplus S_1 - \lambda I_{mm_1} & 0 \\ 0 & D_0 \oplus S_2 - \lambda I_{mm_2} \end{bmatrix}$$

$$A_1^{(3)} = \begin{bmatrix} (D_0 + (1-\gamma)D_1) \oplus S_1 - \lambda I_{mm_1} & 0 \\ 0 & (D_0 + (1-\gamma)D_1) \oplus S_2 - \lambda I_{mm_2} \end{bmatrix}$$

$$A_1^{(4)} = A_1^{(3)} + \lambda I_{m(m_1+m_2)} \qquad A_1^{(5)} = \begin{bmatrix} 0 & \lambda I_{mm_2} \end{bmatrix}_{mm_2 \times m(m_1+m_2)}$$

$$A_1^{(6)} = \lambda I_{m(m_1+m_2)} \qquad A_1^{(7)} = \begin{bmatrix} I_m \otimes (S_1^0 \beta_2) \\ 0 \end{bmatrix}_{m(m_1+m_2) \times mm_2}$$

Let $q = 1 - p$.

$$A_1^{(8)} = \begin{bmatrix} q(I_m \otimes (S_1^0 \beta_1)) & p(I_m \otimes (S_1^0 \beta_2)) \\ 0 & 0 \end{bmatrix}, \qquad A_1^{(9)} = \begin{bmatrix} B_0^{(8)} & 0 \\ 0 & 0 \end{bmatrix}$$

$$A_0 = \begin{bmatrix} A_0^{(1)} & 0 & 0 \\ 0 & I_{M-1} \otimes A_0^{(2)} & 0 \\ 0 & 0 & I_{n-k-M+2} \otimes A_0^{(3)} \end{bmatrix}$$

where $A_0^{(1)} = D_1 \otimes I_{m_2}$

$$A_0^{(2)} = \begin{bmatrix} D_1 \otimes I_{m_1} & 0 \\ 0 & D_1 \otimes I_{m_2} \end{bmatrix} \qquad A_0^{(3)} = \gamma A_0^{(2)}$$

$$A_2 = \begin{bmatrix} A_2^{(1)} & 0 & 0 \\ 0 & I_{L-1} \otimes A_2^{(2)} & 0 \\ 0 & 0 & I_{n-k-L+2} \otimes A_2^{(3)} \end{bmatrix}$$

where

$$A_2^{(1)} = I_m \otimes (S_2^0 \beta_2) \qquad A_2^{(2)} = \begin{bmatrix} 0 & 0 \\ q(I_m \otimes (S_2^0 \beta_1)) & p(I_m \otimes (S_2^0 \beta_2)) \end{bmatrix}$$

$$A_2^{(3)} = \begin{bmatrix} 0 & 0 \\ I_m \otimes (S_2^0 \beta_1) & 0 \end{bmatrix}$$

$A_2^{(2)}$ and $A_2^{(3)}$ are square matrices of order $m(m_1 + m_2)$.

4.2. Stability condition

The generator matrix $A = A_0 + A_1 + A_2$ is given by

$$A = \begin{bmatrix} \hat{A}_1 & A_1^{(5)} & & & & & & & & \\ A_1^{(7)} & \hat{A}_2 & A_1^{(6)} & & & & & & & \\ & A_1^{(8)} & \hat{A}_2 & A_1^{(6)} & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & A_1^{(8)} & \hat{A}_2 & A_1^{(6)} & & & & \\ & & & & A_1^{(8)} & \hat{A}_3 & A_1^{(6)} & & & \\ & & & & & A_1^{(9)} & \hat{A}_3 & A_1^{(6)} & & \\ & & & & & & \ddots & \ddots & & \\ & & & & & & A_1^{(9)} & \hat{A}_3 & A_1^{(6)} \\ & & & & & & & A_1^{(9)} & \hat{A}_4 \end{bmatrix}$$

where

$$\hat{A}_1 = (D_0 + D_1 - \lambda I_m) \oplus (S_2 + S_2^0\, \beta_2)$$

$$\hat{A}_2 = \begin{bmatrix} (D_0 + D_1 - \lambda I_m) \oplus S_1 & 0 \\ I_m \otimes (q(S_2^0\, \beta_1)) & (D_0 + D_1 - \lambda I_m) \oplus (S_2 + p(S_2^0\, \beta_2)) \end{bmatrix}$$

$$\hat{A}_3 = \begin{bmatrix} (D_0 + D_1 - \lambda I_m) \oplus S_1 & 0 \\ I_m \otimes (S_2^0 \beta_1) & (D_0 + D_1 - \lambda I_m) \oplus S_2 \end{bmatrix}$$

$$\hat{A}_4 = \begin{bmatrix} (D_0 + D_1) \oplus S_1 & 0 \\ I_m \otimes (S_2^0 \beta_1) & (D_0 + D_1) \oplus S_2 \end{bmatrix}$$

The stationary probability vector π of A, partitioned as

$$\pi = (\pi(0), \pi(1), \pi(2), \ldots, \pi(n-k+1))$$

where the subvector $\pi(0)$ contains mm_2 entries and the subvectors $\pi(i)$ for $1 \leq i \leq n-k+1$ contains $m(m_1+m_2)$ entries, satisfies the equations

$$\pi(0)\hat{A}_1 + \pi(1)A_1^{(7)} = 0 \tag{4.1}$$

$$\pi(0)A_1^{(5)} + \pi(1)\hat{A}_2 + \pi(2)A_1^{(8)} = 0 \tag{4.2}$$

$$\pi(i)A_1^{(6)} + \pi(i+1)\hat{A}_2 + \pi(i+2)A_1^{(8)} = 0, \quad 1 \leq i \leq L-2 \tag{4.3}$$

$$\pi(i)A_1^{(6)} + \pi(i+1)\hat{A}_3 + \pi(i+2)A_1^{(9)} = 0, \quad L-1 \leq i \leq n-k-1 \tag{4.4}$$

$$\pi(n-k)A_1^{(6)} + \pi(n-k+1)\hat{A}_4 = 0 \tag{4.5}$$

together with the normalizing condition

$$\pi e = 1. \tag{4.6}$$

The equations from (4.1) to (4.5) implies

$$\pi(0) = \left[\pi(1)A_1^{(7)}\right]\left[(-\hat{A}_1)^{-1}\right] \tag{4.7}$$

$$\pi(1) = \left[\pi(0)A_1^{(5)} + \pi(2)A_1^{(8)}\right]\left[(-\hat{A}_2)^{-1}\right] \tag{4.8}$$

$$\pi(i+1) = \left[\pi(i)A_1^{(6)} + \pi(i+2)A_1^{(8)}\right]\left[(-\hat{A}_2)^{-1}\right],$$
$$1 \le i \le L-2 \qquad (4.9)$$

$$\pi(i+1) = \left[\pi(i)A_1^{(6)} + \pi(i+2)A_1^{(9)}\right]\left[(-\hat{A}_3)^{-1}\right],$$
$$L-1 \le i \le n-k-1 \qquad (4.10)$$

$$\pi(n-k+1) = \left[\pi(n-k)A_1^{(6)}\right]\left[(-\hat{A}_4)^{-1}\right] \qquad (4.11)$$

The invertibility of the matrices $(D_0 + D_1 - \lambda I_m) \oplus (S_2 + S_2^0\, \beta_2)$, $(D_0 + D_1 - \lambda I_m) \oplus S_1$, $(D_0 + D_1 - \lambda I_m) \oplus (S_2 + p(S_2^0\, \beta_2))$, $(D_0 + D_1 - \lambda I_m) \oplus S_2$ follows from the fact that they are strictly diagonally dominant. The invertibility of the matrix $(D_0 + D_1) \oplus S_1$ can be proved as follows.

Suppose that $(D_0 + D_1) \oplus S_1$ is not invertible, then there exists a non-negative vector $u \neq 0$ such that

$u[(D_0 + D_1) \oplus S_1] = 0$ (ie) $u[(D_0 + D_1) \otimes I_{m_1} + I_m \otimes S_1] = 0$

Multiplying both sides of the above equation with $e_m \otimes I_{m_1}$, we get

$$[u(I_m \otimes S_1)](e_m \otimes I_{m_1}) = 0, \quad \text{since } [(D_0 + D_1) \otimes I_{m_1}](e_m \otimes I_{m_1}) = 0.$$

(ie) $u[e_m \otimes S_1] = 0$

If we partition u as $u = (u_1, u_2, \ldots, u_m)$, where each u_i is a row vector containing m_1 elements, the above equation implies that

$$(u_1 + u_2 + \ldots + u_m)S_1 = 0$$

Now since S_1 is invertible, this implies that

$$u_1 + u_2 + \ldots + u_m = 0$$

since each $u_i \geq 0$, above equation implies $u_i = 0 \; \forall_i$

$\Rightarrow \; u = 0$

which contradicts the assumption that $u \neq 0$.

Hence $(D_0 + D_1) \oplus S_1$ is invertible.

Similarly $(D_0 + D_1) \oplus S_2$ is invertible.

The matrices $\hat{A}_2$, $\hat{A}_3$ and $\hat{A}_4$ have the general form $\begin{bmatrix} H_1 & 0 \\ H_2 & H_3 \end{bmatrix}$ where H_1 and H_3 are invertible. The inverse of such a matrix is given by

$$\begin{bmatrix} H_1^{-1} & 0 \\ -(H_3^{-1} \, H_2 H_1^{-1}) & H_3^{-1} \end{bmatrix}$$

which makes it easier to find the inverses $(\hat{A}_2)^{-1}$, $(\hat{A}_3)^{-1}$ and $(\hat{A}_4)^{-1}$.

The equations from (4.6) to (4.11) are well suited for Block Gauss-Seidel iteration procedure which can now be used to find the vector π.

Now the stability condition can be stated as follows:

The process $\mathcal{H}$ will be positive recurrent if and only if $\pi A_0 e < \pi A_2 e$, where

$$\pi A_0 e = \pi(0) \left[(D_1 e_m) \otimes e_{m_2}\right] + \left[\sum_{i=1}^{M-1} \pi(i) + \sum_{i=M}^{n-k+1} \gamma\pi(i)\right] \begin{bmatrix} (D_1 e_m) \otimes e_{m_1} \\ (D_1 e_m) \otimes e_{m_2} \end{bmatrix}$$

$$\pi A_2 e = \pi(0) \left[e_m \otimes S_2^0\right] + \left[\sum_{i=1}^{n-k+1} \pi(i)\right] \begin{bmatrix} 0 \\ e_m \otimes S_2^0 \end{bmatrix}$$

4.3. Stationary distribution

Since the model is studied as a level independent QBD Markov Process, its stationery distribution (when it exists) has a matrix geometric solution. Under the assumption of the existence of the stationary distribution, let the stationary vector x of Q be partitioned by the levels as $\mathbf{x} = (x(0), x(1), x(2), \ldots)$. Then $x(i)$s are given by

$$x(i) = x(1)R^{i-1} \quad \text{for } i \geq 2$$

where R is the minimal non-negative solution to the matrix quadratic equation

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

The vectors $x(0)$ and $x(1)$ are obtained by solving the equations

$$x(0)B_0 + x(1)B_2 = 0$$

$$x(0)B_1 + x(1)[A_1 + RA_2] = 0$$

subject to the normalizing condition

$$x(0)\mathbf{e} + x(1)(I - R)^{-1}\mathbf{e} = 1.$$

To compute the R matrix numerically we used the logarithmic reduction algorithm (see Latouche and Ramaswami [41]).

Departure process of external customers :

We define the departure process of external customers as the sequence of times $\{\tau_m : m \geq 0\}$ at which the external units leave the system due to a service completion with $\tau_0 \equiv 0$. To study this sequence, it is enough to study the interdeparture times of external customers $\{\bar{\tau}_m = \tau_m - \tau_{m-1} : m \geq 1\}$. Since the random variables $\bar{\tau}_1, \bar{\tau}_2, \ldots$ are identically distributed when the process $\mathcal{H}$ is positive recurrent, we focus on $\bar{\tau}_1$ and determine its distribution under the assumption of positive recurrence of $\mathcal{H}$.

Let $F(t) = P(\bar{\tau}_1 \leq t)$ be the distribution function of $\bar{\tau}_1$ and $\Phi(\theta) = E[e^{-\theta\bar{\tau}_1}]$, $Re(\theta) \geq 0$, be its Laplace-Stieltjes transform.

Conditioning on the state of the process $\mathcal{H}$ at time τ_0, we can write

$$F(t) = \sum_{i=0}^{\infty} x(i)\ F_i(t), \tag{4.12}$$

$$\Phi(\theta) = \sum_{i=0}^{\infty} x(i)\ \Phi_i(\theta), \tag{4.13}$$

where $F_0(t)$ and $\Phi_0(\theta)$ are column vectors with J_1 entries, $F_i(t)$ and $\Phi_i(\theta)$ are column vectors with J_2 entries for $i \geq 1$. The entries of $F_i(t)$ and $\Phi_i(\theta)$ are defined as the conditional distribution functions and conditional Laplace-Stieltjes transforms respectively of $\bar{\tau}_1$, given that the state of the process $\mathcal{H}$ at time τ_0 is in the level $l(i)$ for $i \geq 0$. Since the process $\mathcal{H}$ is level independent, we see that

$F_2(t) = F_3(t) = F_4(t) = \ldots = F_1(t)$ and

$\Phi_2(\theta) = \Phi_3(\theta) = \Phi_4(\theta) = \ldots = \Phi_1(\theta)$.

After arranging the state in the level $l(i)$, $i \geq 1$, lexicographically, we rename them as $(i,1),(i,2),\ldots(i,J_2)$ and the states in the level $l(0)$ as $(0,1),(0,2),\ldots(0,J_1)$. Now to find $F_1(t)$ and $\Phi_1(\theta)$ we suppose that at time τ_0 the process $\mathcal{H}$ is in the state $(1,j)$, $1 \leq j \leq J_2$. Then since the transitions in the level independent QBD process $\mathcal{H}$ due to the arrival process of external customers will not affect the departure process, the time $\bar{\tau}_1$ can be thought of as the time until absorption in a finite continuous time Markov chain $\mathcal{H}_1$ with state space $\{\Delta\} \cup \{1, 2, \ldots, J_2\}$, where Δ is an absorbing state, and with infinitesimal generator

$$Q_1 = \begin{bmatrix} 0 & 0 \\ A_2 e & \bar{A}_1 \end{bmatrix}$$

where

$$\bar{A}_1 = \begin{bmatrix} \bar{A}_1^{(1)} & A_1^{(5)} & & & & & & \\ A_1^{(7)} & \bar{A}_1^{(2)} & A_1^{(6)} & & & & & \\ & A_1^{(8)} & \bar{A}_1^{(2)} & A_1^{(6)} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & A_1^{(8)} & \bar{A}_1^{(2)} & A_1^{(6)} & & & \\ & & & A_1^{(9)} & \bar{A}_1^{(2)} & A_1^{(6)} & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & A_1^{(9)} & \bar{A}_1^{(2)} & A_1^{(6)} & & \\ & & & & A_1^{(9)} & \bar{A}_1^{(2)} & A_1^{(6)} & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & A_1^{(9)} & \bar{A}_1^{(2)} & A_1^{(6)} \\ & & & & & & A_1^{(9)} & \bar{A}_1^{(3)} \end{bmatrix}.$$

with

$$\bar{A}_1^{(1)} = I_m \otimes S_2 - \lambda I_{mm_2}$$

$$\bar{A}_1^{(2)} = \begin{bmatrix} I_m \otimes S_1 - \lambda I_{mm_1} & 0 \\ 0 & I_m \otimes S_2 - \lambda I_{mm_2} \end{bmatrix}$$

$$\bar{A}_1^{(3)} = \bar{A}_1^{(2)} + \lambda I_{m(m_1+m_2)}$$

The process $\mathcal{H}_1$ is obtained from the process $\mathcal{H}$ as follows: since the process $\mathcal{H}$ is at $(1, j)$, $1 \leq j \leq J_2$ at time τ_0, we suppose that the process $\mathcal{H}_1$ starts at state j.Now corresponding to each transition in $\mathcal{H}$ brought by the arrival process of external customers, that is the transitions governed by the matrices D_0 and D_1, we suppose that no transition occurs in $\mathcal{H}_1$. That is corresponding to these transitions in $\mathcal{H}$, there is a sojourn in $\mathcal{H}_1$. Corresponding to those transitions in $\mathcal{H}$ between states in the same level, which are not governed by the matrix D_0, there is a transition in $\mathcal{H}_1$ governed by the matrix $\bar{A}_1$. The moment there is a transition in the process $\mathcal{H}$ which results in a decrease of level of $\mathcal{H}$ by 1 unit, the departure of an internal customer occurs and we suppose that an absorption to the state Δ occurs in $\mathcal{H}_1$ with rates governed by the column matrix $A_2\mathbf{e}$.

Thus given that the process $\mathcal{H}$ is in state $(1, j)$, $1 \leq j \leq J_2$, the time $\bar{\tau}_1$ is the time until absorption in the process $\mathcal{H}_1$ with generator matrix Q_1 and with initial probability vector $\underline{\alpha_1} = (0, \alpha_1)$ where α_1 is a row vector containing J_2 entries whose j^{th} entry is 1 and all other entries are zeros; that is $\bar{\tau}_1$ has a PH distribution with representation $(\alpha_1, \bar{A}_1)$. Hence the j^{th} entry of the column matrix $F_1(t)$, namely $F_{1j}(t)$ is given by

$$F_{1j}(t) = 1 - \alpha_1[\exp(\bar{A}_1 t)]\text{e}.$$

Note that $\alpha_1(\exp(\bar{A}_1 t))\text{e}$ is the j^{th} entry of the column matrix $(\exp(\bar{A}_1 t))\text{e}$. Thus we have

$$F_1(t) = \text{e} - [\exp(\bar{A}_1 t)]\text{e} \tag{4.14}$$

Also the j^{th} entry of $\Phi_1(\theta)$, namely $\Phi_{1j}(\theta)$ is given by

$$\Phi_{1j}(\theta) = \alpha_1(\theta I - \bar{A}_1)^{-1} A_2 \text{e}$$

and therefore

$$\Phi_1(\theta) = (\theta I - \bar{A}_1)^{-1} A_2 \text{e}. \tag{4.15}$$

Now to find $F_0(t)$ and $\Phi_0(\theta)$ we proceed in a similar way. Suppose that at time τ_0 the process $\mathcal{H}$ is in state $(0, j)$, $1 \leq j \leq J_1$. Then the time $\bar{\tau}_1$ can be thought of as the time until absorption in the process $\mathcal{H}_2$ with state space,

$\{\Delta\} \cup \{\overline{(0,1)}, \overline{(0,2)}, \ldots, \overline{(0,J_1)}, 1, 2, \ldots, J_2\}$, where Δ is an absorbing state, and with infinitesimal generator

$$Q_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & B_0 & B_1 \\ A_2 e & 0 & \bar{A}_1 \end{bmatrix}$$

Like the process $\mathcal{H}_1$, the process $\mathcal{H}_2$ is constructed from the process $\mathcal{H}$ as follows: since the process $\mathcal{H}$ is assumed to be in state $(0, j)$ at time τ_0, we suppose the process $\mathcal{H}_2$ starts in the state $\overline{(0, j)}$, $1 \leq j \leq J_1$. Now corresponding to each transition in the process $\mathcal{H}$

from $(0, j_1)$ to $(0, j_2)$, there is a transition in the process $\mathcal{H}_2$ from $\overline{(0, j_1)}$ to $\overline{(0, j_2)}$ at the rate $(B_0)_{j_1 j_2}$. Corresponding to each transition in $\mathcal{H}$ from $(0, j_1)$ to $(1, j_2)$, $1 \leq j_1 \leq J_1$; $1 \leq j_2 \leq J_2$, there is a transition in $\mathcal{H}_2$ from $\overline{(0, j_1)}$ to j_2 at the rate $(B_1)_{j_1 j_2}$. After the process $\mathcal{H}$ reaches the level $l(1)$, corresponding to each transition in $\mathcal{H}$ brought by the arrival process of external customers, we suppose that there is no transition in the process $\mathcal{H}_2$. Corresponding to those transitions in $\mathcal{H}$ within the same level, which are not governed by the matrix D_0, there is a transition in $\mathcal{H}_2$ governed by the matrix $\bar{A}_1$. When a transition which results in a decrease of level by 1 unit occurs in the process $\mathcal{H}$, the departure of an external customer occurs and we suppose that an absorption to the state Δ occurs in the process $\mathcal{H}_2$; with absorption rates governed by the column matrix $A_2 e$. Thus the conditional distribution of $\bar{\tau}_1$ given that at time τ_0 the process $\mathcal{H}$ is in state $(0, j)$, $1 \leq j \leq J_1$, is PH-type with representation $(\alpha_2, \bar{\bar{A}}_1)$ where α_2 is a row vector containing $J_1 + J_2$ entries whose j^{th} entry is 1 and all other entries are zero; and

$$\bar{\bar{A}}_1 = \begin{bmatrix} B_0 & B_1 \\ 0 & \bar{A}_1 \end{bmatrix}.$$

Hence the j^{th} entry of the column matrix $F_0(t)$, namely $F_{0j}(t)$ is given by

$$F_{0j}(t) = 1 - \alpha_2[\exp(\bar{\bar{A}}_1 t)](\mathrm{e}_{J_1+J_2})$$

and therefore

$$F_0(t) = \mathrm{e}_{J_1} - [I_{J_1} \ 0_{J_1 \times J_2}][\exp(\bar{\bar{A}}_1 t)](\mathrm{e}_{J_1+J_2}) \tag{4.16}$$

