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

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 $A = A_0 + A_1 + A_2$ 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*

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

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

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 $T\mathbf{e} + T^0 = 0$. The initial probability vector of Q is given by $(\boldsymbol{\alpha}, \alpha_{m+1})$ where α_{m+1} is a scalar and $\boldsymbol{\alpha}\mathbf{e} + \alpha_{m+1} = 1$. 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 $(\boldsymbol{\alpha}, T)$ 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 $(\boldsymbol{\alpha}, T)$ of order m then

(1) The distribution function of X is given by

$$F(x) = P(X \leq x) = 1 - \boldsymbol{\alpha} \exp(Tx)\mathbf{e}.$$

(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

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

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

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

where D_0 is an $m \times m$ matrix, $D_0\mathrm{e} + \mathrm{d} = 0$ 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 $(\boldsymbol{\alpha}_i, 0)$, where $\boldsymbol{\alpha}_i$ is an m-dimensional row vector with $\boldsymbol{\alpha}_i\mathrm{e} = 1$, 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 $(D_1)_{i,j} = (\mathrm{d})_i(\boldsymbol{\alpha}_i)_j$ $1 \leq i, j \leq m$. 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

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

her D_0 is an $m \times m$ matrix, $D_0\mathbf{e} + \mathbf{d} = 0$ 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 $(\boldsymbol{\alpha}_i, 0)$, where $\boldsymbol{\alpha}_i$ is an m-dimensional row vector and $\boldsymbol{\alpha}_i\mathbf{e} = 1$, 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 indcpendent 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