Finite-Source Queueing Systems and their Applications

János Sztrik

September 24, 2004

Contents

Part I

Finite-Source Queueing Systems

Chapter 1

Introduction

1.1 Queueing systems

A single station queueing system consists of a queueing buffer of finite or infinite size and one or more identical servers. Such an elementary queueing system is also referred to as a service station or, simply, as a node. First we start with a short description of queueing systems, see for example, [7, 10, 18, 23, 46].

A server can only serve one customer at a time and hence, it is either in "busy" or "idle" state. If all servers are busy upon the arrival of a customer, the newly arriving customer is buffered, assuming that buffer space is available, and waits for its turn. When the customer currently in service departs, one of the waiting customers is selected for service according to a *queueing (or scheduling) discipline*. An elementary queueing system is further described by an arrival process, which can be characterized by its sequence of interarrival time random variables $\{A_1, A_2, \cdots\}$. It is common to assume that the sequence of interarrival times is independent and identically distributed, leading to an arrival process that is known as a renewal process. The distribution function of interarrival times can be continuous or discrete.

The average interarrival time is denoted by $E[A] = \overline{T}_A$ and its reciprocal by the average arrival rate λ:

$$\lambda = \frac{1}{\overline{T}_A}. \tag{1.1}$$

The most common interarrival time distribution is the exponential, in which case the arrival process is Poisson. The sequence $\{B_1, B_2, \cdots\}$ of service times of successive jobs also needs to be specified. We assume that this sequence is also a set of independent random variables with a common distribution function.

The mean service time $E[B]$ is denoted by b and its reciprocal by the service rate μ:

$$\mu = \frac{1}{b}\,. \tag{1.2}$$

However, there are many practical situations when the request's arrivals do not form a renewal process, that is the arrivals may depend on the number of customers, request, jobs staying at the service facility. This happens in the case of **finite-source queueing systems** .

Let us consider some specific examples following in the order of their appearance in practice.

Example 1.1.1. *Consider a set of N machines that operate independently of each other. After a random time they may break down and need repair by one or multiple operatives (repairmen) for a random time. The repair is carried out by a specific discipline and after having been fixed each machine renew its operation. It is assumed that the server can handle only one machine at a time. Besides the usual main characteristics in reliability theory we would like to know the* **distribution of the failure-free operation time** *of the whole system, that is distribution of time while the number of stopped machines exceeds a given limit supposing certain initial conditions, usually, that all the machines are operating.*

Example 1.1.2. *Suppose a single unloader system at which trains arrive which bring coal from various mines. There are N trains involved in the coal transport. The coal unloader can handle only one train at a time and the unloading time per train is a random variable. The unloader is subject to random breakdowns when it is in operation. The operating time and the time to repair a broken unloader are also random variables. The unloading of the train is resumed as soon as the repair of the unloader is completed. An unloaded train returns to the mines for another load of coal. The time for a train to complete a trip from the unloader to the mines and back is assumed to be a random variable, too.*

Example 1.1.3. *N terminals request to use of a computer (server) to process transactions. The length of time that the terminal takes to generate a request for the computer is called "thinking" time. The length of time from the instant a terminal generates a transaction until the computer completes the transaction (and instantaneously responds by communicating this fact to the user at the terminal) is called "response time". We would like to know, for example, the rate at which transactions are processed (which in steady-state equals the rate at which they are generated) is called "throughput", which is one the most important performance measures showing the system's processing power.*

Example 1.1.4. *Let us consider a memory system where N disk units share a disk controller (server) and transmit information when they find the controller idle. Unsatisfied requests are repeated after a disk's rotation which can be modelled as a constant repetition interval.*

Example 1.1.5. *In trunk mobile systems, telephone lines are interfaced with the radio system at the repeaters which serve dispatch type mobile subscribers and telephone line users. Let us consider a system which serves two different types of communication traffic (i) dispatch traffic has short average service time and (ii) interconnect traffic of telephone line users. Both types of users are assumed to arrive from a finite population. The dispatch users are allowed to access all repeaters while interconnect users can occupy only a fixed number of repeaters. A service sharing algorithm to derive blocking probabilities of dispatch and interconnect users and average dispatch delay is to be find.*

Example 1.1.6. *Let us examine the dynamic behavior of a local area network based on the non-persistent Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol. In such a network a finite number, say* N*, of users (or active terminals) are connected by a single channel (bus). Under the specific protocol, if a terminal has a message ready for transmission, the terminal immediately senses the channel to see whether it is idle or busy. If the channel is sensed busy, it re-senses the channel after a random amount of time. On the other hand, if the channel is sensed idle, it starts transmitting the message immediately. Due to non-zero propagation delay, within a certain time interval after the terminal starts transmitting the message, other terminals (if any) with messages ready for transmission may also sense the channel idle and transmit their messages. This phenomenon is referred to as a collision. Each terminal involved in a collision abandons its transmission and re-sense the channel at a later time as if it had sensed a busy channel. A collision usually lasts for a certain amount of time during which no terminals are allowed to transmit, that is, a recovery time is needed by the channel to be free again. This kind of system can be modeled as retrial queueing system with server's vacation.*

As we could see all the above mentioned examples have a common characteristic: We have a queueing system in which requests for service are generated by a finite number N of identical or heterogeneous sources and the requests are handled by a single or multiple server(s). The service times of the requests generated by the sources are random variables. It is assumed that the server can handle only one request at a time and uses specified service discipline. New requests for service can be generated only by idle sources, which are sources having no previous request waiting or being served at the server. A source idle at the present time will generate a request independently of the states of the other sources after a random time with given distribution.

It is easy to see, that in homogeneous case this system can be considered as a closed queueing network with two nodes one with an infinite server (source) and another one with a single or multiple servers (service facility). Similarly, in heterogeneous case it can be viewed as a closed network consisting of $N + 1$ nodes where each request has it own node where to it returns after having been serviced at the " central " node representing the service facility.

Depending on the assumptions on source, service times of the requests and the service disciplines applied at the service facility, there is a great number of queueing models at different level to get the main steady-state performance measures of the system. It is also easy to see, that depending on the application we can use the terms *request, customer, machine, message, job* equivalently. The above mentioned models (problems) are referred to as *machine repair, machine repairmen, machine interference, machine service, unloader problem, terminal model, quasi-random input processes, finite-source or population models* , respectively.

1.1.1 Kendall's notation

The following notation, known as Kendall's notation, is widely used to describe elementary queueing systems:

A/B/m/K/N - queueing discipline,

where A indicates the distribution of the interarrival times, B denotes the distribution of the service times, m is the number of servers, K is the capacity of the system, that is the maximum number of customers staying at the facility (sometimes in the queue), and N denotes the number of sources.

The following symbols are normally used for A and B:

M	Exponential distribution (Markovian or memoryless property)
E_k	Erlang distribution with k phases
H_k	Hyperexponential distribution with k phases
C_k	Cox distribution with k phases
D	Deterministic distribution, i.e., the interarrival time or service time is constant
G	General distribution

The *queueing discipline or service strategy* determines which job is selected from the queue for processing when a server becomes available.

As an example of Kendall's notation, the expression

M/G/1 - LCFS preemptive resume (PR)

describes an elementary queueing system with exponentially distributed interarrival times, arbitrarily distributed service times, and a single server. The queueing discipline is LCFS where a newly arriving job interrupts the job currently being processed and replaces it in the server. The servicing of the job that was interrupted is resumed only after all jobs that arrived after it have completed service.

M/G/1/K/N

describes a finite-source queueing system with exponentially distributed source times, arbitrarily distributed service times, and a single server. There are N request in the system and they are accepted for service iff the number of requests staying at the server is less than K. The rejected customers return to the source and start a new source time with the same distribution. It should be noted that as a special case of this situation the $M/G/1/N/N$ system could be considered. However, in this case we use the traditional $M/G/1//N$ notation, that is the missing letter, as usual in this framework, means infinite capacity.

It is natural to extend this notation to heterogeneous requests, too. The case when we have different requests is denoted by $\rightarrow$. So, the

$$\vec{M}/\vec{G}/1/K/N$$

denotes the above system with different arrival rates and service times.

1.2 Performance measures for finite-source systems

1.2.1 Homogeneous systems

For the better understanding let us consider an $M/G/1//N$ system without server vacations treated in details in [42]. One of the performance measures in our system is the *mean message response time* $E[T]$ defined as the mean time from the arrival of a new message to its service completion, that is, the mean time a message spends in the service facility. Since the mean time that each message takes to complete cycle of staying in the source and staying in the service facility is $E[T] + 1/\lambda$, the *throughput* γ of the system, which is defined as the mean number of messages served per unit time in the whole system, is given by $N/(E[T] + 1/\lambda)$. If P_0 denotes the probability that the server is idle at an arbitrary time, then $\rho' = 1 - P_0$ is the *carried load* or *server utilization*, namely, the long run fraction of the time that the server is busy. Thus, the throughput is also given by $(1 - P_0)/b$. By equating these two expressions for the throughput, we get

$$\gamma = \frac{N}{E[T] + 1/\lambda} = \frac{1 - P_0}{b} = \frac{\rho'}{b} \tag{1.3}$$

Hence we have

$$E[T] = \frac{Nb}{1 - P_0} - \frac{1}{\lambda} \tag{1.4}$$

If $E[L]$ denotes the mean number of messages in the service facility at an arbitrary time, we also have the relationship

$$\gamma = \lambda(N - E[L]) \tag{1.5}$$

that equates the throughput to the mean number of messages arriving per unit of time. Thus we get

$$E[L] = N - \frac{1 - P_0}{\lambda b} = \gamma E[T] \tag{1.6}$$

which is an example of **Little's theorem** applied to those messages that are accepted by the service facility. The ratio

$$E = \frac{N - E[L]}{N} = \frac{\gamma}{N\lambda} = \frac{1 - P_0}{N\lambda b} \tag{1.7}$$

is called the *machine availability* in machine interference models, since it represents the expected fraction of time that a machine remains in working condition, E is the *machine efficiency*, because it is the ratio of the total actual production to what would have been achieved had no stoppage taken place. From (1.3) through (1.5, 1.6), it is clear that performance measures such as ρ' , γ, $E[T]$, and $E[L]$ can be obtained once we have evaulated P_0.

Let $E[\Theta]$ be the mean length of a busy period. Since the state of the system repeats regenerative cycles of a busy period of mean length $E[\Theta]$ and an idle period of mean length $E[I] = 1/(N\lambda)$, the probability P_0 that the server is idle at an arbitrary time is given by

$$P_0 = \frac{E[I]}{E[\Theta] + E[I]} = \frac{1/(N\lambda)}{E[\Theta] + 1/(N\lambda)} \tag{1.8}$$

If π_0 denotes the probability that the service facility is empty after a service completion, $1/\pi_0$ is the mean number of messages that are served during each busy period. This can be seen by considering a long period of time during which a large number of (say **N**) messages are served. Such a period will include $\mathbf{N}\pi_0$ busy periods on the average, because π_0 is the probability that a busy period is terminated after a service completion. Therefore, on the average $1/\pi_0$ messages are served per busy period. Hence, the mean length of a busy period is given by

$$E[\Theta] = \frac{b}{\pi_0} \tag{1.9}$$

From (1.8) and (1.9), we get

$$P_0 = \frac{\pi_0}{\pi_0 + N\lambda b} \tag{1.10}$$

Substituting (1.10) into (1.3),(1.4), and (1.6) we can express the throughput γ, the mean message response time $E[T]$, and the mean number $E[L]$ of messages in the service facility at an arbitrary time in terms of π_0, too, as

$$\gamma = \frac{N\lambda}{\pi_0 + N\lambda b} \qquad ; \qquad E = \frac{1}{\pi_0 + N\lambda b} \tag{1.11}$$

$$E[T] = Nb - \frac{1-\pi_0}{\lambda} \tag{1.12}$$

$$E[L] = N\left(1 - \frac{1}{\pi_0 + N\lambda b}\right) \tag{1.13}$$

We can find π_0 by analyzing a Markov chain of the queue size imbedded at service completion times, or the method of supplementary variables can be applied to obtain P_0.

1.2.2 Asymptotic properties

We can discuss some asymptotic properties of these performance measures without recourse to detailed analysis of the system state. When N is fixed, for $\lambda \approx 0$ we have almost no congestion at the service facility, which means that $\pi_0 \approx 1, P_0 \approx 1, \gamma \approx N\lambda, E \approx 1, E[T] \approx b$ and $E[L] \approx N\lambda b$. As $\lambda \to \infty$, every message whose service has just been completed returns to the facility almost immediately.
Therefore

$$\pi_0 \to 0, \quad P_0 \to 0, \quad \gamma \to 1/b,$$
$$E \to 0, \quad E[T] \to Nb, \quad E[L] \to N.$$

We note that $E[T]$ in (1.4) or (1.12) as a function of N has simple asymptotic forms. When $N = 1$ (which is equivalent to a *loss system* M/G/l/l), we obviously have $\pi_0 = 1$ and $E[T] = b$. As $N \to \infty$, we have $\pi_0 \to 0$ and so

$$E[T] \approx Nb - \frac{1}{\lambda} \qquad \text{as} \qquad N \to \infty \tag{1.14}$$

The value of N, denoted by N^*, at which the two straight lines $E[T] = b$ and the one in (1.14) as a function of N intersect each other is called the *saturation number*. It is given by

$$N^* = 1 + \frac{1}{\lambda b} \tag{1.15}$$

Note that this can be written as $N^* = (b + 1/\lambda)/b$. Therefore, if nature were kind and all messages required exactly b service time and exactly $1/\lambda$ generation time (a deterministic system), then N^* would be the maximum number of messages that could be scheduled without causing mutual interference, see [24] page 209.