Also the j^{th} entry of $\Phi_0(\theta)$, namely $\Phi_{0j}(\theta)$ is given by

$$\Phi_{0j}(\theta) = \alpha_2(\theta I - \bar{\bar{A}}_1)^{-1} \begin{bmatrix} 0 \\ A_2 \mathrm{e} \end{bmatrix}$$

and therefore

$$\Phi_0(\theta) = \begin{bmatrix} I_{J_1} & 0_{J_1 \times J_2} \end{bmatrix} (\theta I - \bar{\bar{A}}_1)^{-1} \begin{bmatrix} 0 \\ A_2 \mathrm{e} \end{bmatrix}$$

Now

$$\theta I - \bar{\bar{A}}_1 = \begin{bmatrix} \theta I - B_0 & -B_1 \\ 0 & \theta I - \bar{A}_1 \end{bmatrix}$$

therefore

$$[\theta I - \bar{\bar{A}}_1]^{-1} = \begin{bmatrix} (\theta I - B_0)^{-1} & (\theta I - B_0)^{-1} B_1 (\theta I - \bar{A}_1)^{-1} \\ 0 & (\theta I - \bar{A}_1)^{-1} \end{bmatrix}$$

which gives

$$\Phi_0(\theta) = \begin{bmatrix} I_{J_1} & 0_{J_1 \times J_2} \end{bmatrix} \begin{bmatrix} (\theta I - B_0)^{-1} B_1 (\theta I - \bar{A}_1)^{-1} A_2 \mathrm{e} \\ (\theta I - \bar{A}_1)^{-1} A_2 \mathrm{e} \end{bmatrix}$$

that is,

$$\Phi_0(\theta) = (\theta I - B_0)^{-1} B_1 (\theta I - \bar{A}_1)^{-1} A_2 \mathrm{e} \tag{4.17}$$

Now,

$$\begin{aligned}
F(t) &= \sum_{i=0}^{\infty} x(i) F_i(t) \\
&= x(0) F_0(t) + [\sum_{i=1}^{\infty} x(i)] F_1(t) \\
&= x(0) F_0(t) + x(1)(I - R)^{-1} F_1(t) \\
&= x(0) \left(\mathrm{e}_{J_1} - \left[[I_{J_1}\ 0_{J_1 \times J_2}][\exp(\bar{\bar{A}}_1 t)](\mathrm{e}_{J_1 + J_2}) \right] \right) \\
&\qquad + x(1)(I - R)^{-1}[\mathrm{e} - [\exp(\bar{A}_1 t)]\mathrm{e}] \\
&= \left[x(0)\mathrm{e}_{J_1} + x(1)(I - R)^{-1}\mathrm{e} \right] - [x(0)\ 0][\exp(\bar{\bar{A}}_1 t)](\mathrm{e}_{J_1 + J_2}) \\
&\qquad - x(1)(I - R)^{-1}[\exp(\bar{A}_1 t)]\mathrm{e}
\end{aligned}$$

$$F(t) = 1 - \left\{[x(0)\ 0][\exp(\bar{\bar{A}}_1 t)]\mathrm{e} + x(1)(I-R)^{-1}[\exp(\bar{A}_1 t)]\mathrm{e}\right\} \tag{4.18}$$

The above relation shows that $F(t)$ is the distribution function of a PH distribution with representation $(\alpha_3, \bar{\bar{A}}_1)$ where $\alpha_3 = (x(0)\ 0\ x(1)(I-R)^{-1})$ is a row vector containing $(J_1 + 2J_2)$ elements and $\bar{\bar{A}}_1 = \begin{bmatrix} \bar{\bar{A}}_1 & 0 \\ 0 & \bar{A}_1 \end{bmatrix}$.

Now

$$\begin{aligned}
\Phi(\theta) &= \sum_{i=0}^{\infty} x(i)\Phi_i(\theta) \\
&= x(0)\Phi_0(\theta) + x(1)(I-R)^{-1}\Phi_1(\theta) \\
&= x(0)(\theta I - B_0)^{-1}B_1(\theta I - \bar{A}_1)^{-1}A_2\mathrm{e} \\
&\qquad + x(1)(I-R)^{-1}(\theta I - \bar{A}_1)^{-1}A_2\mathrm{e}
\end{aligned}$$

$$\Phi(\theta) = \left[x(0)(\theta I - B_0)^{-1}B_1 + x(1)(I-R)^{-1}\right](\theta I - \bar{A}_1)^{-1}A_2\mathrm{e} \tag{4.19}$$

Thus we can conclude that the interdeparture time $\bar{\tau}_1$ has a PH-distribution.

4.3.1. System performance measures.

(1) System reliability which is defined as the probability that there is atleast k operational components is given by

$$\theta_1 = x(0)e^{(0)} + x(1)(I-R)^{-1}e^{(1)}$$

where $e^{(0)}$ is a column vector whose last mm_1 entries are 0s and all other entries are 1s and $e^{(1)}$ is a column vector whose last $m(m_1+m_2)$ entries are 0s and all other entries are 1s.

(2) Probability that system is down $\mathcal{P}_{\text{down}} = 1 - \theta_1$.

(3) Expected number of pooled customers

$$\theta_3 = \sum_{i=1}^{\infty} \sum_{j_1=1}^{n-k+1} \sum_{j_2=1}^{m} \sum_{j_3=1}^{m_1} ix(i, j_1, 0, j_2, j_3)$$

$$+ \sum_{i=1}^{\infty} \sum_{j_1=0}^{n-k+1} \sum_{j_2=1}^{m} \sum_{j_3=1}^{m_2} ix(i+1, j_1, 1, j_2, j_3)$$

(4) Expected loss rate of external customers

$$\theta_4 = (1-\gamma)[x(0)e^{(2)} + x(1)(I-R)^{-1}e^{(3)}]$$

where $e^{(2)}$ and $e^{(3)}$ are column vectors given by

$$\mathbf{e}^{(2)} = \begin{bmatrix} 0 \\ e_{n-k-M+2} \otimes ((D_1 e_m) \otimes e_{m_1}) \end{bmatrix}$$

$$\mathbf{e}^{(3)} = \begin{bmatrix} 0 \\ e_{n-k-M+2} \otimes \begin{bmatrix} (D_1 e_m) \otimes e_{m_1} \\ (D_1 e_m) \otimes e_{m_2} \end{bmatrix} \end{bmatrix}$$

(5) Expected number of transfers from the pool when there is atleast 1 main customer present, per unit time

$$\theta_5 = \sum_{i=1}^{\infty} \sum_{j_1=2}^{L} \sum_{j_2=1}^{m} \sum_{j_3=1}^{m_1} x(i, j_1, 0, j_2, j_3) p S_1^0(j_3)$$

$$+ \sum_{i=2}^{\infty} \sum_{j_1=1}^{L-1} \sum_{j_2=1}^{m} \sum_{j_3=1}^{m_2} x(i, j_1, 1, j_2, j_3) p S_2^0(j_3)$$

(6) Expected number of failed components.

$$\theta_6 = x(0)e^{(4)} + x(1)(1-R)^{-1}e^{(5)}$$

where $e^{(4)}$ and $e^{(5)}$ are column matrices given by

$$e^{(4)} = \begin{bmatrix} 0 \\ e^{(6)} \otimes e_{mm_1} \end{bmatrix} \qquad e^{(5)} = \begin{bmatrix} 0 \\ e^{(6)} \otimes e_{m(m_1+m_2)} \end{bmatrix}$$

$$e^{(6)} = [1, 2, \ldots, n-k+1]^T$$

(7) Probability that the server is found busy with an external customer

$$\theta_7 = \sum_{i=1}^{\infty} \sum_{j_1=0}^{n-k+1} \sum_{j_2=1}^{m} \sum_{j_3=1}^{m_2} x(i, j_1, 1, j_2, j_3)$$

(8) Probability that the server is found idle,

$$\theta_8 = \sum_{j=1}^{m} x(0, j)$$

(9) Probability that the server is found busy, $\mathcal{P}_{\text{busy}} = 1 - \theta_8$

(10) Traffic intensity $\rho = \frac{\pi A_0 e}{\pi A_2 e}$

4.4. A cost function and numerical illustrations

Let C_1 be the cost per unit time incurred if the system is down, C_2, be the holding cost per unit time per customer in the pool, C_3 be the cost due to loss of 1 customer and C_4 profit obtained by serving an external unit when there is atleast one main customer present, and C_5 be the holding cost per unit time of one failed component. We construct a cost function as

$$C = \mathcal{P}_{\text{down}} C_1 + \theta_3 C_2 + \theta_4 C_3 - \theta_5 C_4 + \theta_6 \cdot C_5$$

The common parameters for the following tables are:

$n = 35$, $k = 10$, $\gamma = 0.5$, $p = 0.5$

$$\beta_1 = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix} \qquad S_1^0 = \begin{bmatrix} 3.0 \\ 6.0 \end{bmatrix} \qquad S_1 = \begin{bmatrix} -4.0 & 1.0 \\ 1.0 & -7.0 \end{bmatrix}.$$

$$\beta_2 = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \qquad S_2^0 = \begin{bmatrix} 4.0 \\ 9.5 \end{bmatrix} \qquad S_2 = \begin{bmatrix} -5.0 & 1.0 \\ 1.0 & -10.5 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} -5.5 & 3.5 \\ 1.0 & -3.5 \end{bmatrix} \qquad D_1 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.5 \end{bmatrix}$$

Arrival rate = 2.34615, Correlation = −0.00029.

$C_1 = 1000.0, C_2 = 10.0, C_3 = 25.0, C_4 = 75.0, C_5 = 15.0,$

Table 1 shows that when the component failure rate λ is small, increase in 'L' has not much effect on the probability that the server is found idle. But when λ is 2.5, the probability θ_7 decreases as L increases. The reason for this can be obtained from Table 3 which shows that when λ is 2.5, expected number of pooled customer decreases as L increases. An intuitive reasoning for such a behaviour is that as L increases a pooled customer has a better chance of being selected for service. Note that as we have taken $p = 0.5$, when the number of failed components is $< L$, there is equal probability of selecting a pooled customer for service. Also note that the average service rate is greater than average arrival rate. Table 1 also shows that when $\lambda = 0.1$ and 1.5, increase in 'M' has not much effect on θ_7 but when $\lambda = 2.5$, θ_7 increases with increase in M. As in the previous case, the reasoning for this can be obtained from Table 3 which shows that when λ is 2.5, expected number of pooled customers increases as M increases.

Table 2 shows when $\lambda = 0.1$, increase in L and M has not much effect on θ_6. But when $\lambda = 2.5$, θ_7 increases with increase in L as well as in M.

Table 4 shows that only when $\lambda = 2.5$, variations in L and in M has a considerable effect on ρ. When $\lambda = 2.5$, ρ decreases as L increases and ρ increases as M increases. This can be explained in the same way as the variation in θ_7.

Table 5 shows that cost increases as M increases towards $n - k + 1$, decreases as L increases towards M.

In tables 6 and 7 we compare the model in this chapter with the model where no external customers are allowed.

Let case 1 denote k-out-of-n system where no external customers are allowed and case 2 denote the model discussed in this chapter. Table 6 shows that compared to the increase in the server busy probability, the increase in the system breakdown probability is small. To make these statements more clear, as in chapters 2 and 3, we consider a cost function:

$$\text{ID}_{\text{cost}} = C_{11}\mathcal{P}_{\text{down}} + C_{12}\mathcal{P}_{\text{busy}}$$

where C_{11} is the cost per unit time due to the system breakdown and C_{12} is the profit per unit time due to the server becoming busy.

Table 7 shows that by allowing external customers as described in this chapter, there is a decrease in the value of ID_{cost} even when C_{11} is 1000 times larger than C_{12}, except when $\lambda = 2.5$. Which shows atleast numerically that our goal of idle time utilization without affecting the system reliability is achieved through the model in this chapter.

TABLE 1. Variation in probability that the server is found busy with an external customer θ_7

	$\lambda = 0.1$			$\lambda = 1.5$			$\lambda = 2.5$		
L	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$
3	0.3986	0.3986	0.3986	0.3986	0.3986	0.3986	0.3966	0.3985	.3986
5	0.3986	0.3986	0.3986	0.3986	0.3986	0.3986	0.3951	0.3984	.3986
7	0.3986	0.3986	0.3986	0.3986	0.3986	0.3986	0.3916	0.3981	.3986
9	0.3986	0.3986	0.3986	0.3985	0.3986	0.3986	0.3844	0.3974	.3985
10	0.3986	0.3986	0.3986	0.3984	0.3986	0.3986	0.3786	0.3969	.3985
12		0.3986	0.3986		0.3986	0.3986		0.3946	.3983
14		0.3986	0.3986		0.3986	0.3986		0.3898	.3979
15		0.3986	0.3986		0.3986	0.3986		0.3859	.3975
17			0.3986			0.3986			.3960
19			0.3986			0.3986			.3926

G9094

TABLE 2. Variation in expected number of failed components (θ_6).

	$\lambda = 0.1$			$\lambda = 1.5$			$\lambda = 2.5$		
L	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$
3	0.0346	0.0346	0.0346	0.8469	0.8469	0.8469	2.2362	2.2408	2.2412
5	0.0346	0.0346	0.0346	0.9943	0.9944	0.9944	2.9715	2.9875	2.9887
7	0.0346	0.0346	0.0346	1.0692	1.0693	1.0693	3.7530	3.8058	3.8100
9	0.0346	0.0346	0.0346	1.1017	1.1023	1.1023	4.4620	4.6287	4.6421
10	0.0346	0.0346	0.0346	1.1096	1.1106	1.1106	4.7426	5.0314	5.0554
12		0.0346	0.0346		1.1191	1.1191		5.7934	5.8689
14		0.0346	0.0346		1.1224	1.1225		6.4277	6.6551
15		0.0346	0.0346		1.1232	1.1233		6.6457	7.0319
17			0.0346			1.1241			7.7253
19			0.0346			1.1244			8.2554

TABLE 3. Variation in expected number of pooled customers.

	$\lambda = 0.1$			$\lambda = 1.5$			$\lambda = 2.5$		
L	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$
3	0.3236	0.3236	0.3236	2.0168	2.0186	2.0186	38.0455	42.3705	42.7914
5	0.3235	0.3235	0.3235	1.8161	1.8192	1.8193	33.9457	41.0206	41.7485
7	0.3235	0.3235	0.3235	1.7116	1.7179	1.7180	27.1235	39.1407	40.5645
9	0.3235	0.3235	0.3235	1.6606	1.6733	1.6735	17.5855	36.3427	39.2748
10	0.3235	0.3235	0.3235	1.6445	1.6620	1.6623	12.6822	34.3457	38.5674
12		0.3235	0.3235		1.6502	1.6507		28.3938	36.8954
14		0.3235	0.3235		1.6451	1.6461		19.4375	34.5246
15		0.3235	0.3235		1.6436	1.6450		14.4655	32.8427
17			0.3235			1.6440			27.7254
19			0.3235			1.6435			19.6500

TABLE 4. Variation in Traffic intensity(ρ)

	$\lambda = 0.1$			$\lambda = 1.5$			$\lambda = 2.5$		
L	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$	$M = 10$	$M = 15$	$M = 20$
3	0.408	0.408	0.408	0.6080	0.6081	0.6081	0.9310	0.9355	0.9359
5	0.408	0.408	0.408	0.6080	0.6081	0.6081	0.9271	0.9352	0.9358
7	0.408	0.408	0.408	0.6078	0.6081	0.6081	0.9178	0.9345	0.9358
9	0.408	0.408	0.408	0.6075	0.6081	0.6081	0.8965	0.9328	0.9356
10	0.408	0.408	0.408	0.6073	0.6081	0.6081	0.8777	0.9312	0.9355
12		0.408	0.408		0.6080	0.6081		0.9250	0.9350
14		0.408	0.408		0.6080	0.6081		0.9100	0.9338
15		0.408	0.408		0.6080	0.6081		0.8964	0.9327
17			0.408			0.6081			0.9279
19			0.408			0.6081			0.9162

UNIVERSITY LIBRARY
Cochin - 682 022
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

TABLE 5. Variation of the cost function

L	$\lambda = 0.1$, $M = 10$	$\lambda = 0.1$, $M = 15$	$\lambda = 0.1$, $M = 20$	$\lambda = 1.5$, $M = 10$	$\lambda = 1.5$, $M = 15$	$\lambda = 1.5$, $M = 20$	$\lambda = 2.5$, $M = 10$	$\lambda = 2.5$, $M = 15$	$\lambda = 2.5$, $M = 20$
3	1.5034	1.5034	1.5034	-15.275	-15.262	-15.262	331.929	374.640	378.808
5	1.4998	1.4998	1.4998	-25.211	-25.190	-25.189	269.766	339.347	346.537
7	1.4998	1.4998	1.4998	-27.460	-27.420	-27.419	200.107	317.866	331.916
9	1.4998	1.4998	1.4998	-28.097	-28.020	-28.020	111.255	294.086	323.014
10	1.4998	1.4998	1.4998	-28.233	-28.133	-28.131	67.201	277.540	319.197
12		1.4998	1.4998		-28.229	-28.226		226.456	310.386
14		1.4998	1.4998		-28.262	-28.255		146.599	295.685
15		1.4998	1.4998		-28.270	-28.260		101.752	283.603
17			1.4998			-28.265			242.090
19			1.4998			-28.267			170.629

TABLE 6. Comparison with no retrial case $n = 35$, $k = 10$, $\gamma = 0.7$, $p = 0.5$ other parameters are same as for other tables

			$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$	$\lambda = 2.5$
$\mathcal{P}_{\text{down}}$	Case 1		$< 10^{-13}$	$< 10^{-13}$	$< 10^{-13}$	$.36 \times 10^{-11}$	$.3301 \times 10^{-8}$	$.591 \times 10^{-6}$
	Case 2	$L = 20$, $M = 22$	$< 10^{-13}$	$< 10^{-13}$	-3×10^{-12}	$.7493 \times 10^{-8}$	$.7952 \times 10^{-5}$	$.9996 \times 10^{-3}$
		$L = 20$, $M = 25$	$< 10^{-13}$	$< 10^{-13}$	-3×10^{-12}	$.7494 \times 10^{-8}$	$.7961 \times 10^{-5}$	$.1018 \times 10^{-2}$
		$L = 10$, $M = 25$	$< 10^{-13}$	$< 10^{-13}$	$< 10^{-13}$	$.1013 \times 10^{-9}$	$.8911 \times 10^{-7}$	$.1298 \times 10^{-4}$
$\mathcal{P}_{\text{busy}}$	Case 1		0.02296	0.1148	0.2296	0.3444	0.4592	0.5741
	Case 2	$L = 20$, $M = 22$	0.4216	0.5134	0.6282	0.7431	0.8578	0.9701
		$L = 20$, $M = 25$	0.4216	0.5134	0.6282	0.7431	0.8579	0.9718
		$L = 10$, $M = 25$	0.4216	0.5134	0.6282	0.7431	0.8579	0.9727

TABLE 7. $\text{ID}_{\text{cost}} = C_{11}\mathcal{P}_{\text{down}} - C_{12}\mathcal{P}_{\text{busy}}$

ID_{cost}			$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$	$\lambda = 2.5$
$C_{11} = 100$, $C_{12} = 10$	Case 1		−0.2296	−1.1480	−2.2960	−3.4440	−4.5920	−5.7409
	Case 2	$L = 20, M = 22$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5772	−9.6010
		$L = 20, M = 25$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5782	−9.6162
		$L = 10, M = 25$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5790	−9.7257
$C_{11} = 1000$, $C_{12} = 10$	Case 1		−0.2296	−1.1480	−2.2960	−3.4440	−4.5920	−5.7404
	Case 2	$L = 20, M = 22$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5700	−8.7014
		$L = 20, M = 25$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5710	−8.7000
		$L = 10, M = 25$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5789	−9.7140
$C_{11} = 10000$, $C_{12} = 10$	Case 1		−0.2296	−1.1480	−2.2960	−3.4440	−4.5920	−5.7351
	Case 2	$L = 20, M = 22$	−4.2160	−5.1340	−6.2820	−7.4309	−8.4985	0.2950
		$L = 20, M = 25$	−4.2160	−5.1340	−6.2820	−7.4309	−8.4994	0.4620
		$L = 10, M = 25$	−4.2160	−5.1340	−6.2820	−7.4310	−8.5781	−9.5972

4.5. Comparison of Models in chapters 2, 3 and 4

In tables 8 and 9 we compare the three ways of providing service to external customers which are introduced in Chapters 2, 3 and 4 with the case where no external customers are allowed.

Let I denotes the case of a k-out-of-n system where external customers are not allowed, and let II, III and IV denotes the models in chapters 2, 3 and 4 respectively.

The following parameters are common for I, II, III and IV

$n = 11, k = 4$

$$D_0 = \begin{bmatrix} -5.5 & 3.5 \\ 1.0 & -3.5 \end{bmatrix} \quad D_1 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.5 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} -7.5 & 2.0 \\ 2.1 & -7.7 \end{bmatrix} \quad S_2 = \begin{bmatrix} -5.06 & 2.06 \\ 4.0 & -6.5 \end{bmatrix}$$

$$S_1^0 = \begin{bmatrix} 5.5 \\ 5.6 \end{bmatrix} \quad S_2^0 = \begin{bmatrix} 3.0 \\ 2.5 \end{bmatrix}$$

$$\alpha = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad \beta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

The remaining parameters for II are $\theta = 10.0$

The remaining parameters for III are

$\theta = 10.0, \gamma = 0.7, \delta = 0.7, N = 4, M = 4$

The remaining parameters for IV are $\gamma = 0.7, L = 3, M = 5, p = 0.5$

TABLE 8

		$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
$\mathcal{P}_{\text{down}}$	I	$< 10^{-13}$	$.3956 \times 10^{-8}$	$.4014 \times 10^{-6}$
	II	$< 10^{-13}$	$.1321 \times 10^{-7}$	$.1133 \times 10^{-5}$
	III	$.1801 \times 10^{-7}$	$.3289 \times 10^{-4}$	$.3909 \times 10^{-3}$
	IV	$.112 \times 10^{-10}$	$.1437 \times 10^{-5}$	$.6186 \times 10^{-4}$
$\mathcal{P}_{\text{busy}}$	I	0.0180	0.0901	0.1802
	II	0.4408	0.5129	0.5850
	III	0.7500	0.7941	0.8341
	IV	0.8565	0.9285	0.9994

Table 8 shows the effect of providing service to external customers in a k-out-of-n system as described in Chapters 2, 3, 4. To make these effects more clear, we construct a cost function as

$$\text{ID}_{\text{cost}} = C_{11}\mathcal{P}_{\text{down}} - C_{12}\mathcal{P}_{\text{busy}},$$

where C_{11} is the cost per unit time due to the system becoming non operational and C_{12} is the profit per unit time due to the server becoming busy, whose variation according to table 8 is given in table 9.

TABLE 9

IDcost		$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
	I	--0.1800	-0.9010	-1.8020
$C_{11} = 100$	II	--4.4080	-5.1290	-5.8499
$C_{12} = 10$	III	--7.5000	-7.9377	-8.3019
	IV	--8.5650	-9.2849	-9.9878
	I	--0.1800	-0.9010	-1.8016
$C_{11} = 1000$	II	--4.4080	-5.1290	-5.8489
$C_{12} = 10$	III	--7.5000	-7.9081	-7.9501
	IV	--8.5650	-9.2836	-9.9321
	I	--0.1800	-0.9010	-1.7980
$C_{11} = 10000$	II	--4.4080	-5.1289	-5.8387
$C_{12} = 10$	III	--7.4998	-7.6121	-4.4320
	IV	--8.5650	-9.2706	-9.3754

Table 9 shows that the cost decreases continuously when we allow external customers as in chapters 2, 3, 4 except when C_{11} is 1000 times larger than C_{12} and $\lambda = 0.9$, where the cost in chapter 3 model is more than that in chapter 2, but it is less than the cost in the model where no external customers are allowed. It also shows that cost is minimum for the model described in this chapter where the external customers are kept in a pool of postponed work. From table 8 we see that eventhough $\mathcal{P}_{\text{down}}$ is the least if we consider the model in chapter 2, the server busy probability is the highest for the model described in this chapter which makes that model the best from a server idle time utilization point of view.

CHAPTER 5

On a queueing system with self generation of priorities*

Priority queues have been extensively studied by several researchers (see for example Jaiswal [33] and Takagi [55] for detailed analysis, and Gross and Harris [31] for preliminaries). In such queueing systems, arriving customers are classified as belonging to different priorities. The one with highest priority has better access to the service counter than those with lower priorities. Classification into different levels of priority helps in reducing customer impatience. As an example consider a clinic where patients queue up for appointment with physicians. Patients, while waiting in the system, may become seriously ill (priority generation). At this epoch, any physician who is examining an 'ordinary' patient leaves him to be of service to the emergency (priority) case. At the time of arrival of the customer if one of lower priority is going service then the customer in service may be pushed out to accomodate the one just arrived, provided there is no other customer of priority equal to or greater than that tagged to the present arrival. This manner of pushing out a customer of lower priority in service is reffered to as *pre-emptive priority*. The customer who was pushed out will be taken for service only when the system does not have a customer of higher priority waiting. On the other hand, the service of a lower priority customer may continue even after the arrival of a higher priority customer and the latter is taken for service only after the present service is completed. This type of service discipline is reffered to as non-pre-emptive service. The pre-emptive case can be further divided into pre-emptive resume and pre-emptive repeat services. In pre-emptive ressume, the customer of lower priority will continue getting the remaining part of the service; whereas in the latter case service starts from scratch.

* This chapter was published in Neural Parallel & Scientific Computing, Vol. 13, 2005

In this chapter we introduce a new priority queueing system. The system has 'c' servers. Customers on arrival join a queue if all the servers are busy. At the time of arrival there is no classification of priority levels. However, while waiting (and not undergoing service), customers generate into priority at a constant rate. The interoccurence times of priorities are exponentially distributed random variables with parameter depending on the number of customers waiting for service. If, at the time of priority generation, all servers are busy serving priority generated customers, then the present priority generated unit goes out of the system in search of emergency service. However, if at the epoch where a waiting customer generates into a priority type, there is atleast one ordinary customer undergoing service, then such an ordinary customer is pushed back into the waiting line (as the next customer to be served) and the one generated into priority begins to get service. As an example of the model under discussion consider a clinic where patients queue up for appointment with physicians (c in number). Patients, while waiting in the system, may become seriously ill (priority generation). At this epoch, any physician who is examining an 'ordinary' patient leaves him to service the emergency (priority) case. On the other hand if at the time of priority generation all physicians are busy examining emergency cases, then the present priority generated patient will have to leave the system in search of priority service elsewhere.

This chapter is arranged as follows: Section 5.1 deals with the mathematical modelling of the problem. Section 5.2 provides condition for stability of the system. In section 5.3, the steady state distribution is obtained. In section 5.4, we provide some special cases such as the single server case and constant selfgeneration of priorities (ie, independent of system state). In section 5.5 some performance measures such as expected number of customers in the system are given. Also we introduce two different cost functions and some numerical illustrations are provided.

5.1. Mathematical modelling and analysis

The system we study is described as follows. Customers arrive at a c-server counter according to a Poisson process of rate λ. At the time of arrival, each customer is assumed to be ordinary. However, while waiting in the queue, an ordinary customer may generate into a priority customer at a rate α. Since each waiting ordinary customer generates into priority at the constant rate α, the rate of priority generation is $n\alpha$, when n such customers are waiting. If at that epoch there is any ordinary customer getting service, he is then replaced by the customer who currently turned into a priority case. If there is more than one customer in service, then the one who entered the service most recently is replaced. However, if all the customers in service are priority customers, then the present priority generated unit goes out of the system in search of immediate service. This phenomenon happens at clinics and is also observed as a consequence of customer's impatience resulting in joining a higher priority queue from one of lower priority. An ordinary unit in service is pre-empted by the priority generated unit. The service policy is pre-emptive repeat. Service times of ordinary units are *i.i.d.* exponential random variables with parameter $i\mu$ if there are i $(1 \leq i \leq c)$ ordinary customers in service. Also the service times of priority units are *i.i.d.* exponential random variables with parameter $i\mu_1$, $(1 \leq i \leq c)$ priority customers are in the system.
Let $N(t)$ = # of ordinary customers in the system (including those getting service) at time t and $M(t)$ = # of priority customers in the system.

Under the assumptions on the arrival and service processes, $X(t) = \{(N(t), M(t))|t \in R_+\}$ is a continuous time Markov chain on the state space $Z_+ \times \{0, 1, \ldots, c\}$. The states are arranged in the lexicographic order. The level i denoted by $\underline{i}$ is defined as

$$\underline{i} = \{(i, 0), (i, 1), \ldots (i, c)\}$$

The infinitesimal generator of the process is

$$Q = \begin{bmatrix} B_{00} & A_0 & 0 & 0 & \cdots \\ B_{10} & A_{11} & A_0 & 0 & \cdots \\ 0 & B_{21} & A_{22} & A_0 & \cdots \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

where each entry is a $(c+1) \times (c+1)$ matrix. Here, A_0 governs transition rates from level i to $i+1$ and is given by

$$A_0 = \lambda I_{c+1};$$

$B_{i,i-1}$, $i \geq 1$, governs transitions from level i to $i-1$ and is given by

$$B_{i,i-1} = \begin{array}{c} \\ (i,0) \\ (i,1) \\ \\ \\ (i,c) \end{array} \begin{array}{c} \begin{array}{cccc} (i-1,0) & (i-1,1) & & (i-1,c) \end{array} \\ \begin{pmatrix} b_{00i} & b_{01i} & 0 & \\ 0 & b_{11i} & b_{12i} & \\ & \ddots & \ddots & \\ & & b_{c-1\,c-1\,i} & b_{c-1\,c\,i} \\ & & 0 & b_{cci} \end{pmatrix} \end{array}$$

where $b_{jji} = \{\min(c-j,i)\}\mu$, $0 \leq j \leq c-1$, $b_{cci} = i\alpha$, and $b_{j-1,j,i} = \{\max(0, i-(c-j+1))\}\alpha$, $1 \leq j \leq c$, B_{00}, A_{ii}, $i \geq 1$, governs transitions from level $i \to i$ and are of the form

$$\begin{bmatrix} * & 0 & 0 & & \\ 1\mu_1 & * & 0 & & \\ 0 & 2\mu_1 & * & \cdots & \\ \vdots & & \ddots & \ddots & \\ 0 & & & c\mu_1 & * \end{bmatrix}$$

where in each case '*' is such that $(B_{00} + A_0)e = 0$ and $(B_{i,i-1} + A_{ii} + A_0)e = 0$, $i \geq 1$ and e is the column vector of 1s'.

5.2. Ergodicity

The distinctive nature of self generation of priorities of the above process gives the intuition that the process $X(t)$ will be ergodic. Actually this is the case. We use the following result (Tweedie [57]) to prove this.

Proposition (Tweedie) Let $X(t)$ be a Markov process with discrete state space S and rates of transition q_{sp}, $s, p \in S$, $\sum_p q_{sp} = 0$. Assume that there exist

(1) a function $\phi(s)$, $s \in S$, which is bounded from below; (test function)

(2) a positive number ϵ and a map $s \longrightarrow y_s$ such that

- Variable $y_s \leq \sum_{p \neq s} q_{sp}(\phi(p) - \phi(s)) < \infty$ for all $s \in S$;
- $y_s \leq -\epsilon$ for all $s \in S$ except perhaps a finite number of states.

Then the process $X(t)$ is regular and ergodic.

For the model under investigation, we consider the following test function:

$$\phi(s) \equiv \phi(i, j) = i + aj$$

where a is a parameter which will be determined later. The mean drifts $y_s \equiv y_{ij}$ are given by

$$y_{ij} = \begin{cases} \lambda - c\mu + j\mu - aj\mu_1 + (a-1)i\alpha & \\ \quad +(a-1)(j-c)\alpha, & \text{if } 0 \leq j \leq c-1 \text{ and} \\ & \quad i + j > c, \\ \lambda - ac\mu_1 - i\alpha, & \text{if } j = c,\ i + j > c. \end{cases}$$

Since

$$\lim_{i \to \infty} y_{ij} = L_i = \begin{cases} (a-1)\infty, & \text{if } 0 \leq j \leq c-1, \\ -\infty, & \text{if } j = c, \end{cases}$$

the assumptions of Tweedie's theorem hold if and only if, $a - 1 < 0$ (see Falin and Templeton [23]).

Thus whatever be the system parameters, we see that the process $X(t)$ is regular and ergodic. □

We also note that the system remains finite with probability one. Observe that customers may leave the system without service when the system state is (i, c), $i \geq c$. Given that a change takes place when the system is in state (n, c), it is an arrival with probability $\frac{\lambda}{\lambda+c\mu_1+n\alpha}$ and a departure with probability $\frac{c\mu_1+n\alpha}{\lambda+c\mu_1+n\alpha}$. The second expression goes to 1, whereas the first goes to zero with n increasing.

5.3. Steady state distribution

Since the process under consideration is an level dependant quasi birth and death process (LBQBD), to calculate the steady state distribution, we use the methods described in Bright and Taylor **[13]**. Now if we partition the steady state vector x as $x = (x_0, x_1, x_2, \ldots)$ each entry is a row vector containing $c + 1$ entries, then we can write

$$x_k = x_0 \prod_{l=0}^{k-1} R_l \quad \text{for } k \geq 1$$

where the family of matrices $\{R_k, k \geq 0\}$, is the minimal non-negative solutions to the system of equations:

$$A_0 + R_k A_{k+1,k+1} + R_k[R_{k+1} B_{k+2,k+1}] = 0, \; k \geq 0, \tag{5.1}$$

and x_0 is calculated by solving

$$x_0[B_{00} + R_0 B_{10}] = 0, \tag{5.2}$$

such that

$$x_0 e + x_0 \sum_{k=1}^{\infty} [\prod_{l=0}^{k-1} R_l] e < \infty. \tag{5.3}$$

The calculation of the above infinite sums does not seem to be practical, so we approximate x_ks by $x_k(K^*)$s where $\left(x_k(K^*)\right)_j$, $0 \leq k \leq K^*$, $1 \leq j \leq c + 1$, is defined as the stationary probability that $X(t)$ is in state (k, j) of level k, conditional on $X(t)$ being in the set

$$\{(i,j)|0 \le i \le K^*,\ 1 \le j \le c+1\}$$

Then $x_k(K^*)$, $0 \le k \le K^*$ is given by

$$x_k(K^*) = x_0(K^*) \prod_{l=0}^{k-1} R_l \tag{5.4}$$

where $x_0(K^*)$ is found such that it satisfies (5.2) and

$$x_0(K^*)e + x_0(K^*)\Big[\sum_{k=1}^{K^*} [\prod_{l=0}^{k-1} R_l]\Big]e = 1 \tag{5.5}$$

Here we have that for all $i \ge 1$, and for all k, there exists j such that $[B_{i,i-1}]_{k,j} > 0$. Therefore we can construct a dominating process $\bar{X}(t)$ of $X(t)$ and use it to find the truncation level K^* in the same way as in [13] as follows:

The dominating process $\bar{X}(t)$ has generator

$$\bar{Q} = \begin{bmatrix} B_{00} & A_0 & 0 & 0 & 0 & \cdots \\ 0 & A_{11} & \bar{A}_0 & 0 & 0 & \cdots \\ 0 & \bar{B}_{21} & \bar{A}_{22} & \bar{A}_0 & 0 & \cdots \\ 0 & 0 & \bar{B}_{32} & \bar{A}_{33} & \bar{A}_0 & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{bmatrix}$$

where $(\bar{A}_{kk})_{ij} = (A_{kk})_{ij}$ for $k \ge 2$ and $j \ne i$, $(\bar{A}_0)_{ij} = \frac{\lambda}{c+1}$, $1 \le i, j \le c+1$.

$$(\bar{B}_{l,l-1})_{ij} = \frac{1}{c+1}\{((B_{l-1,l-2})e)_{\min}\} \text{ for } l \ge 2$$

5.4. Some particular cases

(i) Constant priority generation: So far we were dealing with the case of linear priority generation. This is especially the case in most of the practical situations (the clinic example given in the beginning of this chapter falls in this category). However priority generation at constant rate, *ie.*, irrespective of the number waiting, also arises in real life situations. Here is an example. A system operates with n components. The failure rate of components, assuming exponential lifetimes with parameter λ/i, when i components are operational, produces constant priority generation. Here priority generation is identified with the failure of components. Operating components are identified with waiting customers.

So we consider the c server case with constant priority generation. The generator matrix in this case has the tridiagonal form

$$Q = \begin{bmatrix} B_{00} & A_0 & 0 & 0 & & & & & \\ B_{10} & A_{11} & A_0 & 0 & & & & & \\ 0 & B_{21} & A_{22} & A_0 & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & B_{c,c-1} & A_{cc} & A_0 & & & \\ & & & & A_2 & A_1 & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

where $A_0 = \lambda I_{c+1}$

$$B_{i,i-1} = \begin{bmatrix} \tilde{b}_{00i} & \tilde{b}_{01i} & 0 & \cdots \\ 0 & \tilde{b}_{11i} & \tilde{b}_{12i} & \cdots \\ & \ddots & \ddots & \\ & & \tilde{b}_{c-1\,c-1\,i} & \tilde{b}_{c-1,c,i} \\ & & 0 & \tilde{b}_{cci} \end{bmatrix}, 1 \le i \le c,$$

$$\tilde{b}_{jji} = \{\min(c-j,i)\}\mu, \quad 0 \le j \le c-1, \quad \tilde{b}_{cci} = \alpha,$$
$$\tilde{b}_{j-1,j,i} = [\frac{\max\{0, i-(c-j+1)\}}{\max\{1, i-(c-j+1)\}}]\alpha, \quad 1 \le j \le c.$$

$B_{00}, A_{ii}, 1 \le i \le c$, are of the form

$$\begin{bmatrix} * & 0 & 0 & & \\ 1\mu_1 & * & 0 & & \\ 0 & 2\mu_1 & * & & \\ & & \ddots & \ddots & \\ & & & c\mu_1 & * \end{bmatrix}$$

where, in each case '*' is such that $(B_{00} + A_0)e = 0$, $(B_{i,i-1} + A_{ii} + A_0)e = 0, 1 \le i \le c$.

$$A_2 = \begin{bmatrix} c\mu & \alpha & 0 & & \\ 0 & (c-1)\mu & \alpha & & \\ & \ddots & & \ddots & \\ & & & 1\mu & \alpha \\ & & & 0 & \alpha \end{bmatrix}$$

$$A_1 = \begin{bmatrix} -\lambda - c\mu - \alpha & 0 & 0 & \\ 1\mu_1 & -\lambda - (c-1)\mu - \mu_1 - \alpha & 0 & \\ 0 & 2\mu_1 & -\lambda - (c-2)\mu - 2\mu_1 - \alpha & \\ & \ddots & & \\ & & c\mu_1 & -\lambda - c\mu_1 - \alpha \end{bmatrix}$$

Let $A = A_0 + A_1 + A_2$ and $\pi = (\pi_0, \pi_1, \ldots, \pi_c)$ be the steady state probability vector of the generator matrix A.

From the homogeneous system $\pi A = 0$ we get

$$\pi_i = \frac{\pi_0}{i!}(\frac{\alpha}{\mu_1})^i, \quad 0 \le i \le c,$$

Then using the normalizing condition $\pi_0 + \pi_1 + \ldots + \pi_c = 1$ we get

$$\pi_0 = \frac{1}{1 + (\frac{\alpha}{\mu_1}) + \frac{1}{2!}(\frac{\alpha}{\mu_1})^2 + \ldots + \frac{1}{c!}(\frac{\alpha}{\mu_1})^c}$$

The system will be stable if and only if, $\pi A_0 e < \pi A_2 e$

ie., if and only if,

$$\lambda < \alpha + \mu[\pi_0 c + \pi_1(c-1) + \ldots + \pi_{c-1}].$$

Under the stability condition, the steady state probability vector x of the generator matrix Q exists. Let us partition x as $\mathbf{x} = (x(0), x(1), \ldots)$.

From the structure of the matrix Q, we can write that

$$x(c+i) = x(c)R^i, \quad \text{for } i \geq 0,$$

where R is the minimal non-negative solution of the matrix quadratic equation

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

The vectors $x(0), x(1), \ldots, x(c)$ satisfy the equations:

$$x(0)B_{00} + x(1)B_{10} = 0 \tag{5.6}$$

$$x(0)A_0 + x(1)A_{11} + x(2)B_{21} = 0 \tag{5.7}$$

$$\ldots$$

$$x(c-2)A_0 + x(c-1)A_{c-1,c-1} + x(c)B_{c,c-1} = 0 \tag{5.8}$$

$$x(c-1)A_0 + x(c)A_{cc} + x(c+1)A_2 = 0. \tag{5.9}$$

Equation (5.9) can be written as

$$x(c-1)A_0 + x(c)[A_{cc} + RA_2] = 0.$$

ie., $x(c-1) = \frac{-1}{\lambda}x(c)[A_{cc} + RA_2]$, since $A_0 = \lambda I_{c+1}$.
Setting $R_{c-1} := \frac{-1}{\lambda}[A_{cc} + RA_2]$, we obtain

$$x(c-1) = x(c)R_{c-1}.$$

From (5.8), $x(c-2)A_0 + x(c)[R_{c-1}A_{c-1,c-1} + B_{c,c-1}] = 0$

$$ie., \quad x(c-2) = \frac{-1}{\lambda}x(c)[R_{c-1}A_{c-1,c-1} + B_{c,c-1}]$$

Putting $R_{c-2} = \frac{-1}{\lambda}[R_{c-1}A_{c-1,c-1} + B_{c,c-1}]$ we get $x(c-2) = x(c)R_{c-2}$. Similarly

$$x(c-3) = x(c)R_{c-3},$$

where $R_{c-3} = \frac{-1}{\lambda}[R_{c-2}A_{c-2,c-2} + R_{c-1}B_{c-1,c-2}]$.
Thus defining the matrices R_i recursively by

$$R_c = I,$$

$$R_{c-1} = \frac{-1}{\lambda}[A_{cc} + RA_2],$$

$R_{c-i} = \frac{-1}{\lambda}[R_{c-i+1}A_{c-i+1,c-i+1} + R_{c-i+2}B_{c-i+2,c-i+1}], 2 \leq i \leq c,$
we can write $x(i) = x(c)R_i$ for $0 \leq i \leq c-1$.

Now from (5.6) we can write

$$x(c)[R_0B_{00} + R_1B_{10}] = 0.$$

This determines $x(c)$ upto a multiplicative constant which can then be evaluated using the normalizing condition

$$x(c)[\sum_{i=0}^{c-1} R_i]e + x(c)(I-R)^{-1}e = 1.$$

(ii) The single server case with constant priority generation:

This case produces sharper results. We have

$$Q = \begin{bmatrix} B_0 & A_0 & 0 & 0 & \\ B_1 & A_{11} & A_0 & 0 & \\ 0 & A_2 & A_1 & A_0 & \\ \vdots & & \ddots & \ddots & \ddots \end{bmatrix}$$

which is a quasi-Toeplitz matrix. Define $A = A_0 + A_1 + A_2$ and $\pi = (\pi_0, \pi_1)$ with $\pi_0, \pi_1 \geq 0$ and $\pi_0 + \pi_1 = 1$, be the steady state probability vector of the generator matrix A. Then the relation $\pi A = 0$ provides us the marginal probabilities for the system with no priority customer and that of the system with a priority customer. These probabilities are immediately computed as $\pi_0 = \frac{\mu_1}{\alpha+\mu_1}$ and $\pi_1 = \frac{\alpha}{\alpha+\mu_1}$. The system is stable if, and only if, $\pi A_2 e > \pi A_0 e$ (see Neuts [44]), that is if and only if,

$$\frac{\mu\mu_1}{\alpha + \mu_1} + \alpha > \lambda.$$

COROLLARY 5.1. *If there is no priority generation, then $\alpha = 0$, and the stability condition reduces to $\mu > \lambda$ which is the stability condition for the classical $M/M/1$ queue.*

COROLLARY 5.2. *The model considered here generalises the classical queue with reneging as explained below:*

Consider the $M/M/1$ queue with reneging. The reneging rate is linear with parameter α (per unit). Suppose $p(i)$ is the probability of the system being found in state i in the long run.