It should also be mentioned that well-known book of Takagi [42] provides an organized and unified presentation of the analysis techniques for $M/G/1//N$ systems without and with vacations. The $M/G/1/K/N$ and $M/G/m/m/N$ models are also treated there. In Takagi [43] discrete-time $Geo/G/1/K/N$ systems are analyzed. To the best knowledge of the author these books are the most comprehensive ones on this special topic in the existing literature.

1.2.3 Heterogeneous systems

In this section, we study $M/G/1//N$ systems with a heterogeneous population; that is, we assume that messages can be distinguished according to their arrival rates and service time distributions. We consider three models that differ with respect to the population constraint: an individual message model, a multiple finite-source model, and a single finite-source model. In the *individual message model*, each message has a distinct arrival rate and a distinct service time distribution. It is also called a *single buffer model* because of its equivalence to a system of multiple classes of messages in which each class is allotted a single buffer. In the *multiple finite-source model*, see [22] (sec. III.1), there are P classes of messages and the population size of class p is fixed at $N_p(< \infty)$ such that $N = \sum_{p=1}^{P} N_p$. The individual message model is a special case of the multiple finite-source model in which $P = N$ and $N_p = 1$ for $i \leq p \leq N$. In the *single finite source model* the total number of messages in the system is fixed at N, and each message becomes a message of one of P classes with a given probability when it leaves the source.

The multiple finite-source model and the single finite-source model may be associated with *flow control* and *congestion avoidance* mechanisms in computer communication networks. Namely, the multiple finite-source model in which the population size is fixed for each class corresponds to the *window flow control*. Let us first assume that each of the N messages has different characteristics. In terms of machine interference problems, each machine is assumed to have a different breakdown rate and a different repair time distribution. Specifically, let λ_i be the rate at which message i in the source arrives at the service facility, and let $B_i(x)$ be the distribution function (DF) for the service time of message i, where $i = 1, 2, \ldots, N$. We also denote by b_i and $B_i^*(s)$ the mean and Laplace-Stieltjes transform (LST) of $B_i(x)$, respectively. We call this system an *individual message model*. The total arrival rate when all messages are in the source is denoted by

$$\Lambda = \sum_{i=1}^{N} \lambda_i \tag{1.16}$$

We denote by $E[T_i]$ the mean response time of message i, and by γ_i the throughput of message i, that is, the mean number of times that message i is served per unit time, where $i = 1, 2, \ldots, N$. These are related by

$$\gamma_i = \frac{1}{E[T_i] + 1/\lambda_i} \qquad 1 \leq i \leq N \tag{1.17}$$

If Γ_i denotes the mean number of times that message i is served in a busy period of length Θ, the throughput γ_i can also be expressed as

$$\gamma_i = \frac{\Gamma_i}{E[\Theta] + E[I]} \qquad 1 \leq i \leq N \tag{1.18}$$

where

$$E[I] = \frac{1}{\Lambda} \tag{1.19}$$

is the mean lenght of an idle period I, and

$$E[\Theta] = \sum_{j=1}^{N} b_j \Gamma(j) \tag{1.20}$$

is the mean length of an busy period Θ. The carried load (total server utilization) ρ' is given by

$$\rho' = \frac{E[\Theta]}{E[\Theta] + E[I]} = 1 - P_0 \tag{1.21}$$

where P_0 is the probability that the service facility is empty at an arbitrary time. The total throughput γ of the system is given by

$$\gamma = \sum_{i=1}^{N} \gamma_i = \frac{\sum_{i=1}^{N} \Gamma_i}{E[\Theta] + E[I]} \tag{1.22}$$

Hence we can obtain the throughput γ_i and the mean response time $E[T_i]$ once we have calculated $\{\Gamma(j);\ 1 \leq j \leq N\}$, where $i = 1, 2, \ldots, N$. The mean waiting time of message i is given by

$$E[W_i] = E[T_i] - b_i = \frac{1}{\gamma_i} - \frac{1}{\lambda_i} - b_i \qquad 1 \leq i \leq N \tag{1.23}$$

If $P^{(i)}$ denotes the probability that message i is present in the service facility at an arbitrary time, we have

$$P^{(i)} = \frac{E[T_i]}{E[T_i] + 1/\lambda_i} = \gamma_i E[T_i] = 1 - \frac{\gamma_i}{\lambda_i} \qquad 1 \leq i \leq N \tag{1.24}$$

which represents **Little's theorem** for message i in the service facility. In terms of machine repairman problems, $P^{(i)}$ is the probability that machine i is down at an arbitrary time. Alternatively, we can express the mean response time $E[T_i]$ and the throughput γ_i for message i in terms of $P^{(i)}$ as

$$E[T_i] = \frac{P^{(i)}}{\lambda_i(1 - P^{(i)})} \qquad ; \qquad \gamma_i = \lambda_i(1 - P^{(i)}) \tag{1.25}$$

Depending on the assumptions on source, service times of the requests and the service disciplines applied at the service facility, there is a great number of queueing models at different level to get the main steady-state performance measures of the system. It is also easy to see, that depending on the application we can use the terms *request, customer, machine, message, job* equivalently. The above mentioned models (problems) are referred to as *machine repair, machine repairmen, machine interference, unloader problem, terminal model, or quasi-random input processes, finite population models* , respectively.

Because of the page limitation, only the most related references are cited. However, a more detailed Bibliography can be found on this topic in [31].

For additional materials the interested reader is referred to the following basic comprehensive books [1, 7, 10, 11, 12, 17, 18, 22, 23, 24, 25, 26, 28, 40, 41, 42, 43, 44, 47].

The main aim of the following chapters is to show how different methods can be applied in the investigation of finite-source queueing systems. Thus, analytical, numerical and asymptotic approaches are presented.

The classical $M/M/r//N$ model is treated in full details because in this case the waiting and response time distribution functions can explicitly be derived. Then by using the supplementary variable technique closed-form steady-state distributions can be obtained for systems with heterogeneous requests. After that a stable numerical algorithmic approach is introduced which works even in those cases when the calculation of the famous Takács-formulas is failed due to the factorial.

Recent tool-supported modeling techniques are illustrated by using the software package MOSEL for retrial systems with non-reliable servers.

Finally, asymptotic analysis for complex renewable systems with fast repair evolving in random environment is presented. This approach is very effective since state space explosion problems can be avoided by exploring the special structure of the underlying Markov chain.

In many cases the proofs are omitted. However, some theorems are stated because either they follow from well-known theorems or they are important for possible future investigations. In addition, the most important sources of information are listed to draw attention of the interested readers. Finally, some of the recent works of the author are either presented or cited.

Acknowledgement. The author is very grateful to Prof. M. Telek for his valuable comments.

Chapter 2

Analytical Methods

2.1 Homogeneous $M/M/r//N$ systems, the classical model

This section presents the classical queuing theory approach for solving to machine interference problem. It should be noted that this system is analyzed by many authors in different books. It is a classical example for queueing systems with state-dependent arrival rates and it can be treated in the framework of the so-called *birth-and-death processes*. The present problem is descibed in several classical books on queueing systems, for example [1, 9, 10, 23, 18, 20, 46] such to mention the basic ones. Our aim is to show the form of the steady-state probabilities of stopped machines. In the above mentioned works one can find the detailed analysis of waiting time, down time distribution of machines, too. Several numerical examples from real life situations illustrates this interesting system.

It is also proved that in steady-state the arriving machines' distribution in system containing N machines is the same as the outside observer's distribution for the corresponding system with $N-1$ machines, or other words in arrival epochs the distribution is the same as the time-average distribution of system with one machine less .

The assumptions of the model are as follows:

Suppose that there are N machines and r operators, $(r < N)$, and

1. The time between breakdowns (or production time) of any one of the machines is a sample from an *exponential* probability distribution with mean $1/\lambda$, (or mean rate λ). A breakdown is random and is independent of the operating behavior of the other machines. Then, when there are n machines not working at time t,

   Prob (one of the $N-n$ machines goes down in the interval $(t, t+\Delta t)$) $=$ $(N-n)\lambda\Delta t + o(\Delta t)$,

   where Δt is a small increment of time.

2. Any one of the n down machines requires only one of the r operators to fix it. The service time distribution is *exponential* with mean $1/\mu$ for each machine and each operator. The service times are mutually independent and also independent of the number of down machines.

Then

Prob [one of the n down machines is fixed in an interval Δt]

$$= \begin{cases} n\mu\Delta t + o(\Delta t), & \text{for } 1 \leq n \leq r, \\ r\mu\Delta t + o(\Delta t), & \text{for } r < n \leq N. \end{cases}$$

3. The machines are *served in the order of their beakdowns*.

Let

L(t) = the number of down machines at time t

and

$$P_n(t) = Prob(L(t) = n | L(0) = i), \quad n = 0, \ldots, N.$$

Then the stochastic process, ($L(t), t \geq 0$), is a birth-and-death process, with rates

$$\lambda_n = \begin{cases} (N-n)\lambda, & n = 0, 1, \ldots, N, \\ 0, & n > N, \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & n = 1, 2, \ldots, r, \\ r\mu, & n = r+1, \ldots, N. \end{cases}$$

The forward Kolmogorov-equations of the birth-death process are

$$\begin{aligned}
P_0'(t) &= N\lambda P_0(t) + \mu P_1(t), \\
P_n'(t) &= -((N-n)\lambda + n\mu)P_n(t) + (N-n+1)\lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t), \\
&\quad 1 \leq n < r, \\
P_n'(t) &= -((N-n)\lambda + r\mu)P_n(t) + (N-n+1)\lambda P_{n-1}(t) + r\mu P_{n+1}(t), \\
&\quad r \leq n < N, \\
P_N'(t) &= -r\mu P_N(t) + \lambda P_{n-1}(t).
\end{aligned}$$

This finite system of ordinary differential equations can be solved and we can get the transient probabilities.

For the equilibrium values of P_n these derivatives are equal to zero and the equilibrium (or stationary or steady state) values are

$$P_n = \lim_{t \to \infty} P_n(t).$$

The flow balance equations (steady-state equations) become

$$\begin{aligned}
N\lambda P_0 &= \mu P_1, \\
((N-n)\lambda + n\mu)P_0 &= (N-n+1)\lambda P_{n-1} + (n+1)\mu P_{n+1}, \quad 1 < n < r, \\
((N-n)\lambda + r\mu)P_0 &= (N-n+1)\lambda P_{n-1} + r\mu P_{n+1}, \quad r \leq n < N, \\
r\mu P_N &= \lambda P_{N-1}.
\end{aligned}$$

These equations are solved recursively using the relationship

$$
\begin{aligned}
(N-n)\lambda P_n &= (n+1)\mu P_{n+1}, \quad 0 \le n < r,\\
(N-n)\lambda P_n &= r\mu P_{n+1}, \quad r \le n < N.
\end{aligned}
$$

Letting $\rho = \lambda/\mu$ (the servicing factor), the steady-state probabilities are

$$
\begin{aligned}
P_n &= \binom{N}{n}\rho^n P_0 \qquad \text{for } 0 \le n \le r, \qquad (2.1)\\
P_n &= \frac{N!}{(N-n)!r!r^{n-r}}\rho^n P_0 \qquad \text{for } r \le n \le N.
\end{aligned}
$$

where P_0 is obtained by solving $\sum\limits_{n=0}^{N} P_n = 1$ to get

$$
P_0 = \left(\sum_{n=0}^{r} \binom{N}{n}\rho^n + \sum_{n=r+1}^{N} \binom{N}{n} \frac{n!}{r!r^{n-r}}\rho^n \right)^{-1}
$$

In the following only the **main performance measures** of the machine interference problem are mentioned.

1. **The expected (average) number of down machines**

$$
E[L] = \sum_{n=0}^{N} nP_n.
$$

2. **Machine efficiency or machine utilization**

$$
U_m = \frac{N - E[L]}{N},
$$

which is the percentage of average production obtained (or the fraction of total production time on all machines).

3. **Average operator utilization**

$$
U_S = \sum_{n=0}^{N} \frac{nP_n}{r} + \sum_{n=r+1}^{N} P_n
$$

which is fraction of time an operator would be working.

4. **Average number of idle operators**

$$
r - rU_s = \sum_{n=0}^{r} (r-n)P_n.
$$

5. **Average number of machines waiting**

$$\overline{Q} = \sum_{n=r+1}^{N} (n-r)P_n.$$

6. **Average down time of machines**

$$\overline{T} = \frac{E(L)}{\lambda(N-E(L))}.$$

7. **Mean waiting time of machines**

$$\overline{W} = \frac{\overline{Q}}{\lambda(N-E(L))}.$$

By dividing measure 4 by the number of operators, r, and measure 5 by the number of machines, N, some related measures can be obtained

- **Coefficient of loss for operator**

$$\frac{\sum_{n=0}^{N}(r-n)P_n}{r}$$

or percentage of idle operators.