Now consider the system described here with linear priority generation. Then

$$p(i) = \frac{(i-1)\alpha}{(i-1)\alpha + \mu + \lambda}\pi(i,0) + \pi(i-1,1)$$

This can be extended to the multiserver case as well.

To obtain the system state probabilities, we proceed to calculate G (a stochastic matrix) from

$$A_0G^2 + A_1G + A_2 = 0 \tag{5.10}$$

where

$$G = \begin{bmatrix} G_{00} & G_{01} \\ G_{10} & G_{11} \end{bmatrix}$$

(see Latouche and Ramaswami [41])

Since G is stochastic, we have $G_{01} = 1 - G_{00}$ and $G_{11} = 1 - G_{10}$ and in the present case we have G as a 2×2 matrix, its elements are computed easily using (5.10):

On substituting A_0, A_1, A_2, G and G^2 in (5.10) we note that G_{00} satisfies

$$(G_{00})^3\lambda\mu - (G_{00})^2(-\alpha(\mu_1 + \lambda - \mu) + \mu(-\mu_1 + 2\lambda + \mu))$$
$$- G_{00}(\alpha^2 + \alpha(\mu_1 + \lambda) + (2\mu_1 - \lambda - 2\mu)\mu) - \mu(-\alpha - \mu_1 + \mu) = 0$$

and G_{11} satisfies

$$G_{11} = \frac{1}{\lambda}[(\lambda + \alpha + \mu - \lambda G_{00}) - \frac{\alpha}{1 - G_{00}}]$$

Thus, for parameters λ, α, μ, μ_1 satisfying the stability condition, we can calculate G.

We can find G in another way also. Note that here G has distinct characteristic values. For, if G has only one characteristic value namely 1, then, since G is stochastic, we can see that G is the identity matrix and it cannot satisfy equation (5.10). Thus G is diagonalizable. Now we can find the characteristic value of G, other than 1, by solving the equation:

$$\det(x^2A_0 + xA_1 + A_2) = 0 \text{ for a root less than 1.}$$

The characteristic vectors of G corresponding to a characteristic value λ are the characteristic vectors of the matrix $G(\lambda) = \lambda^2A_0 + \lambda A_1 + A_2$ corresponding to the characteristic value 0. Finding these characteristic vectors corresponding to the two characteristic values, we can get G.

Now we can compute the rate matrix R from the relation $R = A_0(-A_1 - A_0G)^{-1}$. Now the system state probabilities are given by $\mathbf{x} = (x(0), x(1), \ldots)$ where

$$x(i) = (y(i,0), y(i,1)).$$

Now $x(i)$s satisfy

$$x(i) = x(1)R^{i-1},\ i \geq 1.$$

Now to calculate $x(0)$ and $x(1)$ we use the equations

$$x(0)B_0 + x(1)B_1 = 0,$$

$$x(0)A_0 + x(1)[A_{11} + RA_2] = 0,$$

together with the normalizing condition $\sum_{i=0}^{\infty} x(i)e = 1$.

5.5. System performance measures

Here we obtain some of the important measures of performance of the system in the long run in the single server case with rate of priority generation depending on the level of the process. These provide us information about the various characteristics of the system.

The performance measures that we concentrate on are

(i) Average number of customers in the system is given by
$\sum_{i=0}^{\infty}\{i\pi(i,0) + (i+1)\pi(i,1)\}$.

(ii) Average number of customers lost per unit time is $\sum_{i=1}^{\infty} i\alpha\pi(i,1)$

(iii) Hence the number of priority generated units in unit time is
$\{\sum_{i=1}^{\infty} i\alpha\pi(i,1) + \sum_{i=2}^{\infty}(i-1)\alpha\pi(i,0)\}$

These measures can be utilized to construct the following cost functions in the case of single server system.

TABLE 1. Case 1: $\lambda = 1.0$, $\alpha = 0.8$, $\mu = 1.1$, $C_1 = 50.0$, $C_2 = 10.0$, $C_3 = 10.0$, $C = 10.0$, $h = 4.0$

μ_1	0.6	0.8	1.0	1.2	1.4
$F_1(\alpha, \mu_1)$	-0.472	-3.419	-5.110	-6.204	-6.969
$F_2(\alpha, \mu_1)$	11.928	9.373	7.976	7.113	6.536

TABLE 2. Case 1: $\lambda = 1.0$, $\mu = 1.1$, $\mu_1 = 1.1$, $C_1 = 50.0$, $C_2 = 10.0$, $C_3 = 10.0$, $C = 10.0$, $h = 4.0$

α	0.8	0.9	1.0	1.1	1.2
$F_1(\alpha, \mu_1)$	-5.710	-5.645	-5.585	-5.528	-5.476
$F_2(\alpha, \mu_1)$	7.498	7.347	7.226	7.127	7.044

Case 1. Nonconstant priority generation

Here

$$F_1(\alpha, \mu_1) = -\left(\sum_{i=2}^{\infty} \pi(i,0)\frac{(i-1)\alpha}{\mu}\right) C_1 + \left(\sum_{i=1}^{\infty} \pi(i,1)\frac{i\alpha}{\mu_1}\right) C_2 + \left(\sum_{i=2}^{\infty} \pi(i,0)(i-1)\alpha\right) C_3$$

where the first term on the right sides represents the revenue to the system by way of serving priority units over unit time. The second term represents the loss to the system due to priority generated customers leaving the queue when a priority customer is getting service. The last term represents the cost to the system due to pre-emption of service of ordinary customers.

We consider another cost function

$$F_2(\alpha, \mu_1) = \left(\sum_{i=1}^{\infty} \pi(i,1)\frac{i\alpha}{\mu_1}\right) C + h\left[\sum_{i=1}^{\infty} i\pi(i,0) + \sum_{i=0}^{\infty}(i+1)\pi(i,1)\right]$$

The first term on the right side represents loss to the system due to priority generated customers leaving the system when a priority customer is receiving service and the second one represents holding cost of customers in the system.

TABLE 3. Case 2: (Variations of μ_1) where $\lambda = 1$, $\alpha = 1$, $\mu = 1.1$, $c_1 = 20, c_2 = 10, c_3 = 15, C = 10$.

μ_1	$F_1(\alpha,\mu_1)$	$F_2(\alpha,\mu_1)$			Average number of customers in the system	Average number of customers lost per unit time	Average number of priority units generated.
		$h=1$	$h=5$	$h=12$			
1.1	1.054	3.829	11.62	25.254	1.948	0.207	0.467
1.2	1.166	3.487	11.012	24.182	1.881	0.193	0.456
1.3	1.284	3.212	10.515	23.295	1.826	0.180	0.447
1.4	1.402	2.988	10.102	22.551	1.778	0.169	0.439
1.5	1.518	2.803	9.754	21.918	1.738	0.1598	0.431

TABLE 4. Case 2: (Variations of α) where $\lambda = 1$, $\mu = 1.1$, $\mu_1 = 1.1$, $c_1 = 20, c_2 = 10, c_3 = 15, C = 10$.

α	$F_1(\alpha,\mu_1)$	$F_2(\alpha,\mu_1)$			Average number in the system	Average number of customers lost	Average number of priority units generated.
		$h=1$	$h=5$	$h=12$			
1	1.054	3.829	11.62	25.254	1.948	0.207	0.467
1.1	1.195	3.808	10.937	23.414	1.782	0.223	0.484
1.15	1.261	3.80366	10.652	22.637	1.712	0.230	0.491
1.2	1.323	3.80309	10.398	21.938	1.649	0.237	0.498
1.25	1.384	3.8053	10.169	21.307	1.591	0.244	0.505
1.3	1.441	3.80959	9.964	20.734	1.539	0.2498	0.511
1.4	1.548	3.8231	9.6096	19.736	1.447	0.261	0.522
1.9	1.962	3.92931	8.5319	16.587	1.151	0.306	0.562

Case 2. Constant priority generation:

$$F_1(\alpha,\mu_1) = -\left(\sum_{i=2}^{\infty}\pi(i,0)\frac{\alpha}{\mu}\right)C_1 + \left(\sum_{i=1}^{\infty}\pi(i,1)\frac{\alpha}{\mu_1}\right)C_2 + \left(\sum_{i=2}^{\infty}\pi(i,0)\alpha\right)C_3$$

$$F_2(\alpha,\mu_1) = \left(\sum_{i=1}^{\infty}\pi(i,1)\frac{\alpha}{\mu_1}\right)C + h\left[\sum_{i=1}^{\infty}i\pi(i,0) + \sum_{i=0}^{\infty}(i+1)\pi(i,1)\right]$$

Table 1 shows that as μ_1 increases, both F_1 and F_2 decrease. This may be due to the fact that as μ_1 increases, the mean service time of a priority customer decreases and as a result the loss to the system due to priority generated customers leaving the system decreases. Table 2 shows that as α increases, F_1 increases which can be attributed to the fact that as α increases, pre-emption rate of ordinary customer in service as well as loss rate due to priority generation increases. But as α increases, the overall holding cost decreases which can be regarded as the reason behind the decrease in F_2. Table 3 shows that in the case of constant priority generation as μ_1 increases, F_1 also increases. This may be due to the fact that as μ_1 increases, eventhough the loss rate due to priority generation decreases (see table 3) priority generation rate also decreases (see table 3) so that there is a decrease in the revenue due to serving priority customers. Also note that in this case the priority generation rate is assumed to be independent of the number of ordinary customers in the system.

CHAPTER 6

The impact of self-generation of priorities on multi-server queues with finite capacity*

This chapter deals with multi-server queues with a finite buffer of size N in which units waiting for service generate into priority at a constant rate, independently of other units in the buffer. At the epoch of a unit's priority generation, the unit is immediately taken for service if there is one unit in service which did not generate into priority while waiting; otherwise such a unit leaves the system in search of immediate service elsewhere. The arrival stream of units is a Markovian arrival process (MAP) and service requirements are of phase (PH) type. Our interest is in the continuous-time Markov chain describing the state of the queue at arbitrary times, which constitutes a finite quasi-birth-and-death (QBD) process. We give formulas well suited for numerical computation for a variety of performance measures, including the blocking probability, the departure process, and the stationary distributions of the system state at pre-arrival epochs, at post-departure epochs and at epochs at which arriving units are lost. Illustrative numerical examples show the effect of several parameters on certain probabilistic descriptors of the queue for various levels of congestion.

This chapter is organized as follows. In section 6.1 we start by introducing self-generation of priorities in the MAP/PH,PH/c/$c + N$ queue. In section 6.2 the focus is on the continuous-time Markov chain (CTMC) at arbitrary times which constitutes a finite quasi-birth-and-death (QBD) process. An efficient computational approach to its analysis is then derived. In section 6.3, we give tractable analytical formulas for the departure process, the blocking probability and the stationary distributions at pre-arrival epochs, at

* This chapter was published in Stochastic Models, Vol. 21, No. 2–3, pp. 427–447, 2005

post-departure epochs and at epochs at which arriving units are lost. In section 6.4 the effect of several parameters on probabilistic descriptors of our queue is numerically illustrated.

6.1. The finite capacity MAP/PH,PH/c/c + N queue with self-generation of priorities

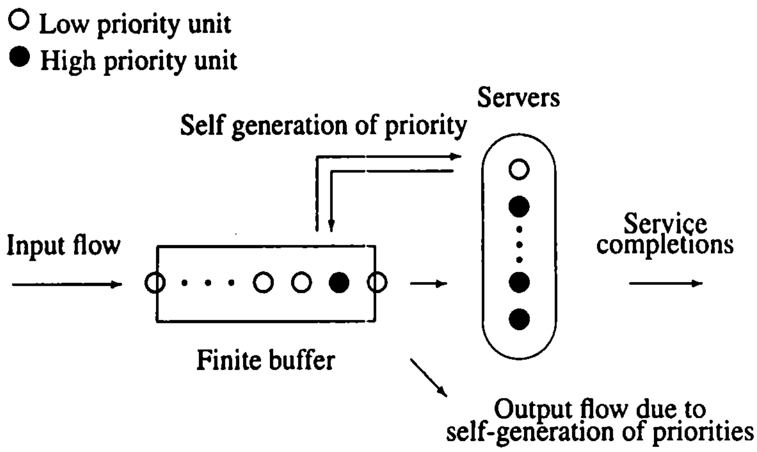

Figure 1. Multi-server queue with finite capacity and self-generation of priorities

Figure 1 depicts the configuration of the priority-generating queue to be investigated in this chapter. We consider a multi-server queue consisting of c servers and a finite buffer of size $N \geq 1$, in which units arrive one at a time according to a Markovian arrival process. Formally, the MAP is parameterized by two $a \times a$ matrices $\mathbf{D}_0$ and $\mathbf{D}_1$, whose sum $\mathbf{D} \equiv \mathbf{D}_0 + \mathbf{D}_1$ is an irreducible infinitesimal generator. The (i, j)th entry of the matrix $\mathbf{D}_1$ corresponds to the transition rate associated with the arrival of one unit when the underlying CTMC makes a transition from the state i to the state j. The matrix $\mathbf{D}_0$ covers the case when there is no arrival. Then the arrival rate of the point process of units is given by $\lambda = \mathbf{d}\mathbf{D}_1\mathbf{e}_a$ where $\mathbf{d}$ is the stationary vector of $\mathbf{D}$ and $\mathbf{e}_a$ is the a-dimensional column vector of 1s.

Arriving units are of homogeneous nature, whence they are identified as low priority units. They are queued in the buffer and treated in order of their arrival. During waiting in the buffer a low priority unit turns into one of high priority at a constant rate $\gamma > 0$, independently of other units in the queue. At the epoch at which a waiting unit generates into high priority, it is immediately taken for service, provided that there is at least one low priority unit in service at that time. Assume that there is an identical chance for assigning the high priority turned unit to any of the servers which are occupied by low priority units. On the contrary if all servers are busy serving high priority units when a waiting unit generates into one of high priority, then the latter leaves the system in search of urgent service elsewhere.

Assume that low priority units are preempted by high priority turned units. Specifically, a low priority unit in service, when preempted by a high priority turned unit, is queued in the buffer according to a predetermined rule and its elapsed service time is lost. Low priority units in service do not turn into the high priority category.

Successive service times of low and high priority units are mutually independent, and follow PH laws with representations (α, T) and (β, S), respectively. Here, T and S are square matrices with negative diagonal elements and nonnegative off-diagonal elements. Assume, without loss of generality, that T and S are stable matrices; see e.g. the books [**41**, **44**] for a review of the main results on PH random variables. For later use, t_0 and s_0 are column vectors of sizes t and s, respectively, defined by $\mathrm{t}_0 = -\mathrm{T}\mathrm{e}_t$ and $\mathrm{s}_0 = -\mathrm{S}\mathrm{e}_s$.

The stream of units, the process of self-generation of priorities and the service times are assumed to be mutually independent.

In what is to follow, $\otimes$ and $\oplus$ stand for Kronecker product and sum respectively (see [40]), I_p is the identity matrix of order p, $0_{p\times q}$ is the zero matrix of dimension $p \times q$ and 0_p is the p-dimensional column vector of 0s. If v is a v-dimensional column vector and w is a w-dimensional row vector, then the product vw is the matrix of dimension $v \times w$ with elements $[\mathrm{vw}]_{ij} = \mathrm{v}_i \mathrm{w}_j$. Given a square matrix V, we define $\mathrm{V}^{\oplus m}$ as the matrix

$$\mathrm{V}^{\oplus m} = \overbrace{\mathrm{V} \oplus \mathrm{V} \oplus \cdots \oplus \mathrm{V}}^{m},\ m \geq 1,$$

and $\mathbf{V}^{\oplus 0}$ as the scalar 0. For a column vector $\mathbf{v}$ with v entries, we denote by $\mathbf{v}^{\oplus m}$ the matrix defined as $\mathbf{v}\otimes\mathbf{I}_{v^{m-1}}+\mathbf{I}_v\otimes\mathbf{v}\otimes\mathbf{I}_{v^{m-2}}+\cdots+\mathbf{I}_{v^{m-1}}\otimes\mathbf{v}$, for $m\geq 1$, and the scalar 1 for $m=0$.

6.2. Basic results of the system

The purpose of this section is to find the stationary vector of the system state at arbitrary times. Let $\xi_l(u)$ and $\xi_h(u)$ be the number of low and high priority units in the system at time u, respectively, and $\eta(u)$ be the phase of the arrival process. Define two vectors $\boldsymbol{\nu}_l(u)$ and $\boldsymbol{\nu}_h(u)$ that record phases of service corresponding, respectively, to low and high priority units in service at time u. Based on the model description of our queue, we see that $\mathcal{X}=\{(\xi_l(u),\xi_h(u),\eta(u),\boldsymbol{\nu}_l(u),\boldsymbol{\nu}_h(u)):u\geq 0\}$ forms a finite QBD process on the state space

$$\mathcal{S}=\bigcup_{n=0}^{c+N} l(n),$$

where the nth level is given by

$$\begin{aligned} l(n) &= \cup_{m=0}^{n}\mathcal{L}(m,n-m), \ \text{for } 0\leq n\leq c,\\ &= \cup_{m=0}^{c}\mathcal{L}'(n-c+m,c-m), \ \text{for } c+1\leq n\leq c+N. \end{aligned}$$

The subsets $\mathcal{L}(m,n-m)$ are defined as $\{(m,n-m,i,j_1,\ldots,j_m,k_1,\ldots,k_{n-m}):1\leq i\leq a,1\leq j_1,\ldots,j_m\leq t,1\leq k_1,\ldots,k_{n-m}\leq s\}$, for $0\leq m\leq n\leq c$, and $\mathcal{L}'(n-c+m,c-m)$ are given by $\{(n-c+m,c-m,i,j_1,\ldots,j_m,k_1,\ldots,k_{c-m}):1\leq i\leq a,1\leq j_1,\ldots,j_m\leq t,1\leq k_1,\ldots,k_{c-m}\leq s\}$, for $0\leq m\leq c<n\leq c+N$. Thus the level $l(n)$ consists of J_n states, where

$$J_n = a\sum_{i=0}^{\min(n,c)} t^{\min(n,c)-i}s^i, \ 0\leq n\leq c+N.$$

If the states in $\mathcal{S}$ are listed in lexicographical order, then transitions among subsets $\mathcal{L}(m,n-m)$ and $\mathcal{L}'(n-c+m,c-m)$ are summarized as follows:

(a) For $0 \le n \le c$, the matrices $\mathbf{I}_a \otimes \mathbf{t}_0^{\oplus m} \otimes \mathbf{I}_{s^{n-m}}$, $\mathbf{I}_{at^m} \otimes \mathbf{s}_0^{\oplus n-m}$ and $\mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus n-m}$ record transition rates to states in $\mathcal{L}(m-1, n-m)$ for $1 \le m \le n$, in $\mathcal{L}(m, n-m-1)$ for $0 \le m \le n-1$, and in $\mathcal{L}(m, n-m)$ for $0 \le m \le n$, respectively, starting from states in $\mathcal{L}(m, n-m)$.

(b) For $0 \le n \le c-1$, the matrix $\mathbf{D}_1 \otimes \mathbf{I}_{t^m} \otimes \boldsymbol{\alpha} \otimes \mathbf{I}_{s^{n-m}}$ records transition rates to states in $\mathcal{L}(m+1, n-m)$ for $0 \le m \le n$, starting from states in $\mathcal{L}(m, n-m)$.

(c) The matrix $\mathbf{D}_1 \otimes \mathbf{I}_{t^m s^{c-m}}$ records transition rates to states in $\mathcal{L}'(m+1, c-m)$ for $0 \le m \le c$, starting from states in $\mathcal{L}(m, c-m)$.

(d) For $c+1 \le n \le c+N-1$, the matrices $\mathbf{D}_1 \otimes \mathbf{I}_{t^m s^{c-m}}$ and $\mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m} - (n-c)\gamma \mathbf{I}_{at^m s^{c-m}}$ record transition rates to states in $\mathcal{L}'(n-c+m+1, c-m)$ for $0 \le m \le c$, and in $\mathcal{L}'(n-c+m, c-m)$ for $0 \le m \le c$, respectively, starting from states in $\mathcal{L}'(n-c+m, c-m)$.

(e) For $c+1 \le n \le c+N$, the matrices $\mathbf{I}_a \otimes (\mathbf{t}_0\boldsymbol{\alpha})^{\oplus m} \otimes \mathbf{I}_{s^{c-m}}$, $\mathbf{I}_{at^m} \otimes \boldsymbol{\alpha} \otimes \mathbf{s}_0^{\oplus c-m}$ and $\mathbf{I}_a \otimes \left(\frac{(n-c)\gamma}{m}\mathbf{e}_t^{\oplus m}\right) \otimes \mathbf{I}_{s^{c-m}} \otimes \boldsymbol{\beta}$ record transition rates to states in $\mathcal{L}'(n-c+m-1, c-m)$ for $1 \le m \le c$, in $\mathcal{L}'(n-c+m, c-m-1)$ for $0 \le m \le c-1$, and in $\mathcal{L}'(n-c+m-1, c-m+1)$ for $1 \le m \le c$, respectively, starting from states in $\mathcal{L}'(n-c+m, c-m)$. Similarly, the matrix $(n-c)\gamma \mathbf{I}_{as^c}$ records transition rates to states of $\mathcal{L}'(n-c-1, c)$, starting from states in $\mathcal{L}'(n-c, c)$.

(f) The matrix $\mathbf{D} \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m} - N\gamma \mathbf{I}_{at^m s^{c-m}}$ records transition rates to states of $\mathcal{L}'(N+m, c-m)$ for $0 \le m \le c$, starting from states in $\mathcal{L}'(N+m, c-m)$.

Then the CTMC $\mathcal{X}$ possesses the following infinitesimal generator:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{A}_0 & & & & & & & \\ \mathbf{C}_1 & \mathbf{B}_1 & \mathbf{A}_1 & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & \mathbf{C}_{c-1} & \mathbf{B}_{c-1} & \mathbf{A}_{c-1} & & & & \\ & & & \mathbf{C}_c & \mathbf{B}_c & \mathbf{A}_c & & & \\ & & & & \mathbf{C}_{c+1} & \mathbf{B}_{c+1} & \mathbf{A}_c & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & \mathbf{C}_{c+N-1} & \mathbf{B}_{c+N-1} & \mathbf{A}_c \\ & & & & & & & \mathbf{C}_{c+N} & \mathbf{B}_{c+N} + \mathbf{A}_c \end{bmatrix} \quad (6.1)$$

where

$$\mathbf{A}_n = [\mathbf{0}_{J_n \times as^{n+1}}, \mathbf{U}_n], \text{ for } 0 \le n \le c-1,$$

$$= \mathbf{U}_c, \text{ for } n = c,$$

$$\mathbf{B}_n = \mathbf{D}_0, \text{ for } n = 0,$$

$$= diag[\mathbf{D}_0 \oplus \mathbf{S}^{\oplus n}, \mathbf{D}_0 \oplus \mathbf{T} \oplus \mathbf{S}^{\oplus n-1}, ..., \mathbf{D}_0 \oplus \mathbf{T}^{\oplus n}], \text{ for } 1 \le n \le c,$$

$$= \mathbf{B}_c + \begin{bmatrix} \mathbf{0}_{as^c \times (J_c - at^c)} & \mathbf{0}_{as^c \times at^c} \\ \mathbf{U}_n & \mathbf{0}_{(J_c - as^c) \times at^c} \end{bmatrix} - (n-c)\gamma \mathbf{I}_{J_c},$$

for $c+1 \le n \le c+N$,

$$\mathbf{C}_n = \begin{bmatrix} \mathbf{V}_n \\ \mathbf{0}_{at^n \times J_{n-1}} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{as^n \times J_{n-1}} \\ \mathbf{W}_n \end{bmatrix}, \text{ for } 1 \le n \le c,$$

$$= \begin{bmatrix} \mathbf{0}_{(J_c - at^c) \times as^c} & \mathbf{V}_n \\ \mathbf{0}_{at^c \times as^c} & \mathbf{0}_{at^c \times (J_c - as^c)} \end{bmatrix} + \mathbf{W}_n, \text{ for } c+1 \le n \le c+N.$$

The expressions for $\mathbf{U}_n$, $\mathbf{V}_n$ and $\mathbf{W}_n$ are as follows.

Expressions for the blocks $\mathbf{U}_n$. The matrices $\mathbf{U}_n$ are easily written as follows:

(a) For $0 \le n \le c-1$, $\mathbf{U}_n = diag[\mathbf{D}_1 \otimes \boldsymbol{\alpha} \otimes \mathbf{I}_{s^n}, \mathbf{D}_1 \otimes \mathbf{I}_t \otimes \boldsymbol{\alpha} \otimes \mathbf{I}_{s^{n-1}}, ..., \mathbf{D}_1 \otimes \mathbf{I}_{t^n} \otimes \boldsymbol{\alpha}]$.

(b) For $n = c$, $\mathbf{U}_n = diag[\mathbf{D}_1 \otimes \mathbf{I}_{s^c}, \mathbf{D}_1 \otimes \mathbf{I}_{ts^{c-1}}, ..., \mathbf{D}_1 \otimes \mathbf{I}_{t^c}]$.

(c) For $c+1 \le n \le c+N$, $\mathbf{U}_n = diag\,[\mathbf{I}_a \otimes ((n-c)\gamma \mathbf{e}_t) \otimes \mathbf{I}_{s^{c-1}} \otimes \boldsymbol{\beta}$, $\mathbf{I}_a \otimes \left(\frac{(n-c)\gamma}{2} \mathbf{e}_t^{\oplus 2}\right) \otimes \mathbf{I}_{s^{c-2}} \otimes \boldsymbol{\beta}, ..., \mathbf{I}_a \otimes \left(\frac{(n-c)\gamma}{c} \mathbf{e}_t^{\oplus c}\right) \otimes \boldsymbol{\beta}\Big]$.

Expressions for the blocks $\mathbf{V}_n$. The matrices $\mathbf{V}_n$ are given by

(a) For $1 \le n \le c$, $\mathbf{V}_n = diag[\mathbf{I}_a \otimes \mathbf{s}_0^{\oplus n}, \mathbf{I}_{at} \otimes \mathbf{s}_0^{\oplus n-1}, ..., \mathbf{I}_{at^{n-1}} \otimes \mathbf{s}_0]$.

(b) For $c+1 \le n \le c+N$, $\mathbf{V}_n = diag[\mathbf{I}_a \otimes \boldsymbol{\alpha} \otimes \mathbf{s}_0^{\oplus c}, \mathbf{I}_{at} \otimes \boldsymbol{\alpha} \otimes \mathbf{s}_0^{\oplus c-1}, ..., \mathbf{I}_{at^{c-1}} \otimes \boldsymbol{\alpha} \otimes \mathbf{s}_0]$.

Expressions for the blocks $\mathbf{W}_n$. The matrices $\mathbf{W}_n$ have the form

(a) For $1 \le n \le c$, $\mathbf{W}_n = diag[\mathbf{I}_a \otimes \mathbf{t}_0 \otimes \mathbf{I}_{s^{n-1}}, \mathbf{I}_a \otimes \mathbf{t}_0^{\oplus 2} \otimes \mathbf{I}_{s^{n-2}}, ..., \mathbf{I}_a \otimes \mathbf{t}_0^{\oplus n}]$.

(b) For $c+1 \leq n \leq c+N$, $\mathbf{W}_n = diag[(n-c)\gamma \mathbf{I}_{as^c}, \mathbf{I}_a \otimes (\mathbf{t}_0\boldsymbol{\alpha}) \otimes \mathbf{I}_{s^{c-1}}, \mathbf{I}_a \otimes (\mathbf{t}_0\boldsymbol{\alpha})^{\oplus 2} \otimes \mathbf{I}_{s^{c-2}}, ..., \mathbf{I}_a \otimes (\mathbf{t}_0\boldsymbol{\alpha})^{\oplus c}]$.

Denote by $\mathbf{x}$ the stationary vector of $\mathcal{X}$, and partition $\mathbf{x}$ by levels into sub-vectors $\mathbf{x}(n)$ for $0 \leq n \leq c+N$. Observe that, since $\mathcal{S}$ is a finite state space, the stationary probabilities of $\mathcal{X}$ exist and are positive. By using Lemma 2 and Theorem 1 of Ref. [26], we find that $\mathbf{x}(n)$ is determined by

$$\begin{aligned}
\mathbf{x}(n) &= \mathbf{x}(0) \prod_{i=0}^{n-1} \left(\mathbf{A}_i \left(-\mathbf{F}_{i+1}^{-1}\right)\right), \text{ for } 1 \leq n \leq c, \\
&= \mathbf{x}(0) \prod_{i=0}^{c-1} \left(\mathbf{A}_i \left(-\mathbf{F}_{i+1}^{-1}\right)\right) \prod_{j=c}^{n-1} \left(\mathbf{A}_c \left(-\mathbf{F}_{j+1}^{-1}\right)\right), \text{ for } c+1 \leq n \leq c+N,
\end{aligned} \quad (6.2)$$

where $\mathbf{x}(0)$ satisfies $\mathbf{x}(0)\mathbf{F}_0 = \mathbf{0}_a^T$ and the equality

$$\begin{aligned}
1 =& \mathbf{x}(0) \left(\mathbf{e}_{J_0} + \sum_{n=1}^{c} \prod_{i=0}^{n-1} \left(\mathbf{A}_i \left(-\mathbf{F}_{i+1}^{-1}\right)\right) \mathbf{e}_{J_n} + \sum_{n=c+1}^{c+N} \prod_{i=0}^{c-1} \left(\mathbf{A}_i \left(-\mathbf{F}_{i+1}^{-1}\right)\right) \right. \\
& \left. \times \prod_{j=c}^{n-1} \left(\mathbf{A}_c \left(-\mathbf{F}_{j+1}^{-1}\right)\right) \mathbf{e}_{J_c} \right).
\end{aligned} \quad (6.3)$$

The matrices $\mathbf{F}_i$ are recursively determined by

$$\begin{aligned}
\mathbf{F}_i &= \mathbf{B}_{c+N} + \mathbf{A}_c, \text{ for } i = c+N, \\
&= \mathbf{B}_i - \mathbf{A}_c \mathbf{F}_{i+1}^{-1} \mathbf{C}_{i+1}, \text{ for } c+1 \leq i \leq c+N-1, \quad (6.4) \\
&= \mathbf{B}_i - \mathbf{A}_i \mathbf{F}_{i+1}^{-1} \mathbf{C}_{i+1}, \text{ for } 0 \leq i \leq c.
\end{aligned}$$

From these results, we can effortlessly obtain expressions for specific probabilistic descriptors of the state of the queue at arbitrary epochs. Some of them are:

(a) The mean number of units in the system is

$$\mathcal{N} = \sum_{n=1}^{c+N} n\mathbf{x}(n)\mathbf{e}_{J_n}.$$

(b) The mean number of low priority units in the system is

$$\mathcal{N}_{low} = \sum_{n=1}^{c} \mathbf{x}(n)\mathbf{w}_n + \sum_{n=c+1}^{c+N} \mathbf{x}(n)\bar{\mathbf{w}}_{n-c},$$

where

$$\mathbf{w}_n = \begin{bmatrix} 0_{as^n} \\ \mathbf{e}_{ats^{n-1}} \\ 2\mathbf{e}_{at^2s^{n-2}} \\ \cdots \\ n\mathbf{e}_{at^n} \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{w}}_n = \begin{bmatrix} n\mathbf{e}_{as^c} \\ (n+1)\mathbf{e}_{ats^{c-1}} \\ \cdots \\ (c+n)\mathbf{e}_{at^c} \end{bmatrix}.$$

(c) The mean number of high priority units in the system is given by $\mathcal{N}_{high} = \mathcal{N} - \mathcal{N}_{low}$.

(d) The blocking probability is

$$P_{blocking} = \lim_{u\to\infty} P(\xi_l(u) + \xi_h(u) = c + N) = \mathbf{x}(c+N)\mathbf{e}_{J_c}.$$

(e) The marginal distribution of the number of units in the buffer. Let q_m be the stationary probability that there are m units in the buffer, for $0 \le m \le N$. Then it is clear that

$$\begin{aligned} q_m &= \sum_{n=0}^{c} \mathbf{x}(n)\mathbf{e}_{J_n}, \text{ for } m = 0, \\ &= \mathbf{x}(c+m)\mathbf{e}_{J_c}, \text{ for } 1 \le m \le N. \end{aligned}$$

Thus the mean number of units in the buffer is given by

$$\mathcal{N}_{buffer} = \sum_{n=1}^{N} n\mathbf{x}(c+n)\mathbf{e}_{J_c}.$$

A point worth mentioning is that, in order to compute $\mathbf{x}$ from (6.2)–(??), we may first calculate $\mathbf{x}(0)$ satisfying $\mathbf{x}(0)\mathbf{F}_0 = \mathbf{0}_a^T$ and $\mathbf{x}(0)\mathbf{e}_a = 1$, and evaluate $\mathbf{x}(n) = \mathbf{x}(n-1)\mathbf{A}_{\min(n-1,c)}(-\mathbf{F}_n^{-1})$ for $1 \leq n \leq c+N$. Then the stationary vector $\mathbf{x}$ corresponds to the vector $[\mathbf{x}(0), ..., \mathbf{x}(c+N)]$ normalized by $\sum_{n=0}^{c+N} \mathbf{x}(n)\mathbf{e}_{J_n} = 1$ and, as a result, the complexity of our solution is $O(\sum_{n=0}^{c-1} J_n^3 + (N+1)J_c^3)$. Clearly we can compute the above probabilistic descriptors at the same time as we are preparing the evaluation of $\mathbf{x}$.

When the value of $c+N$ and the physical dimensions of the blocks $\mathbf{F}_i$ in (6.4) are moderate, the computation of $\mathbf{x}$ may best be done by progressively storing the blocks $\mathbf{A}_{\min(n-1,c)}(-\mathbf{F}_n^{-1})$ for $n = c+N, ..., 1$, and $\mathbf{F}_0$. To that end we need an array of dimension $J_0^2 + \sum_{n=1}^{c} J_{n-1}J_n + NJ_c^2$. We also notice that the maximum number of memory locations for the entries of $\mathbf{x}$ and other eventually defined blocks is $\sum_{n=0}^{c-1} J_n + (N+1)J_c + J_c^2$. In the numerical examples presented in Subsection 4.4, this procedure is seen to work well both regard to numerical accuracy and speed.

Larger values of $c+N$ or larger physical dimensions of $\mathbf{F}_i$ in (6.4) imply more demanding memory requirements. It might be advisable then to write a driver routine where particular blocks $\mathbf{F}_i$ are built each time that they are handled, stored in an amount of memory locations and destroyed immediately after their handling. In such a case, we need $\sum_{n=0}^{c-1} J_n + (N+1)J_c$ memory locations for entries of $\mathbf{x}$ and $2J_c^2$ additional memory locations to store blocks being eventually handled. The corresponding procedure results in an increase of the complexity, but it helps to reduce the effort required to minimize the storage space. More details about the numerical efficiency of other computational algorithms can be found in Section 5 of the paper [26].

6.3. Performance evaluation

6.3.1. Blocking probability at pre-arrival epochs. Based on the above analysis, if we denote by $Q_{blocking}$ a new arrival's blocking probability, which is defined as the probability that a new unit arrives to find the system completely occupied, then we have that

$$Q_{blocking} = \lambda^{-1}\mathbf{x}(0)\prod_{i=0}^{c-1}\left(\mathbf{A}_i(-\mathbf{F}_{i+1}^{-1})\right)\prod_{j=c}^{c+N-1}\left(\mathbf{A}_c(-\mathbf{F}_{j+1}^{-1})\right)\mathbf{\Delta}_c,$$

where the column vector $\mathbf{\Delta}_c$ is defined from

$$\mathbf{\Delta}_n = \begin{bmatrix} (\mathbf{D}_1\mathbf{e}_a)\otimes\mathbf{e}_{s^n} \\ (\mathbf{D}_1\mathbf{e}_a)\otimes\mathbf{e}_{ts^{n-1}} \\ \cdots \\ (\mathbf{D}_1\mathbf{e}_a)\otimes\mathbf{e}_{t^n} \end{bmatrix}, \; 0 \le n \le c.$$

It should be pointed out that clearly the blocking probability $P_{blocking}$ will not necessarily be the new unit's blocking probability $Q_{blocking}$. Indeed, a similar remark can be made for the stationary vectors at pre-arrival and arbitrary times. Let $\mathbf{y}(n)$ be a row vector whose entries are the stationary probabilities that arriving units see the queue in states of the level $l(n)$, for $0 \le n \le c+N$. Then it immediately follows that

$$\begin{aligned}
\mathbf{y}(n) &= \lambda^{-1}\mathbf{x}(0)\mathbf{D}_1, \text{ for } n = 0, \\
&= \lambda^{-1}\mathbf{x}(0)\prod_{i=0}^{n-1}\left(\mathbf{A}_i(-\mathbf{F}_{i+1}^{-1})\right)\mathbf{U}_n\mathbf{\Gamma}_n, \text{ for } 1 \le n \le c-1, \\
&= \lambda^{-1}\mathbf{x}(0)\prod_{i=0}^{c-1}\left(\mathbf{A}_i(-\mathbf{F}_{i+1}^{-1})\right)\mathbf{U}_c, \text{ for } n = c, \\
&= \lambda^{-1}\mathbf{x}(0)\prod_{i=0}^{c-1}\left(\mathbf{A}_i(-\mathbf{F}_{i+1}^{-1})\right)\prod_{j=c}^{n-1}\left(\mathbf{A}_c(-\mathbf{F}_{j+1}^{-1})\right)\mathbf{U}_c,
\end{aligned}$$

for $c+1 \leq n \leq c+N$, with $\boldsymbol{\Gamma}_n = diag[\mathbf{I}_a \otimes \mathbf{e}_t \otimes \mathbf{I}_{s^n}, \mathbf{I}_{at} \otimes \mathbf{e}_t \otimes \mathbf{I}_{s^{n-1}}, ..., \mathbf{I}_{at^n} \otimes \mathbf{e}_t]$, for $1 \leq n \leq c-1$.

Also, if our interest is in the probability that an arriving unit does not wait in the buffer before entering the service facility, then we have

$$P_{nw} = \lambda^{-1}\mathbf{x}(0)\sum_{n=0}^{c-1}\prod_{i=0}^{n-1}\left(\mathbf{A}_i(-\mathbf{F}_{i+1}^{-1})\right)\boldsymbol{\Delta}_n.$$

6.3.2. Departure process. In this subsection, we present the analysis of the departure process, which is defined as the sequence of times $\{\tau_m : m \geq 0\}$ at which units leave the queue due to a service completion or a self-generation of priority, with $\tau_0 \equiv 0$. Its study amounts to the analysis of the inter-departure times $\{\bar{\tau}_m = \tau_m - \tau_{m-1} : m \geq 1\}$. It should be pointed out that the random variables $\bar{\tau}_1$, $\bar{\tau}_2$, ... are identically distributed since $\mathcal{X}$ is positive recurrent. Thus, we focus on $\bar{\tau}_1$ and determine its distribution through the Laplace-Stieltjes transform

$$\Phi(\theta) = E\left[e^{-\theta\bar{\tau}_1}\right], \; Re(\theta) \geq 0.$$

According to the state of the queue at time τ_0, we may write down

$$\Phi(\theta) = \sum_{n=0}^{c+N}\mathbf{x}(n)\phi_n(\theta),$$

where $\phi_n(\theta)$ is a column vector with J_n entries defined as the conditional Laplace-Stieltjes transforms of $\bar{\tau}_1$, given that the state of $\mathcal{X}$ at time τ_0 is in the level $l(n)$, for $0 \leq n \leq c+N$. Partition $\phi_n(\theta)$ into column vectors as follows:

$$\phi_n(\theta) = \begin{bmatrix} \bar{\varphi}(\theta|0,n) \\ \bar{\varphi}(\theta|1,n-1) \\ \cdots \\ \bar{\varphi}(\theta|n,0) \end{bmatrix}, \text{ for } 0 \leq n \leq c,$$

$$= \begin{bmatrix} \tilde{\varphi}(\theta|n-c,c) \\ \tilde{\varphi}(\theta|n-c+1,c-1) \\ \cdots \\ \tilde{\varphi}(\theta|n,0) \end{bmatrix}, \text{ for } c+1 \leq n \leq c+N,$$

where the above sub-vectors $\bar{\varphi}(\theta|m, n-m)$ and $\tilde{\varphi}(\theta|n-c+m, c-m)$ are indexed by states in $\mathcal{L}(m, n-m)$ and $\mathcal{L}'(n-c+m, n-m)$ respectively.

For initial states in $\mathcal{L}(m, n-m)$ for $0 \leq m \leq n \leq c-1$, the departure process can be seen as the time until absorption in an appropriately defined absorbing finite QBD process with Laplace-Stieltjes transform for the time until absorption satisfying

$$\begin{aligned} \bar{\varphi}(\theta|m, n-m) &= \int_0^\infty e^{-(\theta \mathbf{I}_{at^m s^{n-m}} - \mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus n-m})u} du \\ &\times ((\mathbf{D}_1 \otimes \mathbf{I}_{t^m} \otimes \boldsymbol{\alpha} \otimes \mathbf{I}_{s^{n-m}}) \bar{\varphi}(\theta|m+1, n-m) \\ &+ \mathbf{e}_a \otimes ((\mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus n-m}) \mathbf{e}_{t^m s^{n-m}})) . \end{aligned}$$

To prove this equality, note that there are two essential events whose occurrences clearly determine the further evolution of the absorbing QBD process: the arrival of a new unit and a service completion. Since $\mathbf{T}$ and $\mathbf{S}$ are stable, the spectral radius of the matrix $\theta \mathbf{I}_{at^m s^{n-m}} - \mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus n-m}$ is strictly less than one. From this it follows that such a matrix is invertible for $Re(\theta) \geq 0$. As a result we derive the equality

$$\begin{aligned} \bar{\varphi}(\theta|m, n-m) = & \left(\theta \mathbf{I}_{at^m s^{n-m}} - \mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus n-m}\right)^{-1} \\ & \times ((\mathbf{D}_1 \otimes \mathbf{I}_{t^m} \otimes \boldsymbol{\alpha} \otimes \mathbf{I}_{s^{n-m}}) \bar{\varphi}(\theta|m+1, n-m) \\ & + \mathbf{e}_a \otimes ((\mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus n-m}) \mathbf{e}_{t^m s^{n-m}})) , \quad (6.5) \end{aligned}$$

for $0 \leq m \leq n \leq c-1$.