- **Coefficient of loss for machines**

$$\frac{\sum_{n=r+1}^{N}(n-r)P_n}{N}$$

or percentage of interference time.

In general, there is no closed form solution for these characterisrics. However, for the single server case all of these measures can be expressed as the function of U_S in the following way.

$$P_k = \frac{N!}{(N-k)!}\rho^k P_0, \qquad \rho = \frac{\lambda}{\mu}$$

$$P_0 = \frac{1}{\displaystyle\sum_{k=0}^{N}\frac{N!}{(N-k)!}\rho^k}.$$

If $z = \dfrac{\mu}{\lambda} = \rho^{-1}$, then we get the following very useful relation

$$P_0 = \frac{1}{\displaystyle\sum_{k=0}^{N}\frac{N!}{(N-k)!}\rho^k} = \frac{1}{\displaystyle\sum_{k=0}^{N}\frac{N!}{k!}\rho^{N-k}} = \frac{1}{\displaystyle N!\rho^N\sum_{k=0}^{N}\frac{1}{k!\rho^k}} = \frac{\dfrac{z^N}{N!}}{\displaystyle\sum_{k=0}^{N}\frac{z^k}{k!}} = B(N,z),$$

where $B(N,z)$ is the well-known **Erlang's B formula** , or loss formula. It can easily be seen that the following recurrence relation is valid

$$B(m,z) = \frac{zB(m-1,z)}{m+zB(m-1,z)} \qquad m = 2,3,...$$

$$B(1,z) = \frac{z}{1+z}.$$

Hence the server utilization is

$$U_S = 1 - P_0 = 1 - B(N,z).$$

After some elementary calculations for the performance measures we obtain

$$E[L] = N - \frac{U_S}{\rho},$$

$$\overline{Q} = \sum_{k=1}^{N}(k-1)P_k = \sum_{k=1}^{N} kP_k - \sum_{k=1}^{N} P_k = N - \left(1+\frac{1}{\rho}\right)U_S,$$

$$U_m = \frac{N-E[L]}{N} = \frac{U_S}{N\rho},$$

$$\overline{T} = \frac{E[L]}{\lambda(N-E[L])} = \frac{1}{\mu}\left(\frac{N}{U_S} - \frac{1}{\rho}\right),$$

$$\overline{W} = \frac{\overline{Q}}{\lambda(N-E(L))} = \frac{1}{\mu}\left(\frac{N}{U_S} - \frac{1+\rho}{\rho}\right).$$

Let us denote by $U_S[N]$ the server's utilization emphasizing that the number of machines is N. Hence, we can write

$$U_S[N] = 1 - B\left(N,\frac{1}{\rho}\right) = 1 - \frac{\frac{1}{\rho}B\left(N-1,\frac{1}{\rho}\right)}{N+\frac{1}{\rho}B\left(N-1,\frac{1}{\rho}\right)} =$$

$$= \frac{N}{N+\frac{1}{\rho}B\left(N-1,\frac{1}{\rho}\right)} = \frac{N\rho}{N\rho+B\left(N-1,\frac{1}{\rho}\right)} =$$

$$= \frac{N\rho}{N\rho+1-U_S[N-1]}, \qquad N = 2,3,....$$

with the initial value

$$U_S[1] = \frac{\rho}{1+\rho}.$$

As a result we have a very useful recurrence formula for calculating the performance measures for case of one more machines. To derive the distribution function for the waiting time of a machine, $W[\cdot]$, we reason that an arriving machine (customer) must queue for repair (service) only if $n \geq r$, where n is the number of customers found in the repair system. When this is the case, the arrival must wait for the departure of $(n-r)+1$ customers. (If $n = r$, one customer must depart, if $n = r+1$, two customers must depart, etc.) Let N_a be the number of customers an arriving machine finds in the repair facility so that $q_n = P[N_a = n]$. If $n \geq r$ and $k = n-r$, then $P[q > t|N_q = n]$ is the probability that k or fewer customers depart in an interval of length t. But this probability is given by

$$P[q > t|N_q = n] = e^{-r\mu t} \sum_{i=0}^{k} \frac{(r\mu t)^i}{i!} = Q[k; r\mu t], \tag{2.2}$$

where $Q[k; r\mu t]$ is the cumulative Poisson distribution. Formula (2.2) follows from the fact that the service facility services customers like a single server with an exponential distribution and mean service time $1/(r\mu)$. Thus, the number of customers processed in an interval of length t has a Poisson distribution with mean $r\mu t$. We have

$$P[q > t] = \sum_{n=r}^{N-1} P[q > t|N_q = n]q_n = \sum_{n=r}^{N-1} q_n Q[n-r; r\mu t]. \tag{2.3}$$

To complete the derivation, we show that any N-source birth-and-death queueing system with quasi-random input, the arriving customer's distribution in steady-state is the same as the outside observer's distribution for the corresponding $(N-1)$-source system.

We consider a system with N sources, each source originating requests at rate λ when idle and rate 0 otherwise (quasi-random input). Then the request rate when n sources are busy (in service or waiting for service) is

$$\lambda_n = (N-n)\lambda \quad (n = 0, 1, ..., N). \tag{2.4}$$

To calculate the arriving customer's distribution $\{q_n\}$, by using the Bayes-formula it is easy to see that we get

$$q_n = \frac{(N-n)P_n}{\sum\limits_k (N-k)P_k}. \tag{2.5}$$

In order to emphasize the dependence on the number N of sources, we write $P_n = P_n[N]$ and $q_n = q_n[N]$. Then (2.5) becomes

$$q_n[N] = \frac{(N-n)P_n[N]}{\sum_{k=0}^{N-1}(N-k)P_k[N]} \quad n = 0, 1, ..., N-1. \tag{2.6}$$

Thus, the outside observer's distribution can be written as

$$P_n[N] = \frac{N(N-1)\cdots(N-n+1)\lambda^n}{\mu_1\mu_2\cdots\mu_n}P_0[N] \quad n = 1, 2, ..., N \tag{2.7}$$

and

$$P_0[N] = \left(1 + \sum_{k=1}^{N}\frac{N(N-1)\cdots(N-k+1)\lambda^k}{\mu_1\mu_2\cdots\mu_k}\right)^{-1}. \tag{2.8}$$

Substitution of (2.7) into (2.6) yields

$$q_n[N] = \frac{\frac{N(N-1)\cdots(N-n+1)\lambda^n}{\mu_1\mu_2\cdots\mu_n}P_0[N]}{NP_0[N] + \sum_{k=1}^{N-1}\frac{N(N-1)\cdots(N-k+1)(N-k)\lambda^k}{\mu_1\mu_2\cdots\mu_k}P_0[N]}$$
$$n = 1, 2, ..., N-1. \tag{2.9}$$

After cancellation of the factor $NP_0[N]$ in (2.9), we have

$$q_n[N] = \frac{\frac{(N-1)\cdots(N-n)\lambda^n}{\mu_1\mu_2\cdots\mu_n}}{1 + \sum_{k=1}^{N-1}\frac{(N-1)\cdots(N-k)\lambda^k}{\mu_1\mu_2\cdots\mu_k}} \quad n = 1, 2, ..., N-1. \tag{2.10}$$

Comparison of Equation (2.10) with (2.7) and (2.8) shows that $q_n[N] = P_n[N-1]$ for $j = 1, 2, ..., N-1$. Since we must have

$$\sum_{n=0}^{N-1}P_n[N-1] = \sum_{n=0}^{N-1}q_n[N] = 1, \tag{2.11}$$

we conclude that

$$q_n[N] = P_n[N-1] \quad n = 0, 1, ..., N-1. \tag{2.12}$$

It should be mentioned that this theorem can be generalized to closed queueing networks stating:
In a closed queueing network the (stationary) state probabilities at customer arrival epochs are identical to those of the same network in long-term equilibrium with one customer removed.

The q_n's in (2.3) are written as $q_n[N]$ in the notation (2.12). By

$$P_n = \begin{cases} \binom{N}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, & n = 0, 1, ..., r \\ \frac{n!}{r! r^{n-r}} \binom{N}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, & n = r+1, ..., N, \end{cases} \tag{2.13}$$

$$P_0 = \left(\sum_{k=0}^{r} \binom{N}{k} \left(\frac{\lambda}{\mu}\right)^k + \sum_{k=r+1}^{N} \frac{k!}{r! r^{k-r}} \binom{N}{k} \left(\frac{\lambda}{\mu}\right)^k \right)^{-1} \tag{2.14}$$

when $n \geq r$, we can write, using $z = \mu/\lambda$,

$$\begin{aligned} P_n[N] &= \frac{n!}{r! r^{n-r}} \binom{N}{n} z^n P_0[N] = \frac{n! r^r}{r! r^n} \frac{N! z^n}{(N-n)! n!} P_0[N] \\ &= \frac{r^r}{r!} P_0[N] \frac{\frac{e^{-rz}(rz)^{N-n}}{(N-n)!}}{\frac{e^{-rz}(rz)^N}{N!}} = \frac{r^r}{r!} P_0[N] \frac{P[N-n; rz]}{P[N; rz]}, \end{aligned} \tag{2.15}$$

where $PD(k, rz)$ denotes the Poisson distribution with parameter rz, that is $PD(k, rz) = \frac{(rz)^k}{k!} e^{-rz}$.

Therefore,

$$q_n[N] = P_n[N-1] = \frac{r^r}{r!} P_0[N-1] \frac{PD[N-n-1; rz]}{PD[N-1; rz]}. \tag{2.16}$$

Substituting the above formula for $q_n[N]$ into (2.3) yields

$$\begin{aligned} P[q > t] &= \sum_{n=r}^{N-1} P_n[N-1] Q[n-r; r\mu t] \\ &= \frac{r^r P_0[N-1]}{r! PD[N-1; rz]} \sum_{n=r}^{N-1} PD[N-n-1; rz] Q[n-r; r\mu t] \\ &= \frac{r^r P_0[N-1]}{r! PD[N-1; rz]} \sum_{n=0}^{N-r-1} PD[N-r-1-n; rz] Q[n; r\mu t] \\ &= \frac{r^r P_0[N-1]}{r! PD[N-1; rz]} Q[N-r-1; rz + r\mu t], \end{aligned} \tag{2.17}$$

In the last step of (2.17), we applied the identity

$$\sum_{j=0}^{k} P[k-j; \lambda] Q[j; \mu] = Q[k; \lambda + \mu].$$

Now we can use (2.17) to write

$$W[t] = 1 - \frac{r^r P_0[N-1] Q[N-r-1; r(z+\mu t)]}{r! PD[N-1; rz]}. \tag{2.18}$$

The derivation of the distribution function $R[\cdot]$ of the response time, or sojourn time of requests staying in the system is much more complicated. It is derived in the solutions manual for Kobayashi [25]. The result is

$$R[t] = 1 - C_1 \times e^{-\mu t} + C_2 \times Q[N-r-1; r(z+\mu t)] \quad t \geq 0, \tag{2.19}$$

where

$$C_1 = 1 + C_2 \times Q[N-r-1; rz], \tag{2.20}$$

and

$$C_2 = \frac{r^r P_0[N-1]}{r!(r-1)(N-r-1)! PD[N-1; rz]}. \tag{2.21}$$

Formula (2.14) (with N replaced by $N-1$ everywhere it appears) can be used to calculate $P_0[N-1]$ in (2.18) and (2.21).

To make easier the calculations, as it was shown in Kobayashi [25], $P_0^{-1}[N]$ satisfies the following recurrence relation

$$P_0^{-1}[N] = 1 + \frac{N}{rz} P_0^{-1}[N-1] + \frac{N}{z} \sum_{i=0}^{r-1} \frac{\binom{N-1}{i}}{z^i} \left[\frac{1}{i+1} - \frac{1}{r} \right], \quad N > r$$

with initial value

$$P_0^{-1}[r] = \left(1 + \frac{1}{z}\right)^r.$$

2.2 The $\vec{G}/M/r//N/FCFS$ sytem

Requests emanate from a finite source of size N and are served by one of $r(r \leq N)$ servers at a service facility according to a First-Come-First-Served (FCFS) discipline. The service times of the requests are supposed to be identically and exponentially distributed random variables with means $1/\mu$. After completing service, request i returns to the source and stays there for a random time having general distribution function $F_i(x)$ with density $f_i(x)$. All of these random variables are assumed to be independent of each other.

2.2.1 The mathematical model

Let the random variable $v(t)$ denote the number of requests staying in the source at time t and $(\alpha_1(t), \ldots, \alpha_{v(t)}(t))$ indicate their indices ordered lexicographically. Let us denote by $(\beta_1(t), \ldots, \beta_{N-v(t)}(t))$ the indices of the requests waiting for the service facility in the order of their arrival. Clearly the sets $\{\alpha_1(t), \ldots, \alpha_{v(t)}(t)\}$ and $\{\beta_1(t), \ldots, \beta_{N-v(t)}(t)\}$ are disjoint.

Introduce the process

$$\underline{Y}(t) = (\alpha_1(t), \ldots, \alpha_{v(t)}(t); \beta_1(t), \ldots, \beta_{N-v(t)}(t)).$$

The stochastic process $(\underline{Y}(t), t \geq 0)$ is not Markovian unless the distribution functions $F_i(x)$ are exponential, $i = 1, \ldots, N$.