Similarly, for $0 \leq m \leq c$, we have that

$$\tilde{\varphi}(\theta|m, c-m) = \left(\theta \mathbf{I}_{at^m s^{c-m}} - \mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m}\right)^{-1} \times \left((\mathbf{D}_1 \otimes \mathbf{I}_{t^m s^{c-m}})\tilde{\varphi}(\theta|m+1, c-m) + \mathbf{e}_a \otimes ((\mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m})\mathbf{e}_{t^m s^{c-m}})\right). \quad (6.6)$$

For initial states in $\mathcal{L}'(n-c+m, c-m)$, $0 \leq m \leq c < n \leq c+N-1$, we take into account a third essential event, the self-generation of priorities by units in the buffer. Then a first-passage argument leads to

$$\begin{aligned}\tilde{\varphi}(\theta|n-c+m, c-m) &= \left((\theta + (n-c)\gamma)\mathbf{I}_{at^m s^{c-m}} - \mathbf{D}_0 \oplus \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m}\right)^{-1} \\ &\times \left((\mathbf{D}_1 \otimes \mathbf{I}_{t^m s^{c-m}})\tilde{\varphi}(\theta|n-c+m+1, c-m)\right. \\ &\left.+\mathbf{e}_a \otimes ((\mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m})\mathbf{e}_{t^m s^{c-m}}) + (n-c)\gamma \mathbf{e}_{at^m s^{c-m}}\right). \quad (6.7)\end{aligned}$$

Finally, for initial states in $\mathcal{L}'(N+m, c-m)$, $0 \leq m \leq c$, we note that the departure process remains unaltered when new arrivals occur. Then, for $0 \leq m \leq c$, we readily derive

$$\tilde{\varphi}(\theta|N+m, c-m) = \mathbf{e}_a \otimes \left(\left((\theta + N\gamma)\mathbf{I}_{t^m s^{c-m}} - \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m}\right)^{-1} \times \left(N\gamma \mathbf{I}_{t^m s^{c-m}} + \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m}\right) \mathbf{e}_{t^m s^{c-m}}\right). \quad (6.8)$$

Writing down column vectors $\phi_n(\theta)$ from (6.5)–(6.7), we see that $\phi_n(\theta)$ satisfies the recursive formulas

$$\begin{aligned}\phi_n(\theta) &= \bar{\Lambda}_n(\theta)\left(\Theta_n \phi_{n+1}(\theta) + \omega_n\right), \text{ for } 0 \leq n \leq c-1, \quad (6.9) \\ &= \tilde{\Lambda}_{n-c}(\theta)\left(\mathbf{U}_c \phi_{n+1}(\theta) + \chi + (n-c)\gamma \mathbf{e}_{J_c}\right), \quad (6.10)\end{aligned}$$

for $c \leq n \leq c+N-1$, where

$$\begin{aligned}
\bar{\Lambda}_n(\theta) &= [0_{J_n \times as^{n+1}}, diag[\bar{\Sigma}_n(\theta|0), \bar{\Sigma}_n(\theta|1), ..., \bar{\Sigma}_n(\theta|n)]], \\
\tilde{\Lambda}_{n-c}(\theta) &= diag[\tilde{\Sigma}_{n-c}(\theta|0), \tilde{\Sigma}_{n-c}(\theta|1), ..., \tilde{\Sigma}_{n-c}(\theta|c)], \\
\Theta_n &= [0_{J_n \times as^{n+1}}, U_n],
\end{aligned}$$

$$\omega_n = \begin{bmatrix} 0_{as^{n+1}} \\ e_a \otimes (S^{\oplus n} e_{s^n}) \\ e_a \otimes ((T \oplus S^{\oplus n-1}) e_{ts^{n-1}}) \\ \cdots \\ e_a \otimes (T^{\oplus n} e_{t^n}) \end{bmatrix}, \quad \chi = \begin{bmatrix} e_a \otimes (S^{\oplus c} e_{s^c}) \\ e_a \otimes ((T \oplus S^{\oplus c-1}) e_{ts^{c-1}}) \\ \cdot \\ e_a \otimes (T^{\oplus c} e_{t^c}) \end{bmatrix},$$

with the matrices $\bar{\Sigma}_n(\theta|m)$ and $\tilde{\Sigma}_{n-c}(\theta|m)$ defined as follows:

$$\begin{aligned}
\bar{\Sigma}_n(\theta|m) &= (\theta I_{at^m s^{n-m}} - D_0 \oplus T^{\oplus m} \oplus S^{\oplus n-m})^{-1}, \\
\tilde{\Sigma}_{n-c}(\theta|m) &= ((\theta + (n-c)\gamma) I_{at^m s^{c-m}} - D_0 \oplus T^{\oplus m} \oplus S^{\oplus c-m})^{-1}.
\end{aligned}$$

Equations (6.9) and (6.10) allow us to find $\phi_{c+N-1}(\theta)$ once $\phi_{c+N}(\theta)$ is given, $\phi_{c+N-2}(\theta)$ once $\phi_{c+N-1}(\theta)$ is given, and so on. We therefore find after a brief recursion

$$\begin{aligned}
\phi_n(\theta) = &\prod_{i=n}^{c-1} \left(\bar{\Lambda}_i(\theta)\Theta_i\right) \phi_c(\theta) \\
&+ \bar{\Lambda}_n(\theta) \sum_{i=1}^{c-n} \prod_{j=n}^{c-1-i} \left(\Theta_j \bar{\Lambda}_{j+1}(\theta)\right) \omega_{c-i}, \; 0 \leq n \leq c-1,
\end{aligned}$$

$$= \prod_{i=0}^{N-1+c-n} \left(\tilde{\Lambda}_{n-c+i}(\theta)\mathrm{U}_c\right) \phi_{c+N}(\theta)$$

$$+ \tilde{\Lambda}_{n-c}(\theta) \sum_{i=n-c}^{N-1} \prod_{j=1}^{N-1-i} \left(\mathrm{U}_c\tilde{\Lambda}_{n-c+j}(\theta)\right)$$

$$\times (\chi + (N-1-i+n-c)\gamma \mathbf{e}_{J_c}),\ c \leq n \leq c+N-1.$$

To find $\phi_{c+N}(\theta)$, we use the explicit expressions for its sub-vectors given in (6.8).

In view of (6.9) and (6.10), we derive the following expression for the mean length of the interval between successive times of service completion or self-generation of priorities:

$$E[\bar{\tau}_1] = \sum_{n=0}^{c+N} \mathbf{x}(n)\phi_n^{(1)},$$

where $\phi_n^{(1)}$ is the column vector with J_n entries evaluated by the iterative scheme

$$\begin{aligned}
\phi_n^{(1)} &= \bar{\Lambda}_n^{(1)}(\Theta_n \mathbf{e}_{J_n} + \omega_n) + \bar{\Lambda}_n(0)\Theta_n\phi_{n+1}^{(1)},\ 0 \leq n \leq c-1, \\
&= \tilde{\Lambda}_{n-c}^{(1)}(\mathrm{U}_c\mathbf{e}_{J_c} + \chi + (n-c)\gamma\mathbf{e}_{J_c}) + \tilde{\Lambda}_{n-c}(0)\mathrm{U}_c\phi_{n+1}^{(1)}, \\
&\quad c \leq n \leq c+N-1,
\end{aligned}$$

where

$$\begin{aligned}
\bar{\Lambda}_n^{(1)} &= [\mathbf{0}_{J_n \times a s^{n+1}}, diag[\bar{\Sigma}_n^2(0|0), \bar{\Sigma}_n^2(0|1), ..., \bar{\Sigma}_n^2(0|n)]], \\
\tilde{\Lambda}_{n-c}^{(1)} &= diag[\tilde{\Sigma}_{n-c}^2(0|0), \tilde{\Sigma}_{n-c}^2(0|1), ..., \tilde{\Sigma}_{n-c}^2(0|c)].
\end{aligned}$$

Obviously, by (8), the vector $\phi^{(1)}_{c+N}$ is simply obtained by

$$\phi^{(1)}_{c+N} = \begin{bmatrix} \tilde{\varphi}^{(1)}(N,c) \\ \tilde{\varphi}^{(1)}(N+1,c-1) \\ \\ \tilde{\varphi}^{(1)}(c+N,0) \end{bmatrix},$$

with $\tilde{\varphi}^{(1)}(N+m,c-m) = \mathbf{e}_a \otimes ((N\gamma \mathbf{I}_{t^m s^{c-m}} - \mathbf{T}^{\oplus m} \oplus \mathbf{S}^{\oplus c-m})^{-1}\mathbf{e}_{t^m s^{c-m}})$. Higher moments of $\bar{\tau}_1$ may be obtained in a similar fashion.

6.3.3. System state at departures. Here, we investigate the system state at departure epochs. More precisely, our interest is in epochs at which units without being served leave the system, epochs at which new arrivals are lost, departure epochs due to a self-generation of priority, service completion epochs of low priority units and service completion epochs of high priority units.

In order to proceed with the analysis conveniently, decompose $\mathbf{x}(n)$ as $[\mathbf{x}(0,n), \mathbf{x}(1,n-1), ..., \mathbf{x}(n,0)]$ for $0 \le n \le c$, and as $[\mathbf{x}(n-c,c), \mathbf{x}(n-c+1,c-1), ..., \mathbf{x}(n,0)]$ for $c+1 \le n \le c+N$, where $\mathbf{x}(m,n-m)$ is a row vector with $at^m s^{n-m}$ entries, for $0 \le m \le n \le c$, and $\mathbf{x}(n-c+m,c-m)$ is a row vector with $at^m s^{c-m}$ entries, for $0 \le m \le c < n \le c+N$.

For the sequence of departure epochs of units without being served, we observe that states just after such events are in the subset

$$\mathcal{L}(0,c) \cup \bigcup_{m=1}^{N-1} \mathcal{L}'(m,c) \cup \bigcup_{m=0}^{c} \mathcal{L}'(N+m,c-m).$$

Specifically, we need states in $\mathcal{L}(0,c) \cup \cup_{m=1}^{N-1} \mathcal{L}'(m,c)$ to identify departures of high priority units and states in $\cup_{m=0}^{c} \mathcal{L}'(N+m,c-m)$ for arriving units which are lost. By arranging states in lexicographic order and introducing vectors $\mathbf{z}(m,c)$, for $0 \le m \le N-1$, and

$\mathbf{z}(N+m, c-m)$, for $0 \le m \le c$, we readily derive

$$\begin{aligned} \mathbf{z}(m,c) &= (\bar{\delta}+\tilde{\delta})^{-1}(m+1)\gamma\mathbf{x}(m+1,c),\ 0 \le m \le N-1, \\ \mathbf{z}(N+m,c-m) &= (\bar{\delta}+\tilde{\delta})^{-1}\mathbf{x}(N+m,c-m)(\mathbf{D}_1 \otimes \mathbf{I}_{t^m s^{c-m}}), \\ & \quad 0 \le m \le c, \end{aligned}$$

where $\bar{\delta} = \gamma \sum_{m=1}^{N} m\mathbf{x}(m,c)\mathbf{e}_{as^c}$ and $\tilde{\delta} = \sum_{m=0}^{c} \mathbf{x}(N+m,c-m)((\mathbf{D}_1\mathbf{e}_a) \otimes \mathbf{e}_{t^m s^{c-m}})$.

It is a simple matter to distinguish between departures of high priority turned units and of arriving units which are lost. Based on states in $\mathcal{L}(0,c) \cup \cup_{m=1}^{N-1}\mathcal{L}'(m,c)$, the vector

$$\bar{\mathbf{z}}(m,c) = \bar{\delta}^{-1}(m+1)\gamma\mathbf{x}(m+1,c)$$

records stationary probabilities at departure epochs of high priority turned units, for $0 \le m \le N-1$. Similarly, it is immediately obvious that

$$\tilde{\mathbf{z}}(N+m,c-m) = \tilde{\delta}^{-1}\mathbf{x}(N+m,c-m)(\mathbf{D}_1 \otimes \mathbf{I}_{t^m s^{c-m}})$$

records stationary probabilities at epochs at which arriving units are lost, for $0 \le m \le c$.

We can also derive the stationary vector $\mathbf{z}_l$ at service completions of low priority units by noting that just after such events, states are in the levels $l(n)$, for $0 \le n \le c+N-1$. Hence if we decompose $\mathbf{z}_l$ by sub-levels into sub-vectors $\mathbf{z}_l(m, n-m)$ for $0 \le m \le n \le c$, and $\mathbf{z}_l(n-c+m, c-m)$ for $0 \le m \le c < n \le c+N-1$, then we see that

$$\begin{aligned} &\mathbf{z}_l(m,n-m) = \delta_l^{-1}\mathbf{x}(m+1,n-m)(\mathbf{I}_a \otimes \mathbf{t}_0^{\oplus m+1} \otimes \mathbf{I}_{s^{n-m}}), \\ &\qquad\qquad 0 \le m \le n \le c-1, \\ &\mathbf{z}_l(m,c-m) = \delta_l^{-1}\mathbf{x}(m+1,c-m)(\mathbf{I}_a \otimes (\mathbf{t}_0\boldsymbol{\alpha})^{\oplus m} \otimes \mathbf{I}_{s^{c-m}}), 0 \le m \le c, \\ &\mathbf{z}_l(n-c+m,c-m) = \delta_l^{-1}\mathbf{x}(n-c+m+1,c-m) \\ &\qquad (\mathbf{I}_a \otimes (\mathbf{t}_0\boldsymbol{\alpha})^{\oplus m} \otimes \mathbf{I}_{s^{c-m}}), 0 \le m \le c < n \le c+N-1, \end{aligned}$$

where

$$\delta_l = \sum_{n=0}^{c-1}\sum_{m=0}^{n} \mathbf{x}(m+1, n-m)\left(\mathbf{e}_a \otimes \left(\mathbf{t}_0^{\oplus m+1}\mathbf{e}_{t^m}\right) \otimes \mathbf{e}_{s^{n-m}}\right)$$
$$+ \sum_{n=c}^{c+N-1}\sum_{m=0}^{c} \mathbf{x}(n-c+m+1, c-m)\left(\mathbf{e}_a \otimes \left((\mathbf{t}_0\boldsymbol{\alpha})^{\oplus m}\mathbf{e}_{t^m}\right) \otimes \mathbf{e}_{s^{c-m}}\right).$$

For service completion epochs of high priority units, the corresponding stationary vector $\mathbf{z}_h$ has sub-vectors of the form

$$\mathbf{z}_h(m, n-m) = \delta_h^{-1}\mathbf{x}(m, n-m+1)(\mathbf{I}_{at^m} \otimes \mathbf{s}_0^{\oplus n-m+1}),$$
$$0 \le m \le n \le c-1,$$
$$\mathbf{z}_h(m, c-m) = \delta_h^{-1}\mathbf{x}(m, c-m+1)(\mathbf{I}_{at^{m-1}} \otimes \boldsymbol{\alpha} \otimes \mathbf{s}_0^{\oplus c-m+1}), 1 \le m \le c,$$
$$\mathbf{z}_h(n-c+m, c-m) = \delta_h^{-1}\mathbf{x}(n-c+m, c-m+1)$$
$$(\mathbf{I}_{at^{m-1}} \otimes \boldsymbol{\alpha} \otimes \mathbf{s}_0^{\oplus c-m+1}), 1 \le m \le c < n \le c+N-1,$$

where

$$\delta_h = \sum_{n=0}^{c-1}\sum_{m=0}^{n} \mathbf{x}(m, n-m+1)\left(\mathbf{e}_{at^m} \otimes \left(\mathbf{s}_0^{\oplus n-m+1}\mathbf{e}_{s^{n-m}}\right)\right)$$
$$+ \sum_{n=c}^{c+N-1}\sum_{m=0}^{c} \mathbf{x}(n-c+m, c-m+1)\left(\mathbf{e}_{at^{m-1}} \otimes \left(\mathbf{s}_0^{\oplus c-m+1}\mathbf{e}_{s^{c-m}}\right)\right).$$

6.4. Effect of the self-generation of priorities

We expect that the effect of self-generating a priority on the current amount of work in the queue is threefold:

		$\lambda' = 0.5$			$\lambda' = 2.0$			$\lambda' = 10.0$		
γ	N	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$
0.01	5	$< 10^{-7}$	$< 10^{-5}$	0.0078	0.0081	0.1753	0.5112	0.6025	0.7880	0.8893
	10	$< 10^{-11}$	$< 10^{-8}$	0.0003	0.0004	0.1409	0.5079	0.6059	0.7889	0.8888
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-5}$	0.1311	0.5056	0.6096	0.7899	0.8882
0.1	5	$< 10^{-7}$	$< 10^{-5}$	0.0071	0.0107	0.1901	0.4937	0.6315	0.7960	0.8838
	10	$< 10^{-11}$	$< 10^{-7}$	0.0002	0.0015	0.1727	0.4564	0.6628	0.8013	0.8738
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	0.0003	0.1680	0.4040	0.6846	0.8001	0.8591
0.25	5	$< 10^{-7}$	0.0001	0.0057	0.0163	0.2060	0.4485	0.6717	0.8016	0.8678
	10	$< 10^{-10}$	$< 10^{-7}$	$< 10^{-5}$	0.0045	0.1510	0.3018	0.6982	0.7809	0.8201
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	0.0011	0.0742	0.1491	0.6807	0.7358	0.7596
0.5	5	$< 10^{-6}$	0.0001	0.0034	0.0250	0.1918	0.3428	0.6969	0.7826	0.8224
	10	$< 10^{-10}$	$< 10^{-8}$	$< 10^{-6}$	0.0046	0.0554	0.0987	0.6461	0.6855	0.7009
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	0.0002	0.0049	0.0100	0.5494	0.5706	0.5776
0.75	5	$< 10^{-6}$	0.0001	0.0020	0.0281	0.1523	0.2456	0.6838	0.7421	0.7668
	10	$< 10^{-10}$	$< 10^{-8}$	$< 10^{-7}$	0.0021	0.0159	0.0275	0.5562	0.5780	0.5852
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-5}$	0.0003	0.0006	0.4053	0.4155	0.4184
1.0	5	$< 10^{-6}$	$< 10^{-5}$	0.0012	0.0264	0.1139	0.1738	0.6525	0.6941	0.7101
	10	$< 10^{-11}$	$< 10^{-9}$	$< 10^{-7}$	0.0008	0.0047	0.0080	0.4643	0.4776	0.4815
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-6}$	$< 10^{-5}$	$< 10^{-5}$	0.2804	0.2860	0.2874
2.5	5	$< 10^{-7}$	$< 10^{-5}$	0.0001	0.0080	0.0211	0.0290	0.4213	0.4317	0.4346
	10	$< 10^{-12}$	$< 10^{-11}$	$< 10^{-10}$	$< 10^{-6}$	$< 10^{-5}$	$< 10^{-5}$	0.1129	0.1149	0.1153
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-10}$	$< 10^{-9}$	$< 10^{-9}$	0.0111	0.0113	0.0114
5.0	5	$< 10^{-7}$	$< 10^{-6}$	$< 10^{-5}$	0.0014	0.0029	0.0038	0.1880	0.1911	0.1919
	10	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-8}$	$< 10^{-7}$	$< 10^{-7}$	0.0076	0.0077	0.0078
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
10.0	5	$< 10^{-8}$	$< 10^{-7}$	$< 10^{-7}$	0.0001	0.0002	0.0003	0.0442	0.0449	0.0450
	10	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-9}$	0.0001	0.0001	0.0001
	15	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-8}$	$< 10^{-8}$	$< 10^{-8}$

(i) Depending on whether the mean service time of a high priority unit is less than or greater than that of a low priority unit, the current amount of work in the queue may decrease or increase if there is an idle server available for the high priority turned unit.

(ii) If a low priority unit is pushed out, then the amount of work associated with such a unit increases somewhat as it needs to restart service, though this may be compensated at a collective level by a possible decrease of work associated with the high priority unit.

(iii) A unit may leave without receiving service, thereby decreasing the current amount of work.

To study the influence of γ on the current amount of work, we focus on the blocking probability $P_{blocking}$ and the mean values $\mathcal{N}$, $\mathcal{N}_{high}$ and $\mathcal{N}_{buffer}$. For the numerical examples, we assume that the stream of units is a renewal process with inter-renewal intervals governed by the hyper-exponential law with density function

$$f(x) = \sum_{i=1}^{a} p_i \lambda_i e^{-\lambda_i x}, \; x > 0,$$

with $p_i > 0$, $\sum_{i=1}^{a} p_i = 1$ and $\lambda_i > 0$. We assume that service times of low priority units follow an Erlang$_t(\nu_1)$ distribution, with $\nu_1 > 0$, and that an Erlang$_s(\nu_2)$ distribution, with $\nu_2 > 0$, governs service times of high priority units.

Tables 1-4 list values of $P_{blocking}$, $\mathcal{N}$, $\mathcal{N}_{high}$ and $\mathcal{N}_{buffer}$, for queues with $c = 3$, $a = 4$, $t = 3$, $s = 3$. To explore the effect of (i) on the current amount of work, we assume mean service times of low and high priority units, denoted by μ_l and μ_h respectively, satisfying $\mu_h = q\mu_l$, where $q = 0.5$, 1.0 and 2.0. Our numerical examples are reported for $N = 5$, 10 and 15, and arrival intensities $\lambda_i = i\lambda'$, with $\lambda' = 0.5$, 2.0 and 10.0 (equivalently, with arrival rates of the point process of units $\lambda = 0.96$, 3.84 and 19.2 respectively). The mean service time of high priority units is assumed to be $\mu_h = 0.9$, and initial probabilities $p_i = 0.25$, for $1 \leq i \leq 4$.

When γ is small, it is clear that (i) and (ii) will dominate. Thus, for $q = 0.5$ and 1.0, increasing values of γ are the cause of higher values of $P_{blocking}$, $\mathcal{N}$, $\mathcal{N}_{high}$ and $\mathcal{N}_{buffer}$, whereas for $q = 2.0$ the current amount of work seems to decrease when a self-generation takes place (meaning that the decrease of work associated with high priority units compensates the increase of work associated with low priority units). Such a decrease in the case $q = 2.0$ implies lower values of the descriptors $P_{blocking}$, $\mathcal{N}$, $\mathcal{N}_{high}$ and $\mathcal{N}_{buffer}$, as the reader may note from Tables 1-4. At some point, if γ is large enough, the effect of (iii) will become more influential and will start to dominate, meaning, even if $q \leq 1$, that an increase of γ will reduce the various performance measures.

		$\lambda' = 0.5$			$\lambda' = 2.0$			$\lambda' = 10.0$		
γ	N	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$
0.01	5	0.4345	0.8968	2.1105	2.1184	5.2249	7.1425	7.3922	7.7427	7.8785
	10	0.4345	0.8971	2.2022	2.2180	8.9397	12.123	12.398	12.744	12.877
	15	0.4345	0.8971	2.2092	2.2272	13.098	17.115	17.406	17.745	17.877
0.1	5	0.4347	0.8987	2.0899	2.1751	5.3230	7.0906	7.4519	7.7542	7.8716
	10	0.4347	0.8991	2.1615	2.3481	9.3750	11.941	12.514	12.761	12.858
	15	0.4347	0.8991	2.1645	2.3965	13.932	16.694	17.557	17.759	17.838
0.25	5	0.4350	0.9019	2.0484	2.2852	5.4325	6.9397	7.5310	7.7620	7.8509
	10	0.4350	0.9024	2.0878	2.6425	9.2289	11.146	12.584	12.727	12.781
	15	0.4350	0.9024	2.0882	2.8026	12.170	14.483	17.548	17.645	17.679
0.5	5	0.4355	0.9061	1.9683	2.4644	5.3634	6.5144	7.5822	7.7319	7.7864
	10	0.4356	0.9066	1.9811	2.9378	7.6630	9.1516	12.474	12.547	12.570
	15	0.4356	0.9066	1.9811	3.0206	8.2913	10.023	17.205	17.252	17.265
0.75	5	0.4361	0.9085	1.8953	2.5766	5.0973	6.0290	7.5601	7.6638	7.6989
	10	0.4361	0.9088	1.9000	2.9323	6.3229	7.4976	12.243	12.291	12.304
	15	0.4361	0.9088	1.9000	2.9539	6.4405	7.6720	16.650	16.685	16.694
1.0	5	0.4367	0.9093	1.8347	2.6158	4.7913	5.5857	7.4990	7.5762	7.6003
	10	0.4367	0.9095	1.8367	2.8462	5.4587	6.4060	11.947	11.985	11.994
	15	0.4367	0.9095	1.8367	2.8519	5.4845	6.4449	15.915	15.949	15.957
2.5	5	0.4392	0.9013	1.6266	2.4223	3.6130	4.0685	6.9068	6.9411	6.9494
	10	0.4392	0.9014	1.6267	2.4433	3.6587	4.1264	9.6012	9.6353	9.6431
	15	0.4392	0.9014	1.6267	2.4433	3.6587	4.1265	10.520	10.562	10.571
5.0	5	0.4415	0.8845	1.4875	2.1778	2.9441	3.2400	5.9301	5.9574	5.9639
	10	0.4415	0.8845	1.4875	2.1794	2.9473	3.2440	6.7480	6.7804	6.7881
	15	0.4415	0.8845	1.4875	2.1794	2.9473	3.2440	6.7836	6.8166	6.8244
10.0	5	0.4432	0.8656	1.3891	1.9935	2.5363	2.7473	4.7328	4.7545	4.7597
	10	0.4432	0.8656	1.3891	1.9936	2.5364	2.7475	4.8369	4.8596	4.8651
	15	0.4432	0.8656	1.3891	1.9936	2.5364	2.7475	4.8371	4.8599	4.8654

In Tables 1-4, we also notice that the increase of the arrival rate λ and the decrease of the buffer capacity N have an increasing effect on $P_{blocking}$, $\mathcal{N}$, $\mathcal{N}_{high}$ and $\mathcal{N}_{buffer}$, which corroborates the above intuitive explanation.

		$\lambda' = 0.5$			$\lambda' = 2.0$			$\lambda' = 10.0$		
γ	N	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$
0.01	5	$< 10^{-5}$	0.0002	0.0035	0.0036	0.0221	0.0373	0.0395	0.0426	0.0439
	10	$< 10^{-5}$	0.0002	0.0042	0.0043	0.0542	0.0821	0.0845	0.0876	0.0888
	15	$< 10^{-5}$	0.0002	0.0043	0.0044	0.0912	0.1269	0.1296	0.1326	0.1338
0.1	5	0.0002	0.0030	0.0347	0.0396	0.2281	0.3661	0.3973	0.4239	0.4343
	10	0.0002	0.0030	0.0399	0.0523	0.5601	0.7702	0.8154	0.8358	0.8437
	15	0.0002	0.0030	0.0401	0.0558	0.9039	1.1112	1.1680	1.1828	1.1886
0.25	5	0.0005	0.0078	0.0793	0.1124	0.5639	0.8347	0.9456	0.9911	1.0086
	10	0.0005	0.0079	0.0856	0.1672	1.1661	1.4658	1.6432	1.6634	1.6709
	15	0.0005	0.0079	0.0856	0.1885	1.4974	1.7829	2.0307	2.0401	2.0435
0.5	5	0.0012	0.0160	0.1301	0.2511	0.9669	1.3013	1.5882	1.6320	1.6477
	10	0.0012	0.0161	0.1332	0.3565	1.4400	1.7765	2.2343	2.2451	2.2486
	15	0.0012	0.0161	0.1332	0.3704	1.5192	1.8651	2.4874	2.4913	2.4923
0.75	5	0.0018	0.0236	0.1606	0.3756	1.1673	1.4990	1.9594	1.9931	2.0043
	10	0.0018	0.0237	0.1619	0.4671	1.4402	1.7794	2.4703	2.4768	2.4786
	15	0.0018	0.0237	0.1619	0.4709	1.4558	1.7982	2.6445	2.6468	2.6474
1.0	5	0.0025	0.0305	0.1804	0.4682	1.2545	1.5740	2.1805	2.2063	2.2142
	10	0.0025	0.0305	0.1810	0.5316	1.4083	1.7361	2.5875	2.5922	2.5934
	15	0.0025	0.0305	0.1810	0.5327	1.4118	1.7405	2.7158	2.7177	2.7182
2.5	5	0.0064	0.0578	0.2286	0.6725	1.2921	1.5517	2.5881	2.6003	2.6032
	10	0.0064	0.0578	0.2286	0.6784	1.3027	1.5636	2.7434	2.7489	2.7502
	15	0.0064	0.0578	0.2286	0.6784	1.3027	1.5636	2.7661	2.7712	2.7723
5.0	5	0.0115	0.0796	0.2516	0.7612	1.2629	1.4754	2.6642	2.6765	2.6794
	10	0.0115	0.0796	0.2516	0.7616	1.2636	1.4761	2.7094	2.7200	2.7226
	15	0.0115	0.0796	0.2516	0.7616	1.2636	1.4761	2.7104	2.7210	2.7235
10.0	5	0.0179	0.0982	0.2670	0.8394	1.2611	1.4394	2.6416	2.6549	2.6582
	10	0.0179	0.0982	0.2670	0.8394	1.2611	1.4394	2.6475	2.6607	2.6639
	15	0.0179	0.0982	0.2670	0.8394	1.2611	1.4394	2.6475	2.6608	2.6640

		$\lambda' = 0.5$			$\lambda' = 2.0$			$\lambda' = 10.0$		
γ	N	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$	$q = 0.5$	$q = 1$	$q = 2$
0.01	5	0.0025	0.0328	0.3999	0.4051	2.4647	4.1506	4.3942	4.7427	4.8785
	10	0.0025	0.0330	0.4761	0.4879	6.0307	9.1232	9.3982	9.7442	9.8778
	15	0.0025	0.0330	0.4824	0.4962	10.137	14.115	14.406	14.745	14.877
0.1	5	0.0025	0.0336	0.3867	0.4424	2.5504	4.1000	4.4535	4.7542	4.8716
	10	0.0025	0.0339	0.4461	0.5881	6.4493	8.9422	9.5142	9.7617	9.8589
	15	0.0025	0.0339	0.4487	0.6320	10.958	13.694	14.557	14.759	14.838
0.25	5	0.0026	0.0351	0.3602	0.5166	2.6457	3.9536	4.5320	4.7621	4.8509
	10	0.0026	0.0355	0.3927	0.8205	6.3032	8.1479	9.5847	9.7278	9.7815
	15	0.0026	0.0355	0.3931	0.9660	9.2124	11.483	14.548	14.645	14.679
0.5	5	0.0027	0.0369	0.3104	0.6369	2.5822	3.5463	4.5829	4.7319	4.7864
	10	0.0027	0.0373	0.3210	1.0402	4.7906	6.1659	9.4745	9.5473	9.5707
	15	0.0027	0.0373	0.3210	1.1155	5.4087	7.0355	14.205	14.252	14.265
0.75	5	0.0028	0.0378	0.2676	0.7072	2.3479	3.0898	4.5608	4.6639	4.6989
	10	0.0028	0.0380	0.2715	1.0099	3.5185	4.5405	9.2431	9.2910	9.3043
	15	0.0028	0.0380	0.2715	1.0295	3.6335	4.7140	13.650	13.685	13.694
1.0	5	0.0030	0.0378	0.2343	0.7242	2.0824	2.6810	4.4999	4.5764	4.6005
	10	0.0030	0.0379	0.2360	0.9202	2.7163	3.4864	8.9474	8.9851	8.9945
	15	0.0030	0.0379	0.2360	0.9254	2.7414	3.5249	12.915	12.949	12.957
2.5	5	0.0034	0.0329	0.1358	0.5407	1.1053	1.3587	3.9119	3.9443	3.9523
	10	0.0034	0.0329	0.1358	0.5588	1.1477	1.4140	6.6031	6.6366	6.6442
	15	0.0034	0.0329	0.1358	0.5588	1.1478	1.4141	7.5224	7.5634	7.5728
5.0	5	0.0033	0.0249	0.0812	0.3445	0.5986	0.7116	2.9521	2.9751	2.9806
	10	0.0033	0.0249	0.0812	0.3460	0.6016	0.7153	3.7662	3.7950	3.8019
	15	0.0033	0.0249	0.0812	0.3460	0.6016	0.7153	3.8017	3.8311	3.8381
10.0	5	0.0028	0.0163	0.0452	0.2021	0.3132	0.3615	1.7962	1.8109	1.8144
	10	0.0028	0.0163	0.0452	0.2022	0.3133	0.3617	1.8992	1.9150	1.9188
	15	0.0028	0.0163	0.0452	0.2022	0.3133	0.3617	1.8995	1.9153	1.9191

CHAPTER 7

Retrial queues with self generation of priority of orbital customers

In this chapter we consider a service system with waiting space restricted to one for a special class of customers called priority generated customers. The system consists of one server. If the server is idle at an arrival epoch then that customer is taken for service immediately. Otherwise it proceeds to an orbit of infinite capacity. Each customer in orbit try to access the server at a constant rate θ. Hence if there are n customers the retrial rate is $n\theta$. In addition each customer in orbit generate priority at a constant rate β. Such a customer is termed as priority generated customer. This unit is immediately transferred to the service station provided no such customer is already in wait there. On the other hand such customers leave the system for ever if already a priority generated unit is waiting. Service discipline is non-pre-emptive. That is a customer in service, even when it is ordinary (not a priority generated one), is given full service before the next one is taken for service.

This class of queues occurs in emergency situations (for example in hospitals). For further details one may refer to Krishnamoorthy, Viswanath and Deepak [**36**].

This chapter is arranged as follows : Section 7.1 deals with the mathematical modelling and prove that the system is always stable. Section 7.2 provides steady state distribution of the system. In section 7.3 some performance measures are provided. Also a few numerical illustrations are given in section 7.4.

7.1. Mathematical modelling

Customers arrive to a single server facility according to a Markovian arrival process (MAP) with representation (D_0, D_1) of order m_3. All customers at the time of their arrival are treated as 'ordinary'. Service to ordinary customers is according to a Markovian

Service rule (MSP—Markovian Service Process) with representation (S_1^0, S_1^1) of order m_1 and service to priority generated customers is also according to a MSP with representation (S_2^0, S_2^1) of order m_2. Systems having MSP have been studied by Bocharov [9]. An MSP of order m with representation S^0, S^1 can be described as follows.

Suppose the underlying Markov chain has state space labelled $\{1, 2, \ldots, m\}$ and generator matrix $Q^* = (q_{ij})$. Let the chain be irreducible. After a Sojourn in state i which is exponentially distributed with parameter $\lambda_i \geq -q_{ii}$, one of the following two events could occur

(a) with probability $P_{ij}(1)$, a transition to state j occurs which corresponds to a service completion.

(b) with probability $P_{ij}(0)$, a transition to state j $(j \neq i)$ occurs without a service completion.

If a service completion occurs with a transition to state j and if there is no customer waiting to be served, then we assume that the Markov Chain stays in the state j until another service starts (*ie.*, the chain is assumed to be freezed in state j). When a new service starts, the chain also gets started in state j, and proceeds as described above. We define the matrices

$$S^k = \left(d_{ij}(k)\right) \text{ for } k = 0, 1 \text{ where } d_{ii}(0) = -\lambda_i, \quad 1 \leq i \leq m\,;$$

$d_{ij}(0) = \lambda_i P_{ij}(0)$, $j \neq i$, $1 \leq i, j \leq m$ and $d_{ij}(1) = \lambda_i P_{ij}(1)$, for $1 \leq i, j \leq m$ and $S^0 + S^1 = Q^*$.

Assuming S^0 to be a nonsingular matrix we notice that the service times are finite with probability 1.

Let $N_1(t)$ denote the number of customers in the orbit, and $N_2(t)$, the number of priority customers in the system including the one getting service, if any, at time t. Note that $N_2(t) = 0, 1$ or 2.

Let

$$I(t) = \begin{cases} 0 & \text{if the server is idle at time } t \\ 1 & \text{if an ordinary customer is getting service at time } t \\ 2 & \text{if a priority generated customer is getting service, at time } t. \end{cases}$$

Let $v_1(t)$ and $v_2(t)$ denote the phases of the service processes of ordinary and priority customers, respectively; and $v_3(t)$ denote the phase of the arrival process at time t. Let $X(t) = (N_1(t), N_2(t), I(t), v_1(t), v_2(t), v_3(t))$. Then $\{X(t) | t \geq 0\}$ forms a continuous time Markov chain with state space,

$$\{(i, 0, j, k_1, k_2, k_3) \mid i \geq 0\ ;\ \ j = 0, 1\ ; 1 \leq k_l \leq m_l,\ l = 1, 2, 3\}$$

$$\cup \{(i, 1, j, k_1, k_2, k_3) \mid i \geq 0\ ;\ \ j = 1, 2\ ; 1 \leq k_l \leq m_l,\ l = 1, 2, 3\}$$

$$\cup \{(i, 2, 2, k_1, k_2, k_3) \mid i \geq 0\ ;\ \ 1 \leq k_l \leq m_l,\ l = 1, 2, 3\}$$

Partitioning the above state space into levels $\underline{i}$, where each level $\underline{i}$ correspond to i customers in the orbit, we get the infinitesimal generator of the above Markov chain as

$$Q = \begin{bmatrix} A_{10} & A_0 & 0 & 0 & \dots \\ A_{21} & A_{11} & A_0 & 0 & \dots \\ 0 & A_{22} & A_{12} & A_0 & \dots \\ 0 & 0 & A_{23} & A_{13} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where

$$A_0 = \begin{bmatrix} 0 & 0 \\ 0 & I_4 \end{bmatrix} \otimes \left(I_{m_1 m_2} \otimes D_1 \right),$$

$$A_{2i} = \begin{bmatrix} 0 & i\theta I_M & 0 & i\beta I_M & 0 \\ 0 & 0 & i\beta I_M & 0 & 0 \\ 0 & 0 & i\beta I_M & 0 & 0 \\ 0 & 0 & 0 & 0 & i\beta I_M \\ 0 & 0 & 0 & 0 & i\beta I_M \end{bmatrix}, \qquad i \geq 1,$$

where $M = m_1 m_2 m_3$. For $i \geq 0$,

$$A_{1i} = \begin{bmatrix} B_1 & B_2 & 0 & 0 & 0 \\ B_3 & B_4 & 0 & 0 & 0 \\ 0 & 0 & B_4 & B_3 & 0 \\ B_5 & 0 & 0 & B_6 & 0 \\ 0 & 0 & 0 & B_5 & B_6 \end{bmatrix}$$

where $B_1 = (I_{m_1 m_2} \otimes D_0) - i(\theta + \beta) I_M$, $B_2 = I_{m_1 m_2} \otimes D_1$, $B_3 = S_1^1 \otimes I_{m_2 m_3}$, $B_4 = \left[\left(S_1^0 \otimes I_{m_2}\right) \oplus D_0\right] - i\beta I_M$, $B_5 = (I_{m_1} \otimes S_2^1) \otimes I_{m_3}$ and $B_6 = \left[\left(I_{m_1} \otimes S_2^0\right) \oplus D_0\right] - i\beta I_M$

System stability

THEOREM 7.1. *The system under discussion is always stable.*

PROOF. To prove the theorem, we use a result due to Tweedie [57]. Consider the Lyapunov test function defined by $\phi(s) = i$ if s is a state belonging to level i. The mean drift y_s for a state s belonging to level i is given by

$$\begin{aligned} y_s &= \sum_{p \neq s} q_{sp}[\phi(p) - \phi(s)] \\ &= \sum_{s'} q_{ss'}(\phi(s') - \phi(s)) + \sum_{s''} q_{ss''}\left(\phi(s'') - \phi(s)\right) + \sum_{s'''} q_{ss'''}\left(\phi(s''') - \phi(s)\right) \end{aligned}$$

where s', s'', s''' varies over the states belonging to levels $i-1, i, i+1$ respectively. Then by definition of ϕ, $\phi(s) = i$, $\phi(s') = i - 1$, $\phi(s'') = i$, $\phi(s''') = i + 1$

So that

$$y_s = -\sum_{s'} q_{ss'} + \sum_{s'''} q_{ss'''}$$

$$= \begin{cases} -i(\theta+\beta) + \sum_{s'''} q_{ss'''} & \text{if } s \text{ is a state at which the server is idle} \\ -i\beta + \sum_{s'''} q_{ss'''} & \text{otherwise} \end{cases}$$

We note that $\sum_{s'''} q_{ss'''}$ is bounded by some fixed constant for any s in any level $i \geq 1$. So, let $\sum_{s'''} q_{ss'''} < \kappa$, for some real number $\kappa > 0$, for all states s belonging to level $i \geq 1$. Also since $1 - \delta > 0$, for any $\epsilon > 0$, we can find N' large enough that $y_s < -\epsilon$ for any s belonging to level $i \geq N'$.

Hence by Tweedie's result, the theorem follows. □

REMARK 7.1. *The above theorem can be proved also by noticing the fact that the queueing system under discussion is very much similar to an infinite server queue which is always stable.*

7.2. Steady state distribution

Since the process under consideration is an LDQBD, to calculate the steady state distribution, which always exists, we use the method described in Bright and Taylor [**13**].

By partitioning the steady state probability vector $\mathbf{x}$ as $\mathbf{x} = (x(0), x(1), x(2), \ldots)$ we can write

$$x(k) = x(0) \prod_{l=0}^{k-1} R_l \quad \text{for } k \geq 1$$

where the family of matrices $\{R_k; k \geq 0\}$ are the minimal non-negative solutions to the system of equations

$$A_0 + R_k A_{1k+1} + R_k R_{k+1} A_{2,k+2} = 0, k \geq 0. \tag{7.1}$$

and $x(0)$ is calculated by solving

$$x(0)[A_{10} + R_0 A_{21}] = 0 \tag{7.2}$$

such that

$$x(0)e + x(0)\sum_{k=1}^{\infty}\left[\prod_{l=0}^{k-1} R_l\right] e = 1. \tag{7.3}$$

The calculation of the above infinite sum does not seem feasible. So we approximate $x(k)$s by $x_{K^*}(k)$s where $(x_{K^*}(k))_j$, $0 \le k \le K^*$, $1 \le j \le 5m_1m_2m_3$, is defined as the stationary probability that the Markov chain $X(t)$ is in state (k, j) of level k, conditional on $X(t)$ being in the set $\{(i, j) \mid 0 \le i \le K^*, 1 \le j \le 5m_1m_2m_3\}$. Thus $x_{K^*}(k), 0 \le k \le K^*$ is given by

$$x_{K^*}(k) = x_{K^*}(0)\prod_{l=0}^{k-1} R_l \tag{7.4}$$

where $x_{k^*}(0)$ is found such that it satisfies (7.2) and

$$x_{K^*}(0)\mathrm{e} + x_{K^*}(0)\left[\sum_{k=1}^{K^*}\left[\prod_{l=0}^{k-1} R_l\right]\right]\mathrm{e} = 1. \tag{7.5}$$

Here we have for all $i \ge 1$ and for all k, there exists j such that $(A_{2i})_{k,j} > 0$. So we can construct a process $\bar{X}(t)$ which stochastically dominates $X(t)$ and can use it to find the truncation level K^* in the same way as in Bright and Taylor [13] as follows. The dominating process $\bar{X}(t)$ has generator

$$\bar{Q} = \begin{bmatrix} A_{10} & A_0 & 0 & 0 & 0 & \dots \\ 0 & \bar{A}_{11} & \bar{A}_0 & 0 & 0 & \dots \\ 0 & \bar{A}_{22} & \bar{A}_{12} & \bar{A}_0 & 0 & \dots \\ 0 & 0 & \bar{A}_{23} & \bar{A}_{13} & \bar{A}_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

where
$(\bar{A}_0)_{i,j} = \frac{1}{C}[(A_0 e)_{\max}]$, $(\bar{A}_{2k})_{i,j} = \frac{1}{C}\Big((A_{2,k-1})\mathrm{e}\Big)_{\min}$ for $k \geq 2$, $(\bar{A}_{1k})_{ij} = (A_{1k})_{ij}$, $j \neq i$, $k \geq 1$; and $C = 5m_1m_2m_3$ is the dimension of a level $i \geq 0$.
Since computation of the sequence of matrices $\{R_k\}$ occurring in (7.4) is laborious, requiring tremendous storage space, we use the K^* obtained by the above procedure to define the truncation level for employing the Neuts-Rao [45] procedure in the numerical calculations. Thus we combine the advantages in the two procedures and at the same time get ourselves freed from cumbersome calculations. Besides we are able to maintain atleast the same level of accuracy as obtained in the above two procedures.