Let us also introduce the supplementary variables $\xi_{\alpha_l(t)}$ to denote the random time that request $\alpha_l(t)$ has been spending in the source until time t, $l = 1, \ldots, N$. Define

$$\underline{X}(t) = (\alpha_1(t), \ldots, \alpha_{v(t)}(t); \xi_{\alpha_1(t)}, \ldots, \xi_{\alpha_{v(t)}}(t); \beta_1(t), \ldots, \beta_{N-v(t)}(t)).$$

Then process $(\underline{X}(t), t \geq 0)$ exhibits the Markov property.

Let V_k^N and C_k^N denote the set of all variations and combinations of order k of the integers $1, 2, \ldots, N$ respectively, ordered lexicographically. Then the state space of the process $\underline{X}(t)$ consist of the sets

$$(i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{N-k}), \qquad (i_1, \ldots, i_k) \in C_k^N,$$
$$(j_1, \ldots, j_{N-k}) \in V_{N-k}^n, \qquad x_i \in \mathbf{R}_+, \qquad i = 1, \ldots, k, \quad k = 0, \ldots, N$$

Let $Q_{i_1,\ldots,i_k;j_1,\ldots,j_{N-k}}(x_1, \ldots, x_k; t)$ denote the probability that at time t the process is in state $(i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{N-k})$ if k requests with indices $(i_1, \ldots, i_k)$ have been staying in the source for times $(x_1, \ldots, x_k)$ respectively, while the rest need service and their indices in order of arrival are $j_1, \ldots, j_{N-k})$.

Let λ_i defined by $1/\lambda_i = \int_0^\infty x dF_i(x)$. Then we have:

Theorem 2.2.1. *If $1/\lambda_i < \infty, i = 1, \ldots, N$, then the process $(\underline{X}(t), t \geq 0)$ possesses a unique limiting (stationary) ergodic distribution independent of the initial conditions, namely*

$$Q_{0;j_1,\ldots,j_N} = \lim_{t\to\infty} Q_{0;j_1,\ldots,j_N}(t),$$
$$Q_{i_1,\ldots,i_k;j_1,\ldots,j_{N-k}}(x_1, \ldots, x_k) = \lim_{t\to\infty} Q_{i_1,\ldots,i_k;j_1,\ldots,j_{N-k}}(x_1, \ldots, x_k; t). \tag{2.22}$$

Notice that $\underline{X}(t)$ belongs to the class of piecewise-linear Markov processes, subject to discontinuous changes treated by [17] in detail. Our statement follows from the theorem on page 211 of that monograph.

Let $Q_{i_1,\ldots,i_k,j_1,\ldots,j_{N-k}}$ denote the steady state probability that requests with indices $(i_1,\ldots,i_k)$ are in the source and the order of arrival of the rest to the service facility is $(j_1,\ldots,j_{N-k})$. Furthermore, denote by $Q_{i_1,\ldots,i_k}$ the steady state probability that requests with indices $(i_1,\ldots,i_k)$ are staying at the source.
As it was proved in [35] that these probabilities can be expressed in the following form

$$Q_{i_1,\ldots,i_k} = \frac{(N-k)!}{r!r^{N-r-k}\mu^{N-k}\lambda_{i_1},\ldots,\lambda_{i_k}}C_N, \qquad (2.23)$$
$$(i_1,\ldots,i_k)\in C_k^N,\ k=0,1,\ldots,N-r.$$

Similarly,

$$Q_{i_1,\ldots,i_k} = \frac{1}{\mu^{N-k}\lambda_{i_1}\ldots\lambda_{i_k}}C_N \qquad (2.24)$$
$$(i_1,\ldots,i_k)\in C_k^N,\ k=N-r,\ldots,N. \qquad (2.25)$$

Let $\hat{Q}_k$ and $\hat{P}_l$ denote the steady state probabilities that k requests are staying in the source and l requests are at the service facility, respectively. Clearly

$$Q_{i_1,\ldots,i_N} = Q_{1,\ldots,N} = \hat{Q}_N = \hat{P}_0 \qquad \hat{Q}_k = \hat{P}_{N-k}.$$

It is easy to see that

$$C_n = \hat{Q}_n\lambda_1\ldots\lambda_n \qquad \text{and} \qquad \hat{Q}_k = \sum_{(i_1,\ldots,i_k)\in C_k^N} Q_{i_1,\ldots,i_k},$$

where $\hat{Q}_N$ can be obtained with the aid of the norming condition

$$\sum_{k=0}^{N}\hat{Q}_k = 1.$$

In the homogeneous case, when $\lambda_i = \lambda, \quad i=1,\ldots,N$ relations (2.23) and (2.2.1) yield

$$\hat{Q}_k = \frac{N!}{k!r!r^{N-r-k}}\left(\frac{\lambda}{\mu}\right)^{N-k}\hat{Q}_N \qquad \text{for } 0\le k\le N-r,$$
$$\hat{Q}_k = \binom{N}{k}\left(\frac{\lambda}{\mu}\right)^{N-k}\hat{Q}_n \qquad \text{for } N-r\le k\le N.$$

Thus, the probability that k requests are not in the source is

$$\hat{P}_k = \binom{N}{k}\left(\frac{\lambda}{\mu}\right)^k \hat{P}_0 \qquad \text{for } 0 \le k \le r,$$

$$\hat{P}_k = \frac{N!}{(N-k)!r!r^{k-r}}\left(\frac{\lambda}{\mu}\right)^k \hat{P}_0 \qquad \text{for } r \le k \le N.$$

This is exactly the same result as the one obtained in [8]. The equivalence of the finite-source $E_k/M/1$ to the $M/M/1$ and in addition to that of the $G/M/r$ to the $M/M/r$ model (see 2.1), respectively, are just special cases of the more general result obtained here.

Before determining the main characteristics of the system we need one more theorem. In order to formulate it, we introduce some further notations. Let $Q^{(i)}(P^{(i)})$ denote the steady state probability that request i is in the source (at the service facility) for $i = 1, \ldots, N$. It is clear that the process $(\underline{Y}(t), t \ge 0)$ is a Markov-regenerative process with state space

$$\bigcup_{\substack{(i_1,\ldots,i_k) \in C_k^N, \quad (j_1,\ldots,j_{N-k}) \in V_{N-k}^N, \\ (i_1,\ldots,i_k) \cap (j_1,\ldots,j_{N-k}) = 0, \\ k=0,1,\ldots,N}} \{(i_1,\ldots,i_k; j_1,\ldots,j_{N-k})\}.$$

Let H_i be the event that request i is in the source and $Z_{H_i}(t)$ its characteristic function, that is

$$Z_{H_i}(t) = \begin{cases} 1 & \text{if } \underline{Y}(t) \in H_i \\ 0 & \text{otherwise} \end{cases}$$

Then we have

Theorem 2.2.2.

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T Z_{H_i}(t)dt = \frac{1/\lambda_i}{1/\lambda_i + \overline{W}_i + 1/\mu} = Q^{(i)} = 1 - P^{(i)},$$

where $\overline{W}_i$ denotes the mean waiting time of request i.

The statement is a special case of a theorem concerning the expected sojourn time for semi-Markov processes, (see [45]).

Sometimes we need the long-run fraction of time the request i spends in the source. This happens e.g., in the *machine interference model*. In that case for the utilization of machine i we have

$$U_i = Q^{(i)} = \sum_{k=1}^{n} \sum_{i \in (i_1,\ldots,i_k) \in C_k^N} Q_{i_1,\ldots,i_k}.$$

2.2.2 The main performance measures

(i) **Utilizations**

Utilizations can now be considered for individual servers or for the system as a whole. The process $(\underline{X}(t), t \geq 0)$ is assumed to be in equilibrium. Considering the system as the whole, it will be empty only when there are no requests at the service facility and will be busy at other times. As usual, using renewal-theoretic arguments for the system utilization, that is the long-run fraction of time when at least one server is busy, we have

$$U = 1 - \hat{Q}_N \qquad \text{and} \qquad \hat{Q}_N = \frac{E_{\eta^*}}{E_{\eta^*} + E\delta}$$

where $\eta^* = min(\eta_1, \ldots, \eta_N)$, random variable η_i denotes the source time of request i, $i = 1, \ldots, N$, and $N\delta$ denotes the average busy period of the system.

Thus the expected length of the busy period is given by

$$E\delta = E_{\eta^*} \frac{1 - \hat{Q}_N}{\hat{Q}_N}.$$

In particular, if $F_i(x) = 1 - exp(-\lambda_i x),\ i = 1, \ldots, N$, we get

$$E\delta = \frac{1 - \hat{Q}_N}{\hat{Q}_N} \frac{1}{\sum \lambda_i}.$$

It is also easy to see that for the utilization of a given server, which is called utilization in general, the following relation holds:

$$U_s = \frac{1}{r} \left(\sum_{k=1}^{N} k\hat{P}_k + r \sum_{k=r+1}^{N} \hat{P}_k \right) = \frac{\overline{r}}{r},$$

where $\overline{r}$ denotes the mean number of busy servers.

(ii) **Mean waiting times**

By the virtue of Theorem 2.2.1 we obtain $Q^{(i)} = (1 + \lambda_i \overline{W}_i + \lambda_i/\mu)^{-1}$. Consequently, the average waiting time of request i is

$$\overline{W}_i = (1 - Q^{(i)})(\lambda_i Q^{(i)})^{-1} - 1/\mu.$$

It follows that the mean sojourn time of request i, that is, the sum of waiting and service times , can be obtained by

$$\overline{T}_i = \overline{W}_i + 1/\mu = (1 - Q^{(i)})(\lambda_i Q^{(i)})^{-1} \qquad \text{for } i = 1, \ldots, N. \tag{2.26}$$

Since $\sum_{i=1}^{N}(1 - Q^{(i)}) = \overline{N}$, where $\overline{N}$ denotes the mean number of requests staying at the service facility we have, by reordering and adding (2.26)

$$\sum_{i=1}^{N} \lambda_i \overline{T}_i Q^{(i)} = \overline{N}. \tag{2.27}$$

This is the **Little's formula** for the finite source $\vec{G}/M/r$ queue. In particular, if $F_i(x) = F(x),\ i = 1, \ldots, N$, (2.27) can be written as $\lambda(N - \overline{N})\overline{T} = \overline{N}$, where $(N - \overline{N})$ is the expected number of requests staying in the source.

Using the method of supplementary variables similar problems were treated in [34, 36, 37, 38, 39].

Chapter 3

Numerical Methods

Closed-form solutions for the steady-state probabilities are very rare. Different analytical methods are used to investigate the involved processes and related numerical problems. For the most common procedures and tools the interested reader is referred to [6, 16, 19, 20, 21, 27, 29, 30, 44, 47]

3.1 A recursive method for the $M/G/1//N$ system

In the following the results of [19] are introduced since it give a very stable algorithm for the calculations. Takács [40] gives an explicit expression for the stationary distribution of the number of working (up) machines of the $M/G/1//N$ model. However, for a large number of machines the computation of probabilities using Theorem 2 in [40] (p. 195) may pose problem as it involves many factorials. Even for the simple $M/M/r//N$ model, Gross and Harris [18] (p. 108) makes similar comments and proposed a recursive method for computing probabilities. To obtain the steady state probability distribution of the number of down machines at arbitrary time epoch $P_n(0 \leq n \leq N)$ one can also use the embedded Markov chain technique, see [42].
The objective of this section is to provide an alternative method, using the supplementary variable technique and considering the supplementary variable as the remaining repair time, to obtain $P_n(0 \leq n \leq N)$ for $M/G/1//N$ model which is used to obtain the various system performance measures such as average number of down machines, average waiting time and operator utilization etc. The method is recursive and can be used for several repair time distribution such as mixed generalized Erlang (MGE_h), generalized Erlang (GE_h), hyperexponential (HE_h), generalized hyperexponential (GH_h) and uniform $U(a,b)$ etc. The only input required for efficient evaluation of state probabilities is the Laplace-Stieltjes Transform of the repair time distribution.

3.1.1 The mathematical model

Consider a machine repairman problem with a single repairman and a set of N working machines. Let us assume that the running times of the machines between breakdowns have an exponential distribution with mean $1/\lambda$ and the repair (service) time of the machines are independent and identically distributed random variables having distribution function $B(u)$, probability density function $b(u)$ and a mean repair time b. The state of the system at time t is given by

$N(t) =$ Number of down machines, and
$U(t) =$Remaining repair time for the machine under repair.

Let us define

$$P_0(t) = P(N(t) = 0), \tag{3.1}$$

and

$$\begin{aligned} P_n(u,t)du &= P\{N(t) = n, u < U(t) \leq u + du\}, \\ &\quad u \geq 0, \quad n = 1, 2, \ldots, N. \end{aligned} \tag{3.2}$$

$$P_n(t) = P(N(t) = n) = \int_0^\infty P_n(u,t)du, \quad n = 1, 2, \ldots, N. \tag{3.3}$$

Relating the states of the system at time t and $t + dt$, we obtain

$$\frac{\partial}{\partial t} P_0(t) = -N\lambda P_0(t) + P_1(0,t), \tag{3.4}$$

$$\begin{aligned} \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial u}\right) P_1(u,t) &= -(N-1)\lambda P_1(u,t) + N\lambda P_0(t) b(u) + \\ &\quad + P_2(0,t) b(u), \end{aligned} \tag{3.5}$$

$$\begin{aligned} \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial u}\right) P_r(u,t) &= -(N-r)\lambda P_r(u,t) + (N-r+1)\lambda P_{r-1}(u,t) + \\ &\quad + P_{r-1}(0,t) b(u), \qquad 2 \leq r \leq N-1 \end{aligned} \tag{3.6}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial u}\right) P_N(u,t) = \lambda P_{N-1}(u,t). \tag{3.7}$$

Since we discuss the model in steady state, we let $t \to \infty$ in equations (3.4)-(3.7).
Further define

$$P_n = \lim_{t \to \infty} P_n(t), \quad 0 \leq n \leq N \tag{3.8}$$

$$P_n(u) = \lim_{t \to \infty} P_n(u,t), \quad 1 \leq n \leq N. \tag{3.9}$$

$$B^*(s) = \int_0^\infty e^{-su} dB(u) = \int_0^\infty e^{-su} b(u) du, \tag{3.10}$$

$$\left.\begin{aligned} P_n^*(s) &= \textstyle\int_0^\infty e^{-su} P_n(u) du \qquad & 1 \leq n \leq N \\ P_n &= P_n^*(0) = \textstyle\int_0^\infty P_n(u) du, \qquad & 1 \leq n \leq N. \end{aligned}\right\} \tag{3.11}$$

and

$$\int_0^\infty e^{-su} \frac{\partial}{\partial u} P_n(u) du = s P_n^*(s) - P_n(0). \tag{3.12}$$

From (3.4)-(3.12) and the fact that all derivatives with respect to t are zero, it follows that

$$N\lambda P_0 = P_1(0), \quad (3.13)$$

$$((N-1)\lambda - s)P_1^*(s) = N\lambda P_0 B^*(s) + P_2(0)B^*(s) - P_1(0), \quad (3.14)$$

$$((N-r)\lambda - s)P_r^*(s) = (N-r+1)\lambda P_{r-1}^*(s) + P_{r+1}(0)B^*(s) - P_r(0), \quad 2 \le r \le N-1 \quad (3.15)$$

$$-sP_N^*(s) = \lambda P_{N-1}^*(s) - P_N(0). \quad (3.16)$$

Using (3.13) in (3.14) and then adding (3.14) to (3.16), we obtain

$$\sum_{r=1}^{N} P_r^*(s) = \frac{1-B^*(s)}{s} \sum_{r=1}^{N} P_r(0). \quad (3.17)$$

Taking $s \to 0$ in (3.17), we get

$$\sum_{r=1}^{N} P_r^*(0) = b_1 \sum_{r=1}^{N} P_r(0), \quad (3.18)$$

where $b = -B^{*(1)}(0)$ is mean repair time.

Our main objective is to obtain $P_n \equiv P_n^*(0)(1 \le n \le N)$ from (3.13 -(3.16). To achieve it, our strategy will be to obtain first $P_n(0)(1 \le n \le N)$ and then using it we finally evaluate $P_n^*(0)(1 \le n \le N)$.

Using (3.13) in (3.14) and then setting $s = (N-1)\lambda$ and $s = 0$ respectively in (3.14), we get

$$P_2(0) = \frac{N\lambda(1 - B^*((N-1)\lambda))}{B^*((N-1)\lambda)} P_0, \quad (3.19)$$

and

$$P_1^*(0) = \frac{1}{(N-1)\lambda} P_2(0). \quad (3.20)$$

Now setting $s = (N-r)\lambda$, in (3.15), we obtain

$$P_{r+1}(0) = \frac{1}{B^*((N-r)\lambda)}(P_r(0) - (N-r+1)\lambda P_{r-1}^*((N-r)\lambda)), \quad 2 \le r \le N-1. \quad (3.21)$$

Setting $s = (N-r)\lambda$ in (3.14) for $r = 2, 3, \ldots, N-1$, we get

$$P_1^*((N-r)\lambda) = \frac{1}{(r-1)\lambda}(N\lambda P_0\{B^*((N-r)\lambda) - 1\} + P_2(0)B^*((N-r)\lambda)). \quad (3.22)$$

From equation (3.15) for $r = 3, 4, \ldots, N-1$, we get

$$\begin{aligned} P_j^*((N-r)\lambda) &= \frac{1}{(r-j)\lambda}\Big((N-j+1)\lambda P_{j-1}^*((N-r)\lambda) + \\ &\quad + P_{j+1}(0)B^*((N-r)\lambda) - P_j(0)\Big), \qquad (3.23) \\ &\quad 2 \leq j \leq r-1. \end{aligned}$$

Hence $P_3(0), P_4(0), \ldots, P_N(0)$ can be obtained recursively using (3.19), (3.22), (3.23) and (3.21) in terms of P_0.

Now setting $s = 0$ in (3.15), we get

$$\begin{aligned} P_r^*(0) &= \frac{1}{(N-r)\lambda}\Big((N-r+1)\lambda P_{r-1}^*(0) + P_{r+1}(0) - P_r(0)\Big), \\ &\quad 2 \leq r \leq N-1. \qquad (3.24) \end{aligned}$$

As $P_2(0), P_3(0), \ldots, P_N(0)$ are known, $P_2^*(0), P_3^*(0), \ldots, P_{N-1}(0)$ can be determined recursively using (3.20) and (3.24) in terms of P_0.

Now the only unknown quantity is $P_N^*(0)$ which can be obtained from equation (3.16). To obtain it, differentiate equation (3.16) with respect to s and set $s = 0$, we get

$$P_N^*(0) = -\lambda P_{N-1}^{*(1)}(0). \qquad (3.25)$$

To get $P_{N-1}^{*(1)}(0)$, differentiate (3.15) and (3.14) with respect to s and set $s = 0$.

$$\begin{aligned} P_r^{*(1)}(0) &= \frac{1}{(N-r)\lambda}\Big((N-r+1)\lambda P_{r-1}^{*(1)}(0) \\ &\quad + P_{r+1}(0)B^{*(1)}(0) + P_r^*(0)\Big), \qquad 2 \leq r \leq N-1, \qquad (3.26) \\ P_1^{*(1)}(0) &= \frac{1}{(N-r)\lambda}\Big(N\lambda P_0 B^{*(1)}(0) + P_2(0)B^{*(1)}(0) + P_1^*(0)\Big). \qquad (3.27) \end{aligned}$$

As $P_1^{*(1)}(0)$ is known completely from (3.27), $P_r^{*(1)}(0), (2 \leq r \leq N-1)$ can be determined recursively from (3.26) and hence $P_N^*(0)$ is known from (3.25). So $P_n^*(0)(1 \leq n \leq N)$ is known in terms of P_0, which can be determined using the normalizing condition

$$P_0 + \sum_{n=1}^{N} P_n^*(0) = 1. \qquad (3.28)$$

The steady state probability distribution of the number of down machines at service completion or departure epoch $\pi_n(0 \leq n \leq N-1)$ can also be obtained from $P_r(0)(1 \leq r \leq N)$ and is given by

$$\pi_n = \frac{P_{n+1}(0)}{\sum_{r=1}^{N} P_r(0)}, \quad n = 0, 1, \ldots, N-1. \qquad (3.29)$$

To demonstrate this method we consider a simple example where the repair time distribution is exponential and number of machines (N) is four, i.e. the $M/M/1//4$ model. In this case

$$B^*(s) = \frac{\mu}{\mu + s}$$

From (3.19), we get

$$P_2(0) = \frac{4\lambda(1 - B^*(3\lambda))}{B^*(3\lambda)} P_0.$$

Now from (3.21), we have

$$P_3(0) = \frac{1}{B^*(2\lambda)}(P_2(0) - 3\lambda P_1^*(2\lambda)),$$

where $P_1^*(2\lambda)$ is obtained from (3.22)

$$P_1^*(2\lambda) = \frac{1}{\lambda}(4\lambda P_0\{B^*(2\lambda) - 1\} + P_2(0)B^*(2\lambda)).$$

Now again from (3.21), we have

$$P_4(0) = \frac{1}{B^*(\lambda)}(P_3(0) - 2\lambda P_2^*(\lambda)).$$

where $P_2^*(\lambda)$ is obtained from (3.23)

$$P_2^*(\lambda) = \frac{1}{\lambda}(3\lambda P_1^*(\lambda) + P_3(0)B^*(\lambda) - P_2(0)).$$

To know $P_2^*(\lambda)$ we need $P_1^*(\lambda)$ which can be obtained from (3.22)

$$P_1^*(\lambda) = \frac{1}{2\lambda}(4\lambda P_0\{B^*(\lambda) - 1\} + P_2(0)B^*(\lambda)).$$

From above we get

$$P_2(0) = 12\tfrac{\lambda^2}{\mu}P_0,$$
$$P_1^*(2\lambda) = \tfrac{4\lambda}{\mu+2\lambda}P_0, \quad P_3(0) = 24\tfrac{\lambda^3}{\mu^2}P_0, \quad P_1^*(\lambda) = \tfrac{4\lambda}{\mu+\lambda}P_0,$$
$$P_2^*(\lambda) = 12\tfrac{\lambda^2}{\mu(\mu+\lambda)}P_0, \quad P_4(0) = 24\tfrac{\lambda^4}{\mu^3}P_0.$$

Hence from (3.20) and (3.24), we get

$$P_1^*(0) = \frac{1}{3\lambda}P_2(0) = \frac{4\lambda}{\mu}P_0,$$
$$P_2^*(0) = \frac{1}{2\lambda}P_3(0) = 12\frac{\lambda^2}{\mu^2}P_0,$$
$$P_3^*(0) = \frac{1}{\lambda}P_4(0) = 24\frac{\lambda^3}{\mu^3}P_0.$$

Finally to determine $P_4^*(0)$ we have from (3.25)

$$P_4(0) = -\lambda P_3^{*(1)}(0),$$

where $P_3^{*(1)}(0)$ can be obtained from (3.26)

$$P_3^{*(1)}(0) = \frac{1}{\lambda}(2\lambda P_2^{*(1)}(0) + P_4(0)B^{*(1)}(0) + P_3^*(0)),$$

again $P_2^{*(1)}(0)$ an be obtained from (3.26)

$$P_2^{*(1)}(0) = \frac{1}{2\lambda}(3\lambda P_1^{*(1)}(0) + P_3(0)B^{*(1)}(0) + P_2^*(0)),$$

To know $P_2^{*(1)}(0)$ we need $P_1^{*(1)}(0)$ which can be obtained from (3.27)

$$P_1^{*(1)}(0) = \frac{1}{3\lambda}(4\lambda P_0 B^{*(1)}(0) + P_2(0)B^{*(1)}(0) + P_1^*(0)).$$

From above we get

$$P_1^{*(1)}(0) = -4\frac{\lambda}{\mu^2}P_0, P_2^{*(1)}(0) = -12\frac{\lambda^2}{\mu^3}P_0, P_3^{*(1)}(0) = -24\frac{\lambda^3}{\mu^4}P_0,$$

and hence

$$P_4^*(0) = 24\frac{\lambda^4}{\mu^4}P_0$$

by using $\rho = \frac{\lambda}{\mu}$ we have

$$P_1^*(0) = 4\rho P_0, P_2^*(0) = 12\rho^2 P_0, P_3^*(0) = 24\rho^3 P_0, P_4^*(0) = 24\rho^4 P_0.$$

Since $P_0 + P_1^*(0) + P_2^*(0) + P_3^*(0) + P_4^*(0) = 1$, we get

$$P_0 = \frac{1}{1 + 4\rho + 12\rho^2 + 24\rho^3 + 24\rho^4}.$$

It can be easily seen that this result matches with the expression given in [18] p. 105.

This system has been generalized to $\vec{M}/\vec{G}/1//N/FIFO$ system which can be found in [32, 33].

3.2 Homogeneous finite-source retrial queues with server subject to breakdowns and repairs

Retrial queues have been widely used to model many problems arising in telecommunication networks, computer networks and computer systems, etc. For a systematic account of the fundamental methods and results, and for an accessible classified bibliography on this topic the interested reader is referred to [5], [13], [14], and references therein.

In many practical situations it is important to take into account the fact that the rate of generation of new primary calls decreases as the number of customers in the system increases. This can be done with the help of finite-source, or quasi-random input models. A complete survey on related results can be found in Artalejo [5] for systems of type $M/G/1//K$ and $M/M/c//K$.

In this section finite-source systems with the following assumptions are investigated. Following the widely accepted notation of papers dealing with finite-source retrial queues, we use a different notation as we did in the previous chapters. Consider a single server system, where the primary calls are generated by K, $1 < K < \infty$ homogeneous sources. The server can be in three states: idle, busy and failed. If the server is idle, it can serve the calls of the sources. Each of the sources can be in three states: free, sending repeated calls and under service. If a source is free at time t it can generate a primary call during interval $(t, t + dt)$ with probability $\lambda dt + o(dt)$. If the server is free at the time of arrival of a call then the call starts to be served immediately, the source moves into the under service state and the server moves into busy state. The service finishes during the interval $(t, t + dt)$ with probability $\mu dt + o(dt)$ if the server is available. If the server is busy, then the source starts generation of a Poisson flow of repeated calls with rate ν until it finds the server free. After service the source becomes free, and it can generate a new primary call, and the server becomes idle so it can serve a new call. The server can fail during the interval $(t, t + dt)$ with probability $\delta dt + o(dt)$ if it is idle, and with probability $\gamma dt + o(dt)$ if it is busy. If $\delta = 0, \gamma > 0$ or $\delta = \gamma > 0$ *active or independent breakdowns* can be discussed, respectively. If the server fails in busy state, it either *continues servicing* the interrupted call after it has been repaired or the interrupted request *returns to the orbit*. The repair time is exponentially distributed with a finite mean $1/\tau$. If the server is failed two different cases can be treated. Namely, *blocked sources* case when all the operations are stopped, that is no new primary and repeated calls are generated. In the *unblocked (intelligent) sources* case only service is interrupted but all the other operations are continued (new and repeated calls can be generated). All the random times involved in the model are assumed to be mutually independent of each other.
As it can be seen this systems is rather complicated since it involves two types of failures, continued or repeated service and blocked or unblocked operations during breakdowns.

Our objective is to give the main usual stationary performance and reliability measures of the system and to display the effect of different parameters on them. To achieve this goal a tool called MOSEL (Modeling, Specification and Evaluation Language) developed at the University of Erlangen, Germany, see [6], is used to formulate and solve the problem. We show how this system can be modelled, and how easily performance measures can be represented graphically using IGL (Intermediate Graphical Language). For more detailed information on this topic please read [2].

3.2.1 The underlying Markov chain

The system state at time t can be described with the process $X(t) = (Y(t); C(t); N(t))$, where $Y(t) = 0$ if the server is up, $Y(t) = 1$ if the server is failed, $C(t) = 0$ if the server is idle, $C(t) = 1$ if the server is busy, $N(t)$ is the number of sources with repeated calls at time t. Because of the exponentiality of the involved random variables this process is a Markov–chain with finite state space $S = \{0,1\} \times \{0,1\} \times \{0,1,...,K-1\}$. Since the state space of the process $(X(t), t \geq 0)$ is finite, the process is ergodic for all values of the rate of generation of primary calls, and from now on we will assume that the system is in the steady state.

We define the stationary probabilities:

$$P(q;r;j) = \lim_{t\to\infty} P(Y(t) = q, C(t) = r, N(t) = j),$$
$$q = 0,1, \quad r = 0,1, \quad j = 0,...,K-1.$$

Knowing these quantities the **main performance measures** can be obtained as follows:

- *Utilization of the server*

$$U_S = \sum_{j=0}^{K-1} P(0,1,j)$$

- *Utilization of the repairman*

$$U_R = \sum_{q=0}^{1} \sum_{j=0}^{K-1} P(1,q,j),$$

- *Availability of the server*

$$A_S = \sum_{q=0}^{1} \sum_{j=0}^{K-1} P(0,q,j) = 1 - U_R,$$

- *The mean number of sources of repeated calls*

$$N = E[N(t)] = \sum_{q=0}^{1} \sum_{r=0}^{1} \sum_{j=0}^{K-1} jP(q,r,j),$$

- *The mean number of calls staying in the orbit or in service*

$$M = E[C(t) + N(t)] = \sum_{q=0}^{1}\sum_{r=0}^{1}\sum_{j=0}^{K-1}(r+j)P(q,r,j),$$

- *The mean rate of generation of primary calls*

$$\overline{\lambda} = \begin{cases} \lambda E[K - C(t) - N(t); Y(t) = 0] & \text{for blocked case,} \\ \lambda E[K - C(t) - N(t)] & \text{for unblocked case,} \end{cases}$$

- *The mean response time*

$$E[T] = M/\overline{\lambda},$$

- *The mean waiting time*

$$E[W] = N/\overline{\lambda},$$

- *The blocking probability of a primary call*

$$B = \begin{cases} \frac{\lambda E[K-C(t)-N(t);Y(t)=0;C(t)=1]}{\overline{\lambda}} & \text{for blocked case,} \\ \\ \frac{\lambda E[K-C(t)-N(t);C(t)=1]}{\overline{\lambda}} & \text{for unblocked case.} \end{cases}$$

We used the software tool MOSEL to formulate the model and to calculate the main performance measures. The figures in the next section are automatically generated by the tool.

3.2.2 Numerical examples

In this section we consider some sample numerical results to illustrate the influence of the non–reliable server on the mean response time $E[T]$. The results in the reliable case were validated by the Pascal program given in [14], too.

Input parameters
As it can be seen in the first 3 cases the independent breakdowns are treated, then the state dependent and independent ones are considered. In each case different comparisons are made according to the breakdowns *(dependent, independent)*, service continuation *(continued)* and system operations *(blocked, unblocked)*.

- In Figures 3.1–3.3 we can see the mean response time $E[T]$ for the reliable and the non–reliable retrial system with continuous, non-continuous service after repair, with blocked and unblocked operations during service failure when the primary request generation rate, retrial rate and service rate increase. In these cases, the server's failure rate is independent of the state of the server (busy or

	NT	λ	μ	ν	δ	γ	τ
Figure 1	6	x axis	4	0.4	0.05	0.05	0.1
Figure 2	6	5	10	x axis	0.05	0.05	0.1
Figure 3	6	0.1	x axis	0.4	0.05	0.05	0.1
Figure 4	6	x axis	4	0.4	0.005(0.05)	0.05	0.1
Figure 5	6	5	10	x axis	0.005(0.05)	0.05	0.1
Figure 6	6	0.1	x axis	0.4	0.005(0.05)	0.05	0.1

Table 3.1: Input parameters

idle). Figure 3.1 demonstrates a surprising phenomenon of retrial queues having a maximum of $E[T]$ which was noticed in [15], too. The difference between continuous, non-continuous service, moreover blocked, unblocked (intelligent) systems's operations is clearly shown. However, if the retrial or service rate increases the continuous and non-continuous service result in the same measure, as it was expected, see Figure 3.2, 3.3.

- In Figures 3.4–3.6 the mean response time $E[T]$ is displayed with continuous service after repair but the server's failure rate depends on its state. The system operation is either blocked or unblocked. In Figure 3.4 we can see that the curves of independent failure with blocked operations and dependent failures with unblocked operations intersect each other. In each case the difference between the independent and dependent failures is clearly demonstrated.

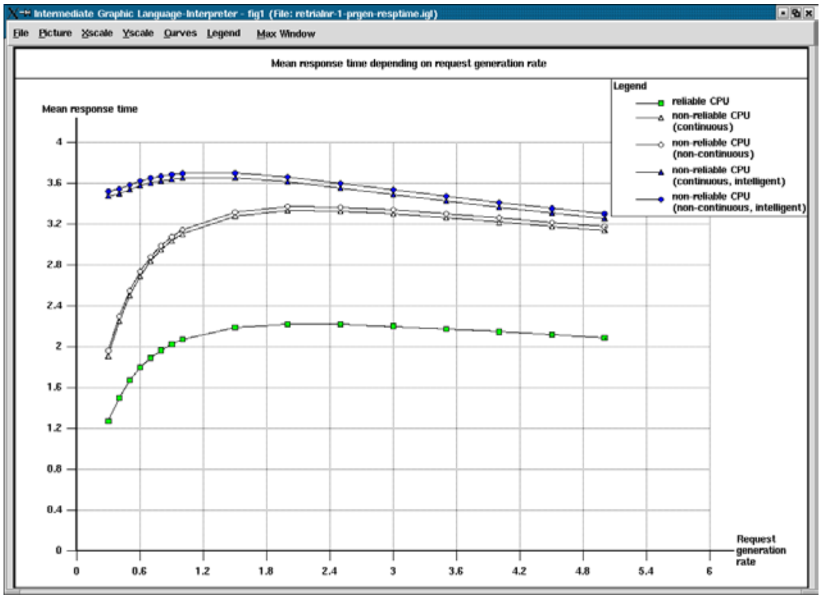

Figure 3.1: $E[T]$ versus primary request generation rate

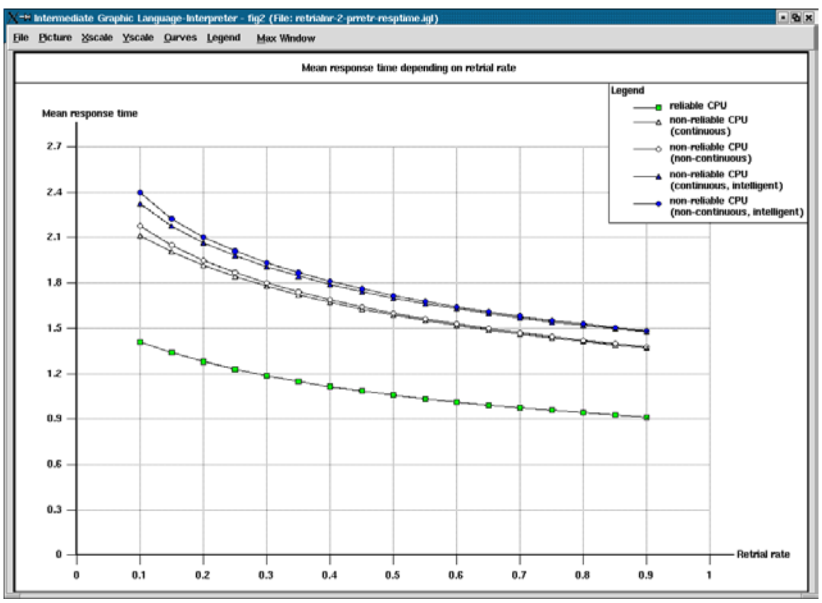

Figure 3.2: $E[T]$ versus retrial rate

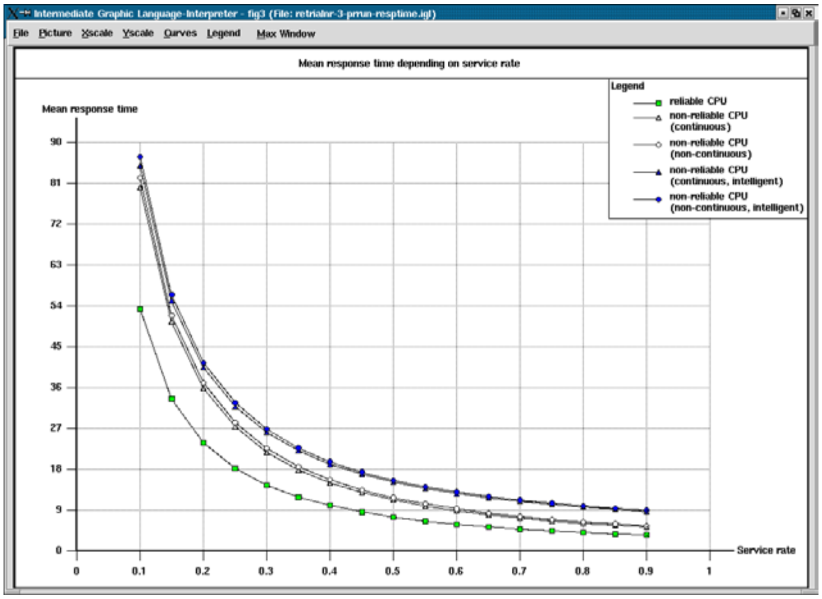

Figure 3.3: $E[T]$ versus service rate

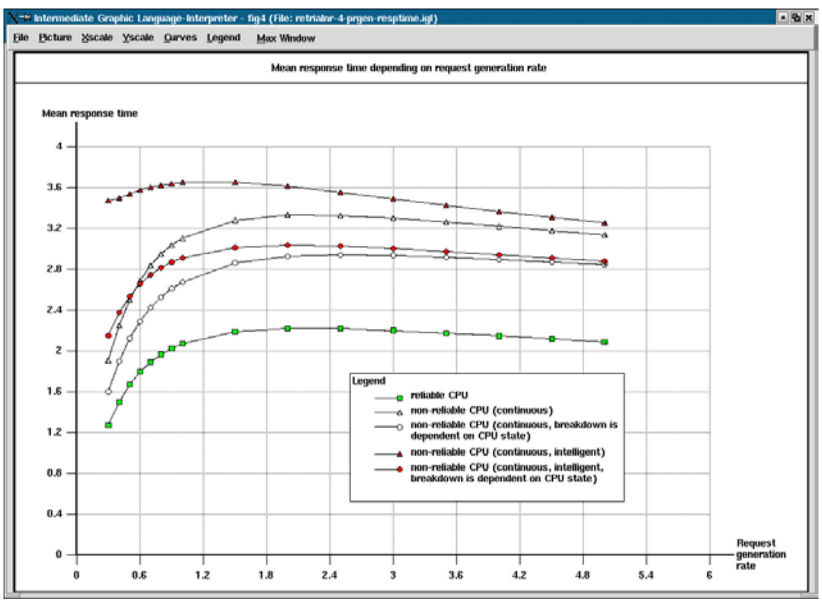

Figure 3.4: $E[T]$ versus primary request generation rate

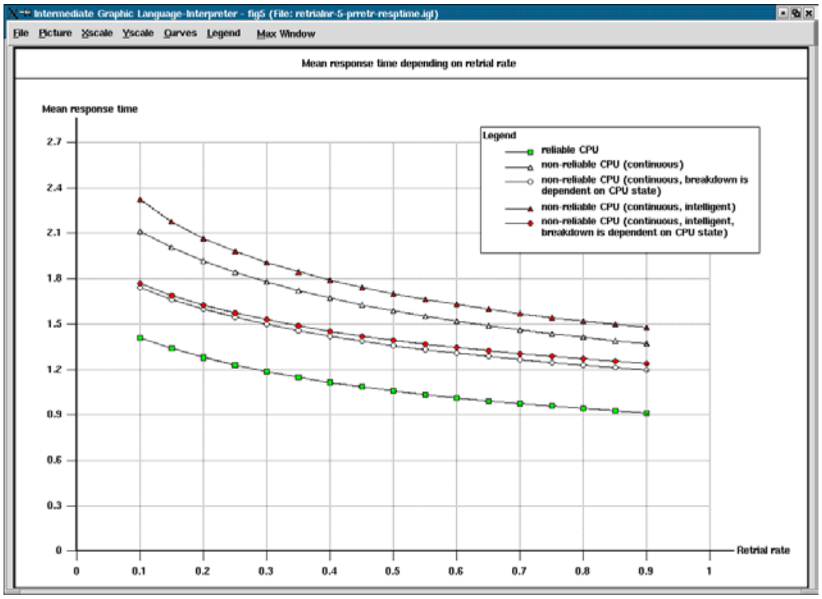

Figure 3.5: $E[T]$ versus retrial rate

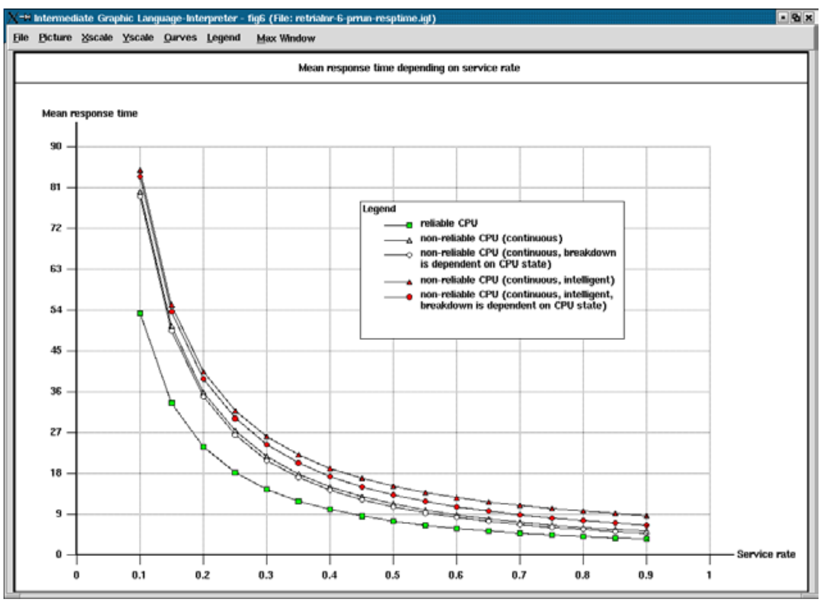

Figure 3.6: $E[T]$ versus service rate

Chapter 4

Asymptotic Methods

4.1 Preliminary results

In this section we give a brief survey of the most related theoretical results due to Anisimov [3, 4], to be applied later on.

Let $(X_\epsilon(k), k \geq 0)$ be a Markov chain with state space

$$\bigcup_{q=0}^{m+1} X_q, \qquad X_i \cap X_j = 0,\ i \neq j, \qquad , i, j = 1, ..., m+1,$$

defined by its transition matrices satisfying the following conditions:

1. $p_\epsilon(i^{(0)}, j^{(0)}) \to p_0(i^{(0)}, j^{(0)})$, as $\epsilon \to 0$,$i^{(0)}, j^{(0)} \in X_0$, and $P_0 = \|p_0(i^{(0)}, j^{(0)})\|$ is irreducible;
2. $p_\epsilon(i^{(q)}, j^{(q+1)}) = \epsilon\alpha^{(q)}(i^{(q)}, j^{(q+1)}) + o(\epsilon), i^{(q)} \in X_q,\ j^{(q+1)} \in X_{q+1}$
3. $p_\epsilon(i^{(q)}, f^{(q)}) \to 0$, as $\epsilon \to 0, i^{(q)}, f^{(q)} \in X_q,\ q \geq 1;$
4. $p_\epsilon(i^{(q)}, f^{(z)}) \equiv 0,\ i^{(q)} \in X_q, f^{(x)} \in X_z,\ z - q \geq 2$

In the sequel the set of states X_q is called the q-th level of the chain, $q = 1, \ldots, m+1$. Let us single out the subset of states

$$\langle \alpha_m \rangle = \bigcup_{q=0}^{m} X_q$$

Denote by
$pi_\epsilon(i^{(q)}, \quad i^{(q)} \in X_q \qquad q = 1, ..., m$ the stationary distribution of a chain with transition matrix

$$\|\frac{p_\epsilon(i^{(q)}, j^{(z)})}{1 - \sum_{k^{(m+1)} \in X_{m+1}} p_\epsilon(i^{(q)}, k^{(m+1)})}\|, i^{(q)} \in X_q, j^{(z)} \in X_z,\ q, z \leq m,$$

Furthermore denote by $g_\epsilon(\langle\alpha_m\rangle)$ the steady state probability of exit from $\langle\alpha_m\rangle$, that is

$$g_\epsilon(\langle\alpha_m\rangle) = \sum_{i^{(m)}\in X_m} \pi_\epsilon(i^{(m)}) \sum_{j^{(m+1)}\in X_{m+1}} p_\epsilon(i^{(m)}, j^{(m+1)})).$$

Denote by
$pi_0(i^{(0)} \quad i^{(0)} \in X_0\}$ the stationary distribution corresponding to P_0 and let

$$\overline{\pi_0} = \{\pi_0(i^{(0)}) \quad i^{(0)} \in X_0\}, \qquad \overline{\pi_\epsilon}^{(q)} = \{\pi_\epsilon(i^{(q)}), \quad i^{(q)} \in X_q\}$$

be row vectors. Finally, let the matrix

$$A^{(q)} = \left\|\alpha^{(q)}(i^{(q)}, j^{(q+1)})\right\|,\ i^{(q)} \in X_q,\ j^{(q+1)} \in X_{q+1}, q = 0, \dots, m$$

defined by Condition 2.

Conditions (1)-(4) enables us to compute the main terms of the asymptotic expression for $\overline{\pi_\epsilon}^{(q)}$ and $g_\epsilon(\langle\alpha_m\rangle)$. Namely, we obtain

$$\begin{aligned} \overline{\pi_\epsilon}^{(q)} &= \epsilon^q \overline{\pi_0} A^{(0)} A^{(1)} \dots A^{(q-1)} + o(\epsilon^q)\ q = 1, \dots, m, \\ g_\epsilon(\langle\alpha_m\rangle) &= \epsilon^{m+1} \overline{\pi}_0 A^{(0)} A^{(1)} \dots A^{(m)} \underline{1} + o(\epsilon^{m+1}), \end{aligned} \tag{4.1}$$

where $\underline{1} = (1, ..., 1)^*$ is a column vector, see Anisimov *et al.* [4] pp. 141-153.

Let $(\eta_\epsilon(t), t \geq 0)$ be a Semi-Markov Process (SMP) given by the embedded Markov chain $(X_\epsilon(k), k \geq 0)$ satisfying conditions (1)-(4). Let the times $\tau_\epsilon(j^{(s)}, k^{(z)})$ – transition times from state $j^{(s)}$ to state $k^{(z)}$ – fulfill the condition

$$\mathbf{E} \exp\{i\Theta\beta_\epsilon\tau_\epsilon(j^{(s)}, k^{(z)})\} = 1 + a_{jk}(s, z, \Theta)\epsilon^{m+1} + o(\epsilon^{m+1}),\ (i^2 = -1)$$

where β_ϵ is some normalizing factor.

Denote by $\Omega_\epsilon(m)$ the instant at which the SMP reaches the $(m+1)$-th level for the first time, exit time from $\langle\alpha_m\rangle$ provided $\eta_\epsilon(0) \in \langle\alpha_m\rangle$. Then we have:

Theorem 4.1.1. *[cf. [4] pp. 153] If the above conditions are satisfied then*

$$\lim_{\epsilon\to 0} \mathbf{E} \exp\{i\Theta\beta_\epsilon\Omega_\epsilon(m)\} = (1 - A(\Theta))^{-1},$$

where

$$A(\Theta) = \frac{\sum\limits_{j^{(0)}, k^{(0)}\in X_0} \pi_0(j^{(0)}) p_0(j^{(0)}, k^{(0)}) a_{jk}(0, 0, \Theta)}{\overline{\pi}_0 A^{(0)} A^{(1)} \dots A^{(m)} \underline{1}}.$$

Corollary 4.1.1. *In particular, if $\alpha_{jk}(s, z, \Theta) = i\Theta m_{jk}(s, z)$ then the limit is an exponentially distributed random variable with mean*

$$\frac{\sum\limits_{j^{(0)}, k^{(0)}\in X_0} \pi_0(j^{(0)}) p_0(j^{(0)}, k^{(0)}) m_{jk}(0, 0)}{\overline{\pi_0} A^{(0)} A^{(1)} \dots A^{(m)} \underline{1}}.$$

4.2 Machine interference problem with a random environment

This section is concerned with a queueing model to analyse the asymptotic behavior of the machine interference problem with N machines and a single operative. The running and repair times of each machine are supposed to be exponentially distributed random variables with parameter depending on the state of a varying environment. Assuming that the repair rate is much greater than the failure rate ("fast" service), it is shown that the time until the number of stopped machines first reaches a certain level converges weakly, under appropriate norming, to an exponentially distributed random variable.

4.2.1 The queuing model

Let us consider the machine interference problem with the following assumptions. There are N machines which are looked after by an operative. The system is supposed to operate in a random environment governed by an ergodic Markov chain $(\xi(t), t \geq 0)$ with state space $(1, \ldots, r)$ and with transition rate matrix $(a_{ij}, i, j = 1, \ldots, r, q_i = -a_{ii} = \sum_{j \neq i} a_{jj}$.
Whenever the environmental process is in state i, the probability that an operating machine breaks down in the time interval $(t, t+h)$ is $\lambda(i)h + o(h)$. A stopped machine is immediately repaired unless the operative is busy, otherwise it joins the queue of failed machines. Whenever the environmental process is in state i, the probability that the repairman completes the service in the time interval $(t, t+h)$ is $\mu(i, \epsilon)h + o(h)$. All random variables involved here and the random environment are supposed to be independent of each other.
Let us consider the system under the assumption of 'fast' repair, that is, $\mu(i, \epsilon) \to \infty$ as $\epsilon \to 0$. For simplicity let $\mu(i, \epsilon) = \mu(i)/\epsilon$.

Denote by $Y_\epsilon(t)$ the number of stopped machines at time t, and let

$$\Omega_\epsilon(m) = \inf\{t : t > 0, Y_\epsilon(t) = m + 1 | Y_\epsilon(0) \leq m\},$$

that is, the instant at which the number of failed machines reaches the $(m + 1)$-th level for the first time, provided that at the beginning their number is not greater than $m; m = 1, \ldots, N - 1$.
Denote by $(\pi_k, k = 1, \ldots, r)$ the steady-state distribution of the governing Markov chain $(\xi(t), t \geq 0)$. Now we have:

Theorem 4.2.1. *For the system in question under the above assumptions, independently of the initial state, the distribution of the normalized random variable $\epsilon^m \Omega_\epsilon(m)$ converges weakly to an exponentially distributed random variable with parameter*

$$\Lambda = (m + 1)! \binom{N}{m + 1} \sum_{i=1}^{r} \pi_i \frac{\lambda(i)^{m+1}}{\mu(i)^m}.$$

Proof. It is easy to see that the process $Z_\epsilon(t) = (\xi(t), Y_\epsilon(t))$ is a two-dimensional Markov chain with state space $E = ((i,s), i = 1, \dots, r, s = 0, \dots, N)$. Furthermore let

$$\langle \alpha_m \rangle = ((i,s), i = 1, \dots, r, s = 0, \dots, m).$$

Hence our aim is to determine the distribution of the first exit time of $Z_\epsilon(t)$ from $\langle \alpha_m \rangle$, provided that $Z_\epsilon(0) \in \langle \alpha_m \rangle$. It can easily be verified that the transition probabilities in any time interval $(t, t+h)$ are the following:

$$(i,s) \longrightarrow^h \begin{cases} (j,s) & a_{ij}h + o(h), & i \neq j \\ (i,s+1) & (N-s)\lambda(i)h + o(h), & s = 0, \dots, N-1 \\ (i,s-1) & (\mu(i)/\epsilon)h + o(h), & s = 1, \dots, N. \end{cases}$$

In addition, the sojourn time $\tau_\epsilon(i,s)$ of $Z_\epsilon(t)$ in state (i,s) is exponentially distributed with parameter $a_i i + (N-s)\lambda(i) + \mu(i)/\epsilon$. Thus, the transition probabilities for the embedded Markov chain are

$$p_\epsilon[(i,0),(j,0)] = \frac{a_{ij}}{q_i + N\lambda(i)},$$
$$p_\epsilon[(i,s),(j,s)] = \frac{a_{ij}}{q_i + (N-s)\lambda(i) + \mu(i)/\epsilon},\ s = 1, \dots, N,$$
$$p_\epsilon[(i,0),(i,1)] = \frac{N\lambda(i)}{q_i + N\lambda(i)},$$
$$p_\epsilon[(i,s),(i,s+1)] = \frac{(N-s)\lambda(i)}{q_i + (N-s)\lambda(i) + \mu(i)/\epsilon},\ s = 0, \dots, N-1$$
$$p_\epsilon[(i,s),(i,s-1)] = \frac{\mu(i)/\epsilon}{q_i + (N-s)\lambda(i) + \mu(i)/\epsilon},\ s = 1, \dots, N.$$

As $\epsilon \to 0$ this implies

$$p_\epsilon[(i,0),(j,0)] = \frac{a_{ij}}{q_i + N\lambda(i)},$$
$$p_\epsilon[(i,s),(j,s)] = o(1), s = 1, \dots, N,$$
$$p_\epsilon[(i,0),(i,1)] = \frac{N\lambda(i)}{q_i + N\lambda(i)},$$
$$p_\epsilon[(i,s),(i,s+1)] = \frac{(N-s)\lambda(i)\epsilon}{\mu(i)}(1 + o(\epsilon)),\ s = 1, \dots, N-1,$$
$$p_\epsilon[(i,s),(i,s-1)] \to 1, s = 1, \dots, N.$$

This agrees with the conditions (1)-(4), but here the zero level is the set $((i,0),(i,1), i = 1, \dots, r)$ while the q-th level is $((i,q+1), i = 1, \dots, r)$. Since the level 0 in the limit forms an essential class, the probabilities $\pi_0(i,0), \pi_0(i,1), i = 1, \dots, r$, satisfy the following system of equations

$$\pi_0(j,0) = \frac{\sum_{i \neq j} \pi_0(i,0) a_{ij}}{q_i + N\lambda(i)} + \pi_0(j,1), \tag{4.2}$$

$$\pi_0(j,1) = \frac{\pi_0(j,0)N\lambda(j)}{a_{jj} + N\lambda(j)}. \tag{4.3}$$

By substituting (4.3) to (4.2) we get

$$\frac{\pi_0(j,0)a_{jj}}{a_{jj} + N\lambda(j)} = \frac{\sum_{i\neq j}\pi_0(i,0)a_{ij}}{q_i + N\lambda(i)}. \tag{4.4}$$

Since $\pi_j a_{jj} = \sum_{i\neq j}\pi_i a_{ij}$, from (4.3) and (4.4) we have

$$\pi_0(i,0) = B\pi_i(q_i + N\lambda(i)), \qquad \pi_0(i,1) = B\pi_i N\lambda(i)$$

where B is the normalizing constant, i.e. $1/B = \sum_{i=1}^r \pi_i[q_i + 2N\lambda(i)]$.

Then it is easy to see that the probability of exit from $\langle\alpha_m\rangle$ is

$$\begin{aligned} g_\epsilon(\langle\alpha\rangle) &= \epsilon^m NB\sum_{i=1}^r \pi_i\lambda(i)\prod_{s=1}^m \frac{(N-s)\lambda(i)}{\mu(i)}(1+o(1)) \\ &= \epsilon^m B(m+1)!\binom{N}{m+1}\sum_{i=1}^r \pi_i\frac{\lambda(i)^{m+1}}{\mu(i)^m}(1+o(1)). \end{aligned} \tag{4.5}$$

Taking into account the exponentiality of $\tau_\epsilon(j,s)$ for fixed Θ,we have

$$\begin{aligned} E\exp\{i\epsilon^m\Theta\tau_\epsilon(j,0)\} &= 1 + \epsilon^m\frac{i\Theta}{a_{jj} + N\lambda(j)}(1+o(1)) \\ E\exp\{i\epsilon^m\Theta\tau_\epsilon(j,s)\} &= 1 + o(\epsilon^m), s > 0. \end{aligned}$$

Notice that $\beta_\epsilon = \epsilon^m$ and therefore from Corollary 1 we immediately get the statement that $\epsilon^m\Omega_\epsilon(m)$ converges weakly to an exponentially distributed random variable with parameter

$$\Lambda = (m+1)\binom{N}{m=1}\sum_{i=1}^r \pi_i\frac{\lambda(i)^{m+1}}{\mu(i)^m},$$

which completes the proof. □

Consequently, the asymptotic distribution of $\Omega_\epsilon(m)$ can be determined as follows:

$$P(\Omega_\epsilon(m) > t) = P(\epsilon^m\Omega_\epsilon(m) > \epsilon^m t) \approx \exp(-\epsilon^m\Lambda t),$$

that is, $\Omega_\epsilon(m)$ is asymptotically an exponentially distributed random variable with parameter

$$\epsilon^m(m+1)!\binom{N}{m+1}\sum_{i=1}^r \pi_i\frac{\lambda(i)^{m+1}}{\mu(i)^m} = (m+1)!\binom{N}{m+1}\sum_{i=1}^r \pi_i\frac{\lambda(i)^{m+1}}{(\mu(i)/\epsilon)^m}.$$

In particular, for $m = N-1$, that is, when all machines are stopped we have

$$\epsilon^{N-1}\Lambda = N!\sum_{i=1}^r \pi_i\frac{\lambda(i)^N}{(\mu(i)\epsilon)^{N-1}}. \tag{4.6}$$

Hence the steady-state probability Q_W that at least one machine works is

$$Q_W = \frac{\frac{1}{\epsilon^{N-1}\Lambda}}{\frac{1}{\epsilon^{N-1}\Lambda} + \sum_{i=1}^{r} \pi_i \frac{1}{\mu(i)/\epsilon}} = \frac{1}{1 + N!\left(\sum_{i=1}^{r} \pi_i \frac{\lambda(i)^N}{\mu(i)^{N-1}}\right)\left(\sum_{i=1}^{r} \pi_i \frac{1}{\mu(i)/\epsilon}\right)} \tag{4.7}$$

In the case when there is no random environment we get

$$Q_W = \frac{1}{1 + N!\left(\frac{\lambda}{\mu/\epsilon}\right)^N}. \tag{4.8}$$

As a conclusion we can see how simple formulas can be obtained in the case of ”fast repair”. The advantage of this approach is that even very large state space the asymptotic parameter can be obtained thus the explosion problem can be avoided.

Bibliography

[1] ALLEN, A. O. *Probability, statistics, and queueing theory. With computer science applications. 2nd ed.* Academic Press, Inc., Boston, MA, 1990.

[2] ALMÁSI, B. ROSZIK, J., AND SZTRIK, J. Homogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Computers and Mathematics wit Applications (accepted).*

[3] ANISIMOV, V. Limit distributions of functionals of a semi-Markov process defined on a fixed state space until the first exit time,. *Soviet Mathamatical Doklads. 11* (1970), 1002–1004.

[4] ANISIMOV, V., ZAKUSILO, O., AND DONCHENKO, V. *Elements of queueing theory and asymptotic analysis of systems*. Visha Skola, KIev, 1987.

[5] ARTALEJO, J. Retrial queues with a finite number of sources. *Journal of Korean Mathematical Society. 35* (1998).

[6] BEGAIN, K., BOLCH, G., AND HEROLD, H. *Practical Perfromance Modeling, Application of the MOSEL Language*. Wiley & Sons, New York, 2001.

[7] BOLCH, G., GREINER, S., DE MEER, H., AND TRIVEDI, K. *Queueing Networks and Markov Chains*. Wiley & Sons, New York, 1998.

[8] BUNDAY, B. D., AND SCRATON, R. E. The G/M/r machine interference model. *European Journal of Operational Research, Vol. 4, :399-402* (1980).

[9] CARMICHAEL, D. *Engineering Queues in Contruction and Mining*. Ellis Horwood, Chichester, 1987.

[10] COOPER, R. B. *Introduction to queueing theory. 2nd ed.* Edward Arnold, London, 1981.

[11] DSHALALOW, J. H. *Advances in queueing. Theory, methods, and open problems.* CRC Press., Boca Raton, 1995.

[12] DSHALALOW, J. H. *Frontiers in Queueing : Models and Applications in Science and Engineering*. CRC Press., Boca Raton, 1997.

[13] FALIN, G. A survey of retrial queues. *Queueing Systems* (1990).

[14] FALIN, G., AND TEMPLETON, J. *Retrial queues*. Chapman and Hall, London, 1997.

[15] FALIN, G.I.AND ARTALEJO, J. A finite source retrial queue. *European Journal of Operational Research* (1998).

[16] GERMAN, R. *Performance analysis of communication systems. Modeling with non-Markovian stochastic Petri nets.* John Wiley & Sons, Chichester, 2000.

[17] GNEDENKO, B., AND KOVALENKO, I. *Introduction to queueing theory.* Birkhaeuser, Boston, MA, 1991.

[18] GROSS, D., AND HARRIS, C. M. *Fundamentals of queueing theory. 2nd ed.*. John Wiley & Sons., New York, 1985.

[19] GUPTA, U., AND RAO, T. V. A recursive method to compute the steady state probabilities of the machine interference model: M/G/1/K. *Computers and Operations Research. 21*, 6 (1994), 597–605.

[20] HAVERKORT, B. *Performance of Computer Communication Systems: A Model-Based Approach*. Wiley & Sons, New York, 1998.

[21] HAVERKORT, B., MARIE, R., RUBINO, G., AND TRIVEDI, K. *Performability Modelling: Techniques and Tools*. Wiley & Sons, New York, 2001.

[22] JAISWAL, N. *Priority Queues*. Academic Press, New York, 1969.

[23] KLEINROCK, L. *Queueing systems. Vol. I. Theory*. John Wiley & Sons, New York, 1975.

[24] KLEINROCK, L. *Queueing systems. Vol. II: Computer applications.* John Wiley & Sons, New York, 1976.

[25] KOBAYASHI, H. *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*. Addison-Wesley, Reading, MA, 1978.

[26] KOVALENKO, I. Rare events in queueing systems, A survey, . *Queueing Systems 16* (1994), 1–49.

[27] LINDEMANN, C. *Stochastic Modeling using DSPNexpress*. Oldenburg, Munich, 1994.

[28] NELSON, R. *Probability, stochastic processes, and queueing theory. The mathematics of computer performance modeling.* Springer-Verlag, New York, 1995.

[29] SAHNER, R., TRIVEDI, K., AND PULIAFITO, A. *Performance and Reliability Analysis of Computer Systems – An Example-Based Approach Using the SHARPE Software Package*. Kluwer Academic Publisher, Boston, M.A., 1996.

[30] STEWART, W. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton, 1995.

[31] SZTRIK, J. http://irh.inf.unideb.hu/user/jsztrik/research/fsqreview.pdf .

[32] SZTRIK, J. On the machine interference. *Publicationes Mathematicae. 30* (1983), 165–175.

[33] SZTRIK, J. Probability model for non-homogeneous multiprogramming computer system. *Acta Cybernetica 6* (1983), 93–101.

[34] SZTRIK, J. A queueing model for multiprogrammed computer systems with different I/O times. *Acta Cybernetica 7* (1985), 127–135.

[35] SZTRIK, J. On the finite-source $\vec{G}/M/r$ queue. *European Journal of Operational Research. 20* (1985), 261–268.

[36] SZTRIK, J. A finite-source queueing model for some manufacturing processes. *Problems of Control and Information Theory. 16,* 6 (1987), 449–457.

[37] SZTRIK, J. Investigation of stationary characteristics of controlled systems $\vec{G}/M/r$ with finite source. *Serdica 14,* 2 (1988), 179–184.

[38] SZTRIK, J. Some contributions to the machine interference problem with heterogeneous machines. *Journal of Information Processing and Cybernetics. 24,* 3 (1988), 137–143.

[39] SZTRIK, J. The $\vec{G}/M/1$ machine-interference model with state-dependent speeds. *Journal of Operational Research Society. 39,* 2 (1988), 201–207.

[40] TAKÁCS, L. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.

[41] TAKAGI, H. *Queueing analysis. A foundation of performance evaluation. Volume 1: Vacation and priority systems, part 1.* North-Holland, Amsterdam, 1991.

[42] TAKAGI, H. *Queueing analysis. A foundation of performance evaluation. Volume 2: Finite Systems*. North-Holland, Amsterdam, 1993.

[43] TAKAGI, H. *Queueing analysis. A foundation of performance evaluation. Volume 3: Discrete-Time Systems*. North-Holland, Amsterdam, 1993.

[44] TIJMS, H. *Stochastic Modelling and Analysis: A Computational Approach*. Wiley & Sons, New York, 1986.

[45] TOMKÓ, J. On sojourn times for semi-Markov processes. *14th European Meeting of Statisticians Poland* (1981), 399–402.

[46] TRIVEDI, K. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. Prentice-Hall, Englewood Cliffs, N.J., 1982.

[47] VAN HOORN, M. *Algorithms and approximations for queueing systems.* Centrum voor Wiskunde en Informatica, Amsterdam, 1984.