Next we discuss a few system performance measures.

7.3. System performance measures

We partition the steady state probability vector x as

$$\mathbf{x} = (x(0), x(1), x(2), \ldots).$$

where each $x(i)$ is partitioned by sublevels as

$$x(i) = \big(y_i(0,0), y_i(0,1), y_i(1,1), y_i(1,2), y_i(2,2)\big)$$

Here $y_i(j,k)$ is a row vector containing $m_1m_2m_3$ entries which corresponds to $N_2(t) = j$ and $I(t) = k$. The following are the performance measures we concentrate on.

(i) The probability a_i that there are i customers in orbit is given by $a_i = x(i)\mathrm{e}$.

(ii) The mean number of customers in the orbit:

$$N_{\text{orbit}} = \sum_{i=1}^{\infty} i x(i)\mathrm{e}.$$

(iii) The probability mass function b_j, $j = 0, 1, 2$ of the number of priority customers in the system

$$b_0 = \sum_{i=0}^{\infty} \sum_{k=0,1} y_i(0,k)\mathrm{e},$$

$$b_1 = \sum_{i=0}^{\infty} \sum_{k=1,2} y_i(1,k)\mathrm{e},$$

$$b_2 = \sum_{i=0}^{\infty} y_i(2,2)\mathrm{e},$$

(iv) The probability that the server is idle is $\mathcal{P}_{\text{idle}} = \sum_{i=0}^{\infty} y_i(0,0)\mathrm{e}$

(v) The overall rate at which the orbiting customers retry for service is given by $\theta_1^* = \theta N_{\text{orbit}}$.

(vi) The rate at which the orbiting customers successfully reach a free server is given by

$$\theta_2^* = \theta\left[\sum_{i=1}^{\infty} i y_i(0,0)\mathrm{e}\right]$$

(vii) The fraction of successful rate of retrials is given by

$$\theta_3^* = \frac{\theta_2^*}{\theta_1^*}$$

7.4. Numerical illustration

$$D_0 = \begin{bmatrix} -4.05 & 1.55 \\ 3.5 & -5.5 \end{bmatrix} \qquad D1 = \begin{bmatrix} 2.05 & 0.45 \\ 1.0 & 1.0 \end{bmatrix} \tag{7.6}$$

Fundamental arrival rate for (7.6) is 2.346

Correlation $= 0.29 \times 10^{-3}$

$$D_0 = \begin{bmatrix} -5.5 & 3.5 \\ 1.0 & -3.5 \end{bmatrix} \qquad D1 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.5 \end{bmatrix} \tag{7.7}$$

Fundamental arrival rate for (7.7) is 2.346

Correlation $= -0.29 \times 10^{-3}$

$$S_{10} = \begin{bmatrix} -5.06 & 2.06 \\ 4.0 & -6.5 \end{bmatrix} \qquad S_1 = \begin{bmatrix} 2.56 & 0.44 \\ 1.0 & 1.5 \end{bmatrix} \tag{7.8}$$

Fundamental service rate for (7.8) is 2.833

Correlation $= 0.42 \times 10^{-3}$

$$S_{10} = \begin{bmatrix} -6.5 & 4.0 \\ 1.5 & -4.5 \end{bmatrix} \qquad S_1 = \begin{bmatrix} 1.5 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \tag{7.9}$$

Fundamental service rate for (7.9) is 2.833

Correlation $= -0.42 \times 10^{-3}$

$$S_{20} = \begin{bmatrix} -5.15 & 2.1 \\ 4.05 & -6.6 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 2.6 & 0.45 \\ 1.0 & 1.55 \end{bmatrix} \tag{7.10}$$

Fundamental service rate for (7.10) is 2.882

Correlation $= 0.41 \times 10^{-3}$

$$S_{20} = \begin{bmatrix} -6.6 & 4.05 \\ 1.55 & -4.6 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 1.55 & 1.0 \\ 1.0 & 2.05 \end{bmatrix} \tag{7.11}$$

Fundamental service rate for (7.11) is 2.882

Correlation $= -0.41 \times 10^{-3}$

TABLE 1

$\beta = 15.0$	θ	N_{orbit}	b_0	b_1	b_2	θ_3^*	$\mathcal{P}_{idle}$
I	11.0	0.0928	.5910	.2857	.1232	0.0493	.3865
	13.0	0.0924	.5921	.2854	.1225	0.0462	.3864
	15.0	0.0921	.5930	.2851	.1219	0.0434	.3863
	17.0	0.0919	.5938	.2848	.1214	0.0410	.3863
	19.0	0.0916	.5945	.2846	.1209	0.0388	.3862
II	11.0	0.0924	.5908	.2858	.1234	0.0496	.3845
	13.0	0.0921	.5919	.2854	.1227	0.0465	.3844
	15.0	0.0918	.5928	.2851	.1221	0.0438	.3843
	17.0	0.0915	.5936	.2848	.1216	0.0413	.3843
	19.0	0.0913	.5943	.2846	.1211	0.0391	.3842

TABLE 2

$\theta = 15.0$	β	N_{orbit}	b_0	b_1	b_2	θ_3^*	$\mathcal{P}_{idle}$
I	11.0	0.1240	.5974	.2819	.1207	0.0484	.3815
	13.0	0.1057	.5949	.2837	.1214	0.0458	.3843
	15.0	0.0921	.5930	.2851	.1219	0.0434	.3863
	17.0	0.0816	.5915	.2862	.1223	0.0413	.3880
	19.0	0.0733	.5904	.2871	.1225	0.0393	.3892
II	11.0	0.1235	.5973	.2820	.1207	0.0488	.3796
	13.0	0.1053	.5947	.2837	.1216	0.0462	.3823
	15.0	0.0918	.5928	.2851	.1221	0.0438	.3843
	17.0	0.0813	.5913	.2862	.1224	0.0416	.3859
	19.0	0.0730	.5901	.2872	.1227	0.0396	.3872

Parameters: arrival $\longrightarrow$ (7.6), service to components $\longrightarrow$ (7.8),

service to externals $\longrightarrow$ (7.10) (I)

Parameters: arrival $\longrightarrow$ (7.7), service to components $\longrightarrow$ (7.8),

service to externals $\longrightarrow$ (7.10) (II)

Table 1 and 2 shows that when the retrial rate θ increases the probability that the server is idle decreases, but when the self generation rate $beta$ increases the server idle probability also increases. They also shows the effect of a small variation in the correlation between two arrival times on the system performance measures.

TABLE 3

$\beta = 15.0$	θ	N_{orbit}	b_0	b_1	b_2	θ_3^*	$\mathcal{P}_{idle}$
III	11.0	0.0928	.5907	.2863	.1230	0.0492	0.3863
	13.0	0.0925	.5917	.2859	.1223	0.0461	0.3862
	15.0	0.0922	.5927	.2856	.1217	0.0434	0.3861
	17.0	0.0919	.5934	.2854	.1212	0.0410	0.3860
	19.0	0.0917	.5942	.2851	.1207	0.0388	0.3860
IV	11.0	0.0928	.5908	.2859	.1233	0.0493	.3860
	13.0	0.0925	.5918	.2856	.1226	0.0462	.3859
	15.0	0.0922	.5927	.2853	.1220	0.0434	.3858
	17.0	0.0920	.5935	.2850	.1215	0.0410	.3857
	19.0	0.0917	.5943	.2848	.1210	0.0388	.3856

TABLE 4

$\theta = 15.0$	β	N_{orbit}	b_0	b_1	b_2	θ_3^*	$\mathcal{P}_{idle}$
III	11.0	0.1240	.5971	.2825	.1204	0.0484	.3813
	13.0	0.1058	.5946	.2842	.1212	0.0458	.3840
	15.0	0.0922	.5927	.2856	.1217	0.0434	.3861
	17.0	0.0817	.5912	.2868	.1220	0.0412	.3877
	19.0	0.0733	.5900	.2877	.1223	0.0392	.3890
IV	11.0	0.1241	.5972	.2821	.1207	0.0484	.3810
	13.0	0.1058	.5946	.2839	.1215	0.0458	.3837
	15.0	0.0922	.5927	.2853	.1220	0.0434	.3858
	17.0	0.0817	.5913	.2864	.1223	0.0413	.3874
	19.0	0.0733	.5901	.2874	.1225	0.0393	.3887

Parameters: arrival $\longrightarrow$ (7.6), service to components $\longrightarrow$ (7.8),

service to externals $\longrightarrow$ (7.11) (III)

Parameters: arrival $\longrightarrow$ (7.6), service to components $\longrightarrow$ (7.9),

service to externals $\longrightarrow$ (7.11) (IV)

TABLE 5

$\beta = 15.0$	θ	N_{orbit}	b_0	b_1	b_2	θ_3^*	$\mathcal{P}_{idle}$
V	11.0	0.0928	.5910	.2857	.1232	0.0493	.3865
	13.0	0.0924	.5921	.2854	.1225	0.0462	.3864
	15.0	0.0921	.5930	.2851	.1219	0.0434	.3863
	17.0	0.0919	.5938	.2848	.1214	0.0410	.3863
	19.0	0.0916	.5945	.2846	.1209	0.0388	.3862
VI	11.0	0.0928	.5907	.2863	.1230	0.0492	0.3863
	13.0	0.0925	.5917	.2859	.1223	0.0461	0.3862
	15.0	0.0922	.5927	.2856	.1217	0.0434	0.3861
	17.0	0.0919	.5934	.2854	.1212	0.0410	0.3860
	19.0	0.0917	.5942	.2851	.1207	0.0388	0.3860

TABLE 6

$\theta = 15.0$	β	N_{orbit}	b_0	b_1	b_2	θ_3^*	$\mathcal{P}_{idle}$
V	11.0	0.1240	.5974	.2819	.1207	0.0484	.3815
	13.0	0.1057	.5949	.2837	.1214	0.0458	.3843
	15.0	0.0921	.5930	.2851	.1219	0.0434	.3863
	17.0	0.0816	.5915	.2862	.1223	0.0413	.3880
	19.0	0.0733	.5904	.2871	.1225	0.0393	.3892
VI	11.0	0.1240	.5971	.2825	.1204	0.0484	.3813
	13.0	0.1058	.5946	.2842	.1212	0.0458	.3840
	15.0	0.0922	.5927	.2856	.1217	0.0434	.3861
	17.0	0.0817	.5912	.2868	.1220	0.0412	.3877
	19.0	0.0733	.5900	.2877	.1223	0.0392	.3890

Parameters: arrival $\longrightarrow$ (7.6), service to components $\longrightarrow$ (7.8),

service to externals $\longrightarrow$ (7.10) (V)

Parameters: arrival $\longrightarrow$ (7.6), service to components $\longrightarrow$ (7.8),

service to externals $\longrightarrow$ (7.11) (VI)

Table 3 and 4; 5 and 6 shows the effect of a small variation in the correlation between two service times in the system performance measures.

Bibliography

[1] J. R. Artalejo. Accessible bibliography of retrial queues. *Mathematical and computer modelling*, 30(3–4):1–6, August 1999.

[2] J. R. Artalejo. A classified bibliography of research on retrial queues: Progress in 1990–1999. *TOP*, 7(2):187–211, December 1999.

[3] J. R. Artalejo and M. J. Lopez-Herrero. The $M|G|1$ retrial queue : An information theoretic approach. In B. D. Choi, editor, *Proceedings of Fifth International Workshop on Retrial Queues*, pages 1–16, September 2004.

[4] F. Baccelli, P. Boyer, and G. Hebuterne. Single-server queues with impatient customers. *Adv. in Appl. Probab.*, 16:887–905, 1984.

[5] F. Baccelli and G. Hebuterne. On queues with impatient customers. In Kylstra F., editor, *Performance'81*, pages 159–179. North-Holland: Amsterdam, 1981.

[6] D. Y. Barrer. Queuing with impatient customers and indifferent clerks. *Operations Research*, 5:644–649, 1957.

[7] D. Y. Barrer. Queuing with impatient customers and ordered service. *Operations Research*, pages 650–656, 1957.

[8] R. Bellman. *Introduction to Matrix Analysis*. McGraw Hill Book Co., New York, 1960.

[9] P. P. Bocharov, R. Manzo, and A. V. Pechinkin. Tandem queues with a Markov flow and blocking. In Khalid Al-Begain, editor, *Proceedings of the ASMTA 2004*, Magdburg, Germany, 2004.

[10] P. P. Bocharov, O. I. Pavlova, and D. A. Puzikova. M/G/1/r Retrial queuing systems with priority of primary customers. *TOP*, 30(3–4):89–98, 1999.

[11] M. Brahimi and D. J. Worthington. Queueing models for out-patient appointment systems—a case study. *J. Oper. Res. Soc.*, 42:733–746, 1991.

[12] L. Breuer, A. N. Dudin, and V. I. Klimenok. A retrial BMAP|PH|N system. *Queueing Systems*, 40:433–457, 2002.

[13] L. Bright and P. G. Taylor. Equilibrium distribution for level-dependent quasi-birth-and-death processes. *Comm. Stat. Stochastic Models*, 11:497–525, 1995.

[14] S. R. Chakravarthy. The batch Markovian arrival process: a review and future work. In A. Krishnamoorthy et al., editor, *Advances in Probability Theory and Stochastic Processes*, pages 21–49. Notable Publications, NJ, 2001.

[15] S. R. Chakravarthy, A. Krishnamoorthy, and P.V. Ushakumari. A k-out-of-n reliability system with an unreliable server and phase type repairs and services: The (N, T) policy. *Appl. Math & Stoch. Anal.*, 14(4):361–380, 2002.

[16] B. D. Choi and Y. Chang. MAP_1, MAP_2/M/c retrial queue with the retrial group of finite capacity and geometric loss. *Mathematical and Computer Modelling*, 30(3–4), 1999.

[17] B. D. Choi and Y. Chang. Single server retrial queues with priority calls. *Mathematical and Computer Modelling*, 30(3–4), 1999.

[18] E. Cinlar. *Introduction to stochastic processes*. New Jersey: Prentice-Hall, 1975.

[19] D. J. Daley. General customer impatience in the queue $GI/G/1$. *Journal of Applied Probability*, 2:186–205, 1965.

[20] T. G. Deepak, V. C. Joshua, and A. Krishnamoorthy. Queues with postponed work. *TOP*, 12(3–4), 2004.

[21] A. N. Dudin and V. I. Klimenok. A retrial BMAP|SM|1 system with linear repeated requests. *Queueing Systems*, 34:47–66, 2000.

[22] A. N. Dudin, V. I. Klimenok, and G. V. Tsarenkov. Characteristics calculation for a single server queueing system with the batch Markovian arrival process, semi-Markovian service and finite buffer. *Automation and Remote Control*, 8:87–101, 2002.

[23] G. Falin and Templeton. *Retrial Queues*. Chapman & Hall, 1997.

[24] P. D. Finch. Deterministic customer impatience in the queueing system $GI/M/1$. *Biometrika*, 47:45–52, 1960.

[25] H. R. Gail, S. L. Hantler, and B. A. Taylor. Analysis of a non-preemptive priority multiserver queue. *Adv. in Appl. Probab.*, 20:852–879, 1988.

[26] D. P. Gaver, P. A. Jacobs, and G. Latouche. Finite birth-and-death models in randomly changing environments. *Adv. in Appl. Probab.*, 16:715–731, 1984.

[27] B. V. Gnedenko and I. N. Kovalenko. *Introduction to Queuing Theory*. Birkhauser Boston Inc., Boston, 2nd edition, 1989.

[28] A. Gómez-Corral. Analysis of a single-server retrial queue with quasi-random input and nonpreemptive priority. *Comput. Math. Appl.*, 43:767–782, 2002.

[29] A. Gómez-Corral. A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. *Annals of Operations Research*, 141, 2006. To appear.

[30] S. C. Graves. The application of queueing theory to continuous perishable inventory systems. *Management Sci.*, 28:400–406, 1982.

[31] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. John Wiley and Sons (Asia), third edition, 2002.

[32] P. Hokstad. A single server queue with constant service time and restricted accessibility. *Management Science*, 25:205–208, 1979.

[33] N. K. Jaiswal. *Priority Queues*. Academic Press, New York, 1968.

[34] V. I. Klimenok. Sufficient condition for existence of 3-dimensional quasi-toeplitz markov chain stationary distribution. *Queues: Flows, Systems, Networks*, 13:142–145, 1997.

[35] A. G. De Kok and H. C. Tijms. A queueing system with impatient customers. *Journal of Applied Probability*, 22:688–696, 1985.

[36] A. Krishnamoorthy, Viswanath C. Narayanan, and T. G. Deepak. On a queueing system with self generation of priorities. *Neural Parallel & Scientific Computing*, 13, 2005.

[37] A. Krishnamoorthy and P. V. Ushakumari. Reliability of k-out-of-n system with repair and retrial of failed units. *TOP*, 7(2):293–304, 1999.

[38] A. Krishnamoorthy, P. V. Ushakumari, and B. Lakshmi. k-out-of-n system with the repair: the N-policy. *Asia Pacific Journal of Operations Research*, 19:47–61, 2002.

[39] C. Langaris. Waiting time analysis of a two-stage queueing system with priorities. *Queueing Syst.*, 14:457–473, 1993.

[40] A. N. Langville and W. J. Stewart. The kronecker product and stochastic automata networks. *J. Comput. Appl. Math.*, 167:429–447, 2004.

[41] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. ASA-SIAM: Philadelphia, 1999.

[42] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Commun. Statist.-Stochastic Models*, 7:1–46, 1991.

[43] S. Nahmias. Perishable inventory theory:a review. *Operations Research*, 30:680–708, 1982.

[44] M. F. Neuts. *Matrix-geometric solutions in stochastic models-An algorithmic approch*. John-Hopkins Univ. Press, 1981. Also published by Dover, 2002.

[45] M. F. Neuts and B. M. Rao. Numerical investigation of a multiserver retrial model. *Queueing systems*, 7:169–190, 1990.

[46] R. Nobel. A discrete-time retrial queueing model with one server. In B. D. Choi, editor, *Proceedings of Fifth International Workshop on Retrial Queues*, pages 39–51, September 2004.

[47] D. Perry. Analysis of a sampling control scheme for a perishable inventory system. *Oper. Res.*, 47:966–973, 1999.

[48] M. W. Sasieni. Double queues and impatient customers with an application to inventory theory. *Operations Research*, 9:771–781, 1961.

[49] D. A. Stanford. Waiting and interdeparture times in priority queues with poisson- and general-arrival streams. *Oper. Res.*, 45:725–735, 1997.

[50] R. E. Stanford. Reneging phenomena in single channel queues. *Mathematics of Operations Research*, 4:162–178, 1979.

[51] R. E. Stanford. On queues with impatience. *Advances in Applied Probability*, 22:768–769, 1990.

[52] A. R. Swensen. On a GI/M/c queue with bounded waiting times. *Operations Research*, 34:895–908, 1986.

[53] L. Takács. Priority queues. *Operations Research*, 12:63–74, 1964.

[54] L. Takács. A single-server queue with limited virtual waiting time. *Journal of Applied Probability*, 11:612–617, 1974.

[55] H. Takagi. *Queueing Analysis–Volume 1: Vacations and Priority Systems*. North-Holland, Amsterdam, 1991.

[56] I. D. S. Taylor and J. G. C. Templeton. Waiting time in a multi-server cutoff-priority queue and its application to an urban ambulance service. *Oper. Res.*, 28:1168–1188, 1980.

[57] R. L. Tweedie. Sufficient conditions for regularity, recurrence and ergodicity of markov processes. *Proc. Camb. Phil. Soc.*, 78, Part I, 1975.

[58] P. V. Ushakumari and A. Krishnamoorthy. Reliability of a k-out-of-n system with general repair, the N-policy. In Janson and Limnios, editors, *Proceedings of IInd International Conference on Semi-Markov Processes and Application*, Cedex, France, Gordon and Breach, 1998.

[59] P. V. Ushakumari and A. Krishnamoorthy. k-out-of-n system with repair: the $\max(N, T)$ policy. *Performance Evaluation*, 57:221–234, 2004.

[60] Q. Wang. Modeling and analysis of high risk patient queues. *European J. Oper. Res.*, 155:502–515, 2004.

[61] Y. Q. Zhao and A. S. Alfa. Performance analysis of a telephone system with both patient and impatient customers. *Telecommunication Systems*, 4:201–215, 1995.

[62] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.*, 48:566–583, 2002.

UNIVERSITY LIBRARY
Cochin - 682 022
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY