Basic Queueing Theory

Dr. János Sztrik

University of Debrecen, Faculty of Informatics

Reviewers:

Dr. József Bíró
Doctor of the Hungarian Academy of Sciences, Full Professor
Budapest University of Technology and Economics

Dr. Zalán Heszberger
PhD, Associate Professor
Budapest University of Technology and Economics

This book is dedicated to my wife without whom this work could have been finished much earlier.

- If anything can go wrong, it will.
- If you change queues, the one you have left will start to move faster than the one you are in now.
- Your queue always goes the slowest.
- Whatever queue you join, no matter how short it looks, it will always take the longest for you to get served.

(Murphy' Laws on reliability and queueing)

Contents

Preface

Modern information technologies require innovations that are based on modeling, analyzing, designing and finally implementing new systems. The whole developing process assumes a well-organized team work of experts including engineers, computer scientists, mathematicians, physicist just to mention some of them. Modern infocommunication networks are one of the most complex systems where the reliability and efficiency of the components play a very important role. For the better understanding of the dynamic behavior of the involved processes one have to deal with constructions of mathematical models which describe the stochastic service of randomly arriving requests. Queueing Theory is one of the most commonly used mathematical tool for the performance evaluation of such systems.

The aim of the book is to present the basic methods, approaches mainly in a Markovian level for the analysis of not too complicated systems. The main purpose is to understand how models could be constructed and how to analyze them. It is assumed the reader has been exposed to a first course in probability theory, however in the text I give a refresher and state the most important principles I need later on. My intention is to show what is behind the formulas and how we can derive formulas. It is also essential to know which kind of questions are reasonable and then how to answer them.

My experience and advice are that if it is possible solve the same problem in different ways and compare the results. Sometimes very nice closed-form, analytic solutions are obtained but the main problem is that we cannot compute them for higher values of the involved variables. In this case the algorithmic or asymptotic approaches could be very useful. My intention is to find the balance between the mathematical and practitioner needs. I feel that a satisfactory middle ground has been established for understanding and applying these tools to practical systems. I hope that after understanding this book the reader will be able to create his owns formulas if needed.

It should be underlined that most of the models are based on the assumption that the involved random variables are exponentially distributed and independent of each other. We must confess that this assumption is artificial since in practice the exponential distribution is not so frequent. However, the mathematical models based on the memoryless property of the exponential distribution greatly simplifies the solution methods resulting in computable formulas. By using these relatively simple formulas one can easily foresee the effect of a given parameter on the performance measure and hence the trends can be forecast. Clearly, instead of the exponential distribution one can use other distributions but in that case the mathematical models will be much more complicated. The analytic

results can help us in checking the results obtained by stochastic simulation. This approach is quite general when analytic expressions cannot be expected. In this case not only the model construction but also the statistical analysis of the output is important.

The primary purpose of the book is to show how to create simple models for practical problems that is why the general theory of stochastic processes is omitted. It uses only the most important concepts and sometimes states theorem without proofs, but each time the related references are cited.

I must confess that the style of the following books greatly influenced me, even if they are in different level and more comprehensive than this material: Allen [3], Gross, Shortle, Thompson and Harris [40], Harchol-Balter [46], Jain [55], Kleinrock [62], Kobayashi and Mark [65], Medhi [75], Nelson [77], Stewart [97], Tijms [117], Trivedi [120].

This book is intended not only for students of computer science, engineering, operation research, mathematics but also those who study at business, management and planning departments, too. It covers more than one semester and has been tested by graduate students at Debrecen University over the years. It gives a very detailed analysis of the involved queueing systems by giving density function, distribution function, generating function, Laplace-transform, respectively. Furthermore, a software package called **QSA** (Queueing Systems Assistance) developed in 2021 is provided to calculate and visualize the main performance measures. In addition, it helps to minimize a quite general mean total cost per unit time with linear objective function. The main advantage that these scripts can be run in all modern devices including smart phones, too, thus the application is very convenient for students and improve the efficiency of a teacher.

I have attempted to provide examples for the better understanding and a collection of exercises with detailed solution helps the reader in deepening her/his knowledge. I am convinced that the book covers the basic topics in stochastic modeling of practical problems and it supports students in all over the world.

I am indebted to Professors József Bíró and Zalán Heszberger for their review, comments and suggestions which greatly improved the quality of the book. I am also very grateful to Tamás Török, Zoltán Nagy, Ferenc Veres, and Hamza Nemouchi for their help in LaTex editing.

All comments and suggestions are welcome at:

`mailto:sztrik.janos@inf.unideb.hu`
`http://irh.inf.unideb.hu/~jsztrik`

Debrecen, 2012, 2021

János Sztrik

Chapter 1

Fundamental Concepts of Queueing Theory

Queueing theory deals with one of the most unpleasant experiences of life, waiting. Queueing is quite common in many fields, for example, in telephone exchange, in a supermarket, at a petrol station, at computer systems, etc. I have mentioned the telephone exchange first because the first problems of queueing theory was raised by calls and Erlang was the first who treated congestion problems in the beginning of 20th century, see Erlang [29,30].

His works inspired engineers, mathematicians to deal with queueing problems using probabilistic methods. Queueing theory became a field of applied probability and many of its results have been used in operations research, computer science, telecommunication, traffic engineering, reliability theory, just to mention some. It should be emphasized that is a living branch of science where the experts publish a lot of papers and books. The easiest way is to verify this statement one should use the Google Scholar for queueing related items. A Queueing Theory Homepage has been created where readers are informed about relevant sources, for example books, softwares, conferences, journals, etc. I highly recommend to visit it at

`http://web2.uwindsor.ca/math/hlynka/queue.html`

There is only a few books and lectures notes published in Hungarian language, I would mention the work of Györfi and Páli [41], Jereb and Telek [57], Kleinrock [62], Lakatos and Szeidl , Telek [69] and Sztrik [104–107]. However, it should be noted that the Hungarian engineers and mathematicians have effectively contributed to the research and applications. First of all we have to mention Lajos Takács who wrote his pioneer and famous book about queueing theory [114]. Other researchers are J. Tomkó, M. Arató, L. Györfi, A. Benczúr, L. Lakatos, L. Szeidl, L. Jereb, M. Telek, J. Bíró, T. Do, and J. Sztrik. The Library of Faculty of Informatics, University of Debrecen, Hungary offer a valuable collection of queueing and performance modeling related books in English, and Russian, too. Please visit:

`https://irh.inf.unideb.hu/user/jsztrik/education/05/3f.html`

I may draw your attention to the books of Takagi [111–113] where a rich collection of references is provided.

1.1 Performance Measures of Queueing Systems

To characterize a queueing system we have to identify the probabilistic properties of the incoming flow of requests, service times and service disciplines. The arrival process can be characterized by the distribution of the **interarrival times** of the customers, denoted by $A(t)$, that is

$$A(t) = P(\text{ interarrival time} < t).$$

In queueing theory these interarrival times are usually assumed to be independent and identically distributed random variables. The other random variable is the **service time**, sometimes it is called service request, work. Its distribution function is denoted by $B(x)$, that is

$$B(x) = P(\text{ service time} < x).$$

The service times, and interarrival times are commonly supposed to be independent random variables.

The **structure of service and service discipline** tell us the **number of servers**, the **capacity of the system**, that is the maximum number of customers staying in the system including the ones being under service. The **service discipline** determines the rule according to the next customer is selected. The most commonly used laws are

- FIFO - First In First Out: who comes earlier leaves earlier, FCFS - First Come First Served
- LIFO - Last Come First Out: who comes later leaves earlier, LCFS - Last Come First Served
- RS - Random Service: the customer is selected randomly, SIRO - Service In Random Order
- Priority without Preemption or Head of Line (HOL), Priority with Preemption / Resume or Repeat
- PS - Processor Sharing

The aim of all investigations in queueing theory is to get the main performance measures of the system which are the probabilistic properties (distribution function, density function, mean, variance) of the following random variables: number of customers in the system, number of waiting customers, utilization of the server/s, response time of a customer, waiting time of a customer, idle time of the server, busy time of a server. Of course, the answers heavily depends on the assumptions concerning the distribution of interarrival times, service times, number of servers, capacity and service discipline. It is quite rare, except for elementary or Markovian systems, that the distributions can be computed. Usually their mean or transforms can be calculated.

For simplicity consider first a single-server system Let ϱ, called **traffic intensity**, be defined as

$$\varrho = \frac{\text{mean service time}}{\text{mean interarrival time}}.$$

Assuming an infinity population system with arrival intensity λ, which is reciprocal of the mean interarrival time, and let the mean service denote by $1/\mu$. Then we have

$$\varrho = \text{arrival intensity} * \text{mean service time } = \frac{\lambda}{\mu}.$$

If $\varrho > 1$ then the systems is overloaded since the requests arrive faster than as the are served. It shows that more server are needed.

Let $\chi(A)$ denote the characteristic function of event A, that is

$$\chi(A) = \begin{cases} 1 & \text{, if } A \text{ occurs,} \\ 0 & \text{, if } A \text{ does not ,} \end{cases}$$

furthermore let $N(t) = 0$ denote the event that at time T the server is idle, that is no customer in the system. Then the **utilization of the server during time** T is defined by

$$\frac{1}{T} \int_0^T \chi\left(N(t) \neq 0\right) dt \ ,$$

where T is a long interval of time. As $T \to \infty$ we get the **utilization of the server** denoted by U_s and the following relations holds with probability 1

$$U_s = \lim_{T\to\infty} \frac{1}{T} \int_0^T \chi\left(N(t) \neq 0\right) dt = 1 - P_0 = \frac{E\delta}{E\delta + Ei},$$

where P_0 is the steady-state probability that the server is idle $E\delta$, Ei denote the mean busy period, mean idle period of the server, respectively.

This formula is a special case of the relationship valid for continuous-time Markov chains and proved in Tomkó [119].

Theorem 1 *Let $X(t)$ be an ergodic Markov chain, and A is a subset of its state space. Then with probability* 1

$$\lim_{T\to\infty} \frac{1}{T} \left(\int_0^T \chi(X(t) \in A) dt \right) = \sum_{i\in A} P_i = \frac{m(A)}{m(A) + m(\overline{A})},$$

where $m(A)$ and $m(\overline{A})$ denote the mean sojourn time of the chain in A and $\overline{A}$ during a cycle,respectively. The ergodic (stationary, steady-state) distribution of $X(t)$ is denoted by P_i.

In an m-server system the mean number of arrivals to a given server during time T is $\lambda T/m$ given that the arrivals are uniformly distributed over the servers. Thus the utilization of a given server is

$$U_s = \frac{\lambda}{m\mu}.$$

The other important measure of the system is the **throughput of the system** which is defined as the mean number of requests serviced during a time unit. In an m-server system the mean number of completed services is $m\varrho\mu$ and thus

$$\text{throughput } = mU_s\mu = .$$

However, if we consider now the customers for a tagged customer the **waiting and response times** are more important than the measures defined above. Let us define by W_j, T_j the waiting, response time of the jth customer, respectively. Clearly the waiting time is the time a customer spends in the queue waiting for service, and response time is the time a customer spends in the system, that is

$$T_j = W_j + S_j,$$

where S_j denotes its service time. Of course, W_j and T_j are random variables and their mean, denoted by $\overline{W_j}$ and $\overline{T_j}$, are appropriate for measuring the efficiency of the system. It is not easy in general to obtain their distribution function.

Other characteristic of the system is the **queue length, and the number of customers in the system**. Let the random variables $Q(t)$, $N(t)$ denote the number of customers in the queue, in the system at time t, respectively. Clearly, in an m-server system we have

$$Q(t) = \max\{0, N(t) - m\}.$$

The primary aim is to get their distributions, but it is not always possible, many times we have only their mean values or their generating function.

1.2 Kendall's Notation

Before starting the investigations of elementary queueing systems let us introduce a notation originated by **Kendall** to describe a queueing system.
Let us denote a system by

$$A \ / \ B \ / \ m \ / \ K \ / \ n / \ D,$$

where

A: distribution function of the interarrival times,

B: distribution function of the service times,

m: number of servers,

K: capacity of the system, the maximum number of customers in the system including the one being serviced,

n: population size, number of sources of customers,

D: service discipline.

Exponentially distributed random variables are notated by M, meaning Markovian or memoryless. Furthermore, if the population size and the capacity is infinite, the service discipline is FIFO, then they are omitted.

Hence $M/M/1$ denotes a system with Poisson arrivals, exponentially distributed service times and a single server. $M/G/m$ denotes an m-server system with Poisson arrivals and generally distributed service times. $M/M/r/K/n$ stands for a system where the customers arrive from a finite-source with n elements where they stay for an exponentially distributed time, the service times are exponentially distributed, the service is carried out according to the request's arrival by r severs, and the system capacity is K.

David G. Kendall, 1918-2007

1.3 Basic Relations for Birth-Death Processes

Since birth-death processes play a very important role in modeling elementary queueing systems let us consider some useful relationships for them. Clearly, arrivals mean birth and services mean death.

As we have seen earlier the steady-state distribution for birth-death processes can be obtained in a very nice closed-form, that is

$$P_i = \frac{\lambda_0 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} P_0, \quad i = 1, 2, \cdots, \quad {P_0}^{-1} = 1 + \sum_{i=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i}. \tag{1.1}$$

Let us consider the distributions at the moments of arrivals, departures, respectively, because we shall use them later on.

Let N_a, N_d denote the state of the process at the instant of births, deaths, respectively, and let $\Pi_k = P(N_a = k), D_k = P(N_d = k), \quad k = 0, 1, 2, \ldots$ stand for their distributions.

By applying the Bayes's theorem it is easy to see that

$$\Pi_k = \lim_{h\to 0} \frac{(\lambda_k h + o(h))P_k}{\sum_{j=0}^{\infty}(\lambda_j h + o(h))P_j} = \frac{\lambda_k P_k}{\sum_{j=0}^{\infty}\lambda_j P_j}. \tag{1.2}$$

Similarly

$$D_k = \lim_{h\to 0} \frac{(\mu_{k+1} h + o(h))P_{k+1}}{\sum_{j=1}^{\infty}(\mu_j h + o(h))P_j} = \frac{\mu_{k+1} P_{k+1}}{\sum_{j=1}^{\infty}\mu_j P_j}. \tag{1.3}$$

Since $P_{k+1} = \dfrac{\lambda_k}{\mu_{k+1}} P_k, \quad k = 0, 1, \ldots,$ thus

$$D_k = \frac{\lambda_k P_k}{\sum_{i=0}^{\infty}\lambda_i P_i} = \Pi_k, \quad k = 0, 1, \ldots. \tag{1.4}$$

In words, the above relation states that the steady-state distributions at the moments of births and deaths are the same. It should be underlined, that it does not mean that it is equal to the steady-state distribution at a random point as we will see later on.

Further essential observation is that in steady-state the mean birth rate is equal to the mean death rate. This can be seen as follows

$$\overline{\lambda} = \sum_{i=0}^{\infty}\lambda_i P_i = \sum_{i=0}^{\infty}\mu_{i+1} P_{i+1} = \sum_{k=1}^{\infty}\mu_k P_k = \overline{\mu}. \tag{1.5}$$

1.4 Optimal Design of Queueing Systems

The ultimate goal of the modeling is to make optimal decision on a given problem. Queueing theory may help to do that. After obtaining the corresponding formulas one can make the decision. Like the descriptive models in classical queueing theory, optimal design models may be classified according to such parameters as the arrival rate(s), the service rate(s), number of servers, the interarrival time and service time distributions, and the queue discipline(s). In addition, the queueing system under study may be a network with several facilities and/or classes of customers, in which case the nature of the ows of the classes among the various facilities must also be specified. What distinguishes an optimal design model from a traditional descriptive model is the fact that some of the parameters are subject to decision and that this decision is made with explicit attention to economic considerations, with the preferences of the decision maker(s) as a guiding principle. The basic distinctive components of a design model are thus:

- the decision variables
- benefits/rewards and costs
- the objective function

Decision variables may include, for example, the arrival rates, the service rates, number of servers, and the queue disciplines at the various service facilities. Typical benefits and costs include rewards to the customers from being served, waiting costs incurred by the customers while waiting for service, and costs to the facilities for providing the service. These benefits and costs may be brought together in an objective function, which quantifies the implicit trade-offs. For example, increasing the service rate will result in less time spent by the customers waiting (and thus a lower waiting cost), but a higher service cost. Each time we dealt with a linear cost/reward structure, in which the objective is to minimize the expected total cost per unit time in steady state. The objective function is calculated and illustrated without any details. In a design problem, the values of the decision variables, once chosen, cannot vary with time nor in response to changes in the state of the system (e.g., the number of customers present). The decision is made with respect to only one variable.
Let us introduce the following costs and benefits/rewards

- CS - cost of service per server per unit time
- CWS - cost of waiting in the system per customer per unit time
- CI - cost of idleness per server per unit time
- CSR - cost of service rate per server per unit time
- CLC - cost of loss per customer per unit time
- R - reward per entering customer per unit time

Our aim is to minimize the following expected total cost per unit time with objective function

$$\begin{aligned} E(Total\ cost) = &(number\ of\ servers) * CS \\ &+ E(number\ of\ customers\ in\ the\ system) * CW \\ &+ E(number\ of\ idle\ servers) * CI + (number\ of\ servers) * CSR \\ &+ E(arrival\ rate) * P(loss/blocking) * CLC \\ &- E(arrival\ rate)(1 - P(loss/blocking) * R. \end{aligned}$$

It is quite a general cost function and it is calculated numerically by giving the respective costs. Depending on the decision parameter this function is illustrated and the user can determine the optimal value of the parameter and the expected total cost.

There are several books on this type of decision making using queueing formulas. In the past years I found the following sources are very useful, Bhat [9], Gross et. al. [40], Harchol-Balter [46], Hillier and Lieberman [51], Kobayashi and Mark [65], Kulkarni [68], Stidham [98], White [126] in which not only the topic is treated but different software tools support the decision, for example MATLAB, Mathematica, Excel.

1.5 Queueing Software and Collection of Problems with Solutions

To solve practical problems the first step is to identify the appropriate queueing system and then to calculate the performance measures. Of course the level of modeling heavily depends on the assumptions. It is recommended to start with a simple system and then if the results do not fit to the problem continue with a more complicated one. Various software packages help the interested readers in different level. The following links worth a visit

`http://web2.uwindsor.ca/math/hlynka/qsoft.html`

I highly recommend an Excel-based software package called QTSPlus to calculate the main performance measures of basic models. It is associated to the book of Gross, Shortle, Thompson and Harris [40] and can be downloaded here

`http://mason.gmu.edu/~jshortle/fqt5th.html`,
`http://mason.gmu.edu/~jshortle/QtsPlus-4-0.zip`
`ftp://ftp.wiley.com/public/sci_tech_med/queueing_theory/`

For practical oriented teaching courses we have also developed a software package called QSA (**Queueing Systems Assistance**) to calculate and visualize the performance measures together with optimal decisions not only for elementary but more advanced queueing systems as well. It is available at

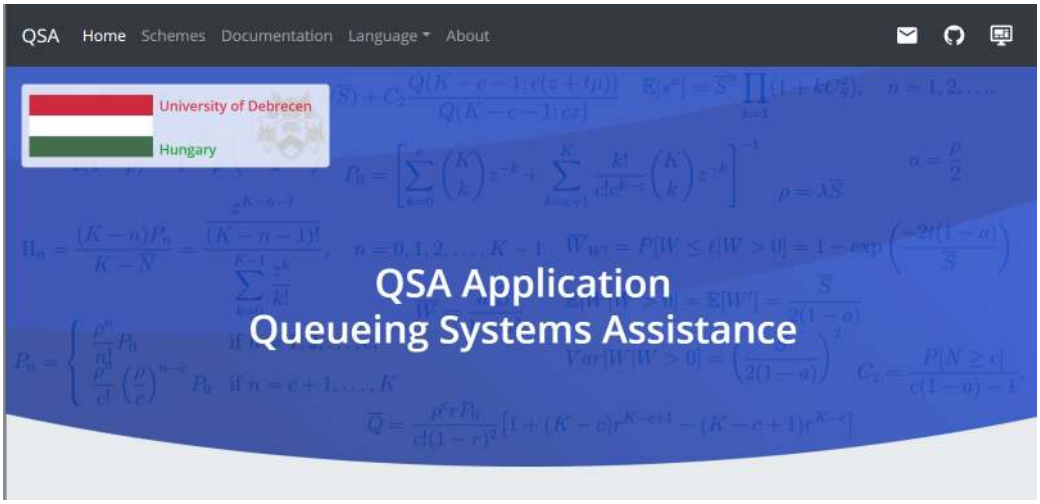
`https://qsa.inf.unideb.hu`

The **main advantages** of QSA over QTSPlus are the following

- It runs on desktops, laptops, smartphones (due to Java)
- It calculates not only the mean but the variance of the corresponding random variables
- It gives the distribution function of the waiting/response times (if possible)
- It visualizes all the main performance measures
- It graphically supports the decision making

Besides the package I have established a **Collection of Problems with Solutions** teaching material in which the problems deliberately listed in random order imitating the practical needs. The material can be downloaded here:

`https://irh.inf.unideb.hu/~jsztrik/education/16/Queueing_Problems_Solutions_2021_Sztrik.pdf`



QSA Welcome page

U_s	0.5	System utilization
$\overline{N}$	1	Average number of customers in the system
var(N)	2	Variance of number of customers in the system
$\overline{Q}$	0.5	Average number of customers in the queue
var(Q)	1.25	Variance of the queue length
$\overline{T}$	1	Mean response time
var(T)	1	Variance of the response time
$\text{var}(T_{LCFS})$	1.25	Variance of the response time, LCFS
$\text{var}(T_{SIRO})$	1.08	Variance of the response time, SIRO
$\overline{W}$	0.5	Average waiting time
var(W)	0.75	Variance of the waiting time

$M/M/1$ System

If the already existing systems are not suitable for your problem then you have to create your own queueing system and then the **creation** starts and the primary aim of the present book is to help this process.

For further readings the interested reader is referred to the following books: Allen [3], Bose [14], Cooper [23], Daigle [25], Gnedenko and Kovalenko [39], Gross, Shortle, Thompson and Harris [40], Harchol-Balter [46], Jain [55], Kleinrock [62], Kobayashi [64, 65], Kulkarni [68], Medhi [75], Nelson [77], Stewart [97], Sztrik [104], Takagi [111–113], Tijms [117], Trivedi [120].

The present book has used some parts of Adan and Reising [1], Allen [3], Daigle [25], Gross and Harris [40], Harchol-Balter [46], Kleinrock [62], Kobayashi [65], Sztrik [104], Tijms [117], Trivedi [120].

Chapter 2

Infinite-Source Queueing Systems

Queueing systems can be classified according to the cardinality of their sources, namely finite-source and infinite-source models. In finite-source models the arrival intensity of the request depends on the state of the system which makes the calculations more complicated. In the case of infinite-source models, the arrivals are independent of the number of customers in the system resulting a mathematically tractable model. In queueing networks each node is a queueing system which can be connected to each other in various way. The main aim of this chapter is to know how these nodes operate.

2.1 The $M/M/1$ Queue

An $M/M/1$ queueing system is the simplest non-trivial queue where the requests arrive according to a Poisson process with rate λ, that is the interarrival times are independent, exponentially distributed random variables with parameter λ. The service times are also assumed to be independent and exponentially distributed with parameter μ. Furthermore, all the involved random variables are supposed to be independent of each other.

Let $N(t)$ denote the number of customers in the system at time t and we shall say that the system is at state k if $N(t) = k$. Since all the involved random variables are exponentially distributed, consequently they have the memoryless property, $N(t)$ is a continuous-time Markov chain with state space $0, 1, \cdots$.

In the next step let us investigate the transition probabilities during time h. It is easy to see that

$$\begin{aligned} P_{k,k+1}(h) = &\left(\lambda h + o(h)\right)\left(1 - \left(\mu h + o(h)\right) + \right. \\ &\left. + \sum_{k=2}^{\infty} \left(\lambda h + o(h)\right)^k \left(\mu h + o(h)\right)^{k-1}, \right. \\ &k = 0, 1, 2, ... \end{aligned}$$

By using the independence assumption the first term is the probability that during h one customer has arrived and no service has been finished. The summation term is the probability that during h at least 2 customers has arrived and at the same time at least 1

has been serviced. It is not difficult to verify the second term is $o(h)$ due to the property of the Poisson process. Thus

$$P_{k,k+1}(h) = \lambda h + o(h).$$

Similarly, the transition probability from state k into state $k-1$ during h can be written as

$$\begin{aligned} P_{k,k-1}(h) &= (\mu h + o(h))\,(1 - (\lambda h + o(h)) + \\ &+ \sum_{k=2}^{\infty} (\lambda h + o(h))^{k-1}\,(\mu h + o(h))^{k} \\ &= \mu h + o(h). \end{aligned}$$

Furthermore, for non-neighboring states we have

$$P_{k,j} = o(h), \qquad \mid k - j \mid \geq 2.$$

In summary, the introduced random process $N(t)$ is a birth-death process with rates

$$\lambda_k = \lambda, \qquad k = 0, 1, 2, ..., \quad \mu_k = \mu, \qquad k = 1, 2, 3....$$

That is all the birth rates are λ, and all the death rates are μ.
As we notated the system capacity is infinite and the service discipline is FIFO.

To get the steady-state distribution let us substitute these rates into formula (1.1) obtained for general birth-death processes. Thus we obtain

$$P_k = P_0 \prod_{i=1}^{k} \frac{\lambda}{\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k, \qquad k \geq 0.$$

By using the normalization condition we can see that this geometric sum is convergent iff $\lambda/\mu < 1$ and

$$P_0 = \left(1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k\right)^{-1} = 1 - \frac{\lambda}{\mu} = 1 - \varrho$$

where $\varrho = \frac{\lambda}{\mu}$. Thus

$$P_k = (1 - \varrho)\varrho^k, \qquad k = 0, 1, 2, ...,$$

which is a modified geometric distribution with success parameter $1 - \varrho$.

In the following we calculate the *the main performance measures of the system*

- *Mean number of customers in the system*

$$\overline{N} = \sum_{k=0}^{\infty} k P_k = (1 - \varrho)\varrho \sum_{k=1}^{\infty} k \varrho^{k-1} =$$

$$= (1-\varrho)\varrho \sum_{k=1}^{\infty} \frac{d\varrho^k}{d\varrho} = (1-\varrho)\varrho \frac{d}{d\varrho}\left(\frac{1}{1-\varrho}\right) = \frac{\varrho}{1-\varrho}.$$

Variance

$$\begin{aligned} Var(N) &= \sum_{k=0}^{\infty}(k-\overline{N})^2 P_k = \sum_{k=0}^{\infty}\left(k - \frac{\varrho}{1-\varrho}\right)^2 P_k \\ &= \sum_{k=0}^{\infty} k^2 P_k + \left(\frac{\varrho}{1-\varrho}\right)^2 - \sum_{k=0}^{\infty} 2k\frac{\varrho}{1-\varrho}P_k \\ &= \sum_{k=0}^{\infty} k(k-1)P_k + \frac{\varrho^2}{(1-\varrho)^2} + \frac{\varrho}{1-\varrho} - 2\left(\frac{\varrho}{1-\varrho}\right)^2 \\ &= (1-\varrho)\varrho^2 \frac{d^2}{d\varrho^2}\sum_{k=0}^{\infty}\varrho^k + \frac{\varrho}{1-\varrho} - \left(\frac{\varrho}{1-\varrho}\right)^2 \\ &= \frac{2\varrho^2}{(1-\varrho)^2} + \frac{\varrho}{1-\varrho} - \left(\frac{\varrho}{1-\varrho}\right)^2 = \frac{\varrho}{(1-\varrho)^2}. \end{aligned}$$

- *Mean number of waiting customers, mean queue length*

$$\overline{Q} = \sum_{k=1}^{\infty}(k-1)P_k = \sum_{k=1}^{\infty} kP_k - \sum_{k=1}^{\infty} P_k = \overline{N} - (1-P_0) = \overline{N} - \varrho = \frac{\varrho^2}{1-\varrho}.$$

Variance

$$Var(Q) = \sum_{k=1}^{\infty}(k-1)^2 P_k - \overline{Q}^2 = \frac{\varrho^2(1+\varrho-\varrho^2)}{(1-\varrho)^2}.$$

- *Server utilization*

$$U_s = 1 - P_0 = \frac{\lambda}{\mu} = \varrho.$$

By using Theorem 1 it is easy to see that

$$P_0 = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + E\delta},$$

where $E\delta$ a *is the mean busy period length of the server*, $\frac{1}{\lambda}$ is the mean idle time of the server. Since the server is idle until a new request arrives which is exponentially distributed with parameter λ. Hence

$$1 - \varrho = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + E\delta},$$

and thus

$$E\delta = \frac{1}{\lambda}\frac{\varrho}{1-\varrho} = \frac{1}{\lambda}\overline{N} = \frac{1}{\mu-\lambda}.$$

In the next few lines we show how this performance measure can be obtained in a different way.
To do so we need the following notations.
Let $\mathbb{E}(\nu_A)$, $\mathbb{E}(\nu_D)$ denote the mean number of customers that have arrived, departed during the mean busy period of the server, respectively. Furthermore, let $\mathbb{E}(\nu_S)$ denote the mean number of customers that have arrived during a mean service time. Clearly

$$\begin{aligned}\mathbb{E}(\nu_D) &= \mathbb{E}(\delta)\mu,\\ \mathbb{E}(\nu_S) &= \frac{\lambda}{\mu},\\ \mathbb{E}(\nu_A) &= \mathbb{E}(\delta)\lambda,\\ \mathbb{E}(\nu_A)+1 &= \mathbb{E}(\nu_D),\end{aligned}$$

and thus after substitution we get

$$\mathbb{E}(\delta) = \frac{1}{\mu-\lambda}.$$

Consequently

$$\begin{aligned}\mathbb{E}(\nu_D) &= \mathbb{E}(\delta)\mu = \frac{1}{1-\varrho}\\ \mathbb{E}(\nu_A) &= \mathbb{E}(\nu_S)\mathbb{E}(\nu_D) = \frac{\lambda}{\mu}\frac{1}{1-\varrho} = \frac{\varrho}{1-\varrho}\\ \mathbb{E}(\nu_A) &= \mathbb{E}(\delta)\lambda = \frac{\varrho}{1-\varrho}.\end{aligned}$$

- *Distribution of the response time of a customer*

  Before investigating the response we show that in any queueing system where the arrivals are Poisson distributed

  $$P_k(t) = \Pi_k(t),$$

  where $P_k(t)$ denotes the probability that at time t the system is a in state k, and $\Pi_k(t)$ denotes the probability that an arriving customers find the system in state k at time t. Let

  $$A(t, t+\Delta t)$$

  denote the event that an arrival occurs in the interval $(t, t+\Delta t)$. Then

  $$\Pi_k(t) := \lim_{\Delta t\to 0} P\left(N(t)=k | A(t,t+\Delta t)\right),$$

  Applying the definition of the conditional probability we have

  $$\Pi_k(t) = \lim_{\Delta t\to 0} \frac{P\left(N(t)=k\ ,\ A(t,t+\Delta t)\right)}{P\left(A(t,t+\Delta t)\right)} =$$

$$= \lim_{\Delta t \to 0} \frac{P\left(A(t, t+\Delta t) | N(t) = k\right) P\left(N(t) = k\right)}{P\left(A(t, t+\Delta t)\right)}.$$

However, in the case of a Poisson process event $A(t, t+\Delta t)$ does not depends on the number of customers in the system at time t and even the time t is irrespective thus we obtain

$$P\left(A(t, t+\Delta t) | N(t) = k\right) = P\left(A(t, t+\Delta t)\right),$$

hence for birth-death processes we have

$$\Pi_k(t) = P\left(N(t) = k\right).$$

That is the probability that an arriving customer find the system in state k is equal to the probability that the system is in state k.

In stationary case applying formula (1.2) with substitutions $\lambda_i = \lambda, \quad i = 0, 1, \ldots$ we have the same result.

If a customer arrives it finds the server idle with probability P_0 hence the waiting time is 0. Assume, upon arrival a tagged customer, the system is in state n. This means that the request has to wait until the residual service time of the customer being serviced plus the service times of the customers in the queue. As we assumed the service is carried out according to the arrivals of the requests. Since the service times are exponentially distributed the remaining service time has the same distribution as the original service time. Hence the waiting time of the tagged customer is Erlang distributed with parameters (n, μ) and the response time is Erlang distributed with $(n+1, \mu)$. Just to remind you the density function of an Erlang distribution with parameters (n, μ) is

$$f_n(x) = \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x}, \qquad x \geq 0.$$

Hence applying the theorem of total probability for the density function of the response time we have

$$f_T(x) = \sum_{n=0}^{\infty} (1-\varrho)\varrho^n \frac{(\mu x)^n}{n!} \mu e^{-\mu x} = \mu(1-\varrho)e^{-\mu x} \sum_{n=0}^{\infty} \frac{(\varrho \mu x)^n}{n!} =$$

$$= \mu(1-\varrho)e^{-\mu(1-\varrho)x}.$$

Its distribution function is

$$F_T(x) = 1 - e^{-\mu(1-\varrho)x}.$$

That is the response time is exponentially distributed with parameter
$\mu(1-\varrho) = \mu - \lambda$.
Hence the expectation and variance of the response time are

$$\overline{T} = \frac{1}{\mu(1-\varrho)}, \quad Var(T) = (\frac{1}{\mu(1-\varrho)})^2.$$

Furthermore

$$\overline{T} = \frac{1}{\mu(1-\varrho)} = \frac{1}{\mu-\lambda} = E\delta.$$

- *Distribution of the waiting time*

Let $f_W(x)$ denote the density function of the waiting time. Similarly to the above considerations for $x > 0$ we have

$$f_W(x) = \sum_{n=1}^{\infty} \frac{(\mu x)^{n-1}}{(n-1)!} \mu e^{-\mu x} \varrho^n (1-\varrho) = (1-\varrho)\varrho\mu \sum_{k=0}^{\infty} \frac{(\mu x \varrho)^k}{k!} e^{-\mu x} =$$

$$= (1-\varrho)\varrho\mu e^{-\mu(1-\varrho)x}.$$

Thus

$$\begin{array}{ll} f_W(0) = 1-\varrho, & \text{if } x = 0, \\ f_W(x) = \varrho(1-\varrho)\mu e^{-\mu(1-\varrho)x}, & \text{if } x > 0. \end{array}$$

Hence

$$F_W(x) = 1 - \varrho + \varrho\left(1 - e^{-\mu(1-\varrho)x}\right) = 1 - \varrho e^{-\mu(1-\varrho)x}.$$

The mean waiting time is

$$\overline{W} = \int_0^{\infty} x f_W(x) dx = \frac{\varrho}{\mu(1-\varrho)} = \varrho E\delta = \overline{N}\frac{1}{\mu}.$$

Since $T = W + S$, in addition W and S are independent we get

$$Var(T) = \frac{1}{(\mu(1-\rho))^2} = Var(W) + \frac{1}{\mu^2},$$

thus

$$Var(W) = \frac{1}{(\mu(1-\rho))^2} - \frac{1}{\mu^2} = \frac{2\rho - \rho^2}{(\mu(1-\rho))^2} = \rho\frac{2}{(\mu(1-\rho))^2} - \frac{\rho^2}{(\mu(1-\rho))^2},$$

that is exactly $\mathbb{E}(W^2) - (\mathbb{E}W)^2$.

Notice that

$$\lambda\overline{T} = \lambda\frac{1}{\mu(1-\varrho)} = \frac{\varrho}{1-\varrho} = \overline{N}. \tag{2.1}$$

Furthermore

$$\lambda\overline{W} = \lambda\frac{\varrho}{\mu(1-\varrho)} = \frac{\varrho^2}{1-\varrho} = \overline{Q}. \tag{2.2}$$

Relations (2.1), (2.2) are called **Little formulas or Little theorem, or Little law** which remain valid under more general conditions.

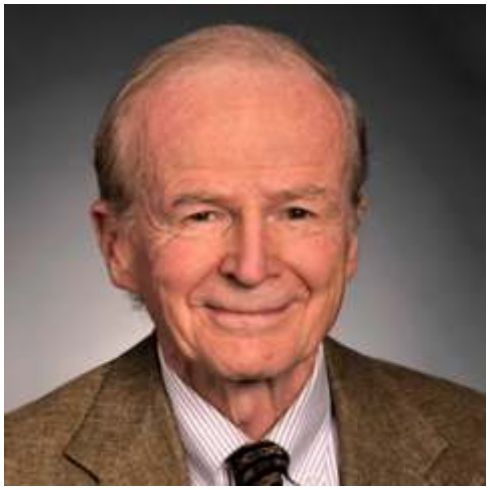


Figure 2.1: John Little, 1928-

It should be noted that in many applications we are dealing with a **required service** denoted by S_R, which does not mean service time. In these cases we involve some kind of capacity (speed, bandwidth) denoted by C. In that case $S_R = CS$ and $E(S) = E(R_S)/C$. It is easy to see if the required service is exponentially distributed with parameter γ than the service time is also exponentially distributed with parameter γC. That is

$$P(S < t) = P(S_R/C < t) = P(S_R < Ct) = 1 - e^{-\gamma Ct}.$$

For example, router A sends 8 packets per second, on the average, to router B. The mean size of a packet is 400 byte (exponentially distributed). The line speed is 64 kbit/s. The utilization of the line (server) is $\rho = 8/s \times 400 \times 8\ bit/(64 \times 1000)\ bit/s = 0.4$.
Or $\rho = \lambda/\mu$, where $\lambda = 8\,packets/s, \quad \mu = 64000\,bit/s/(400{\times}8\,bit/packet) = 20\,packets/s$.
Thus $\lambda/\mu = 8/20 = 0.4$.

Example 1 Economy of Scale
Consider a company that has K terminal rooms. Each terminal room is identical containing as set of terminals/workstations connected by a concentrator to a network. Each set of terminals generates messages to be sent over the concentrator according to a Poisson process with rate λ. Each message requires an exponentially distributed amount of time to be sent by the concentrator with a rate of μ. The company is considering replacing the set of K rooms and K concentrators with one large room and a concentrator that is K times faster.
Comparing **two options***:*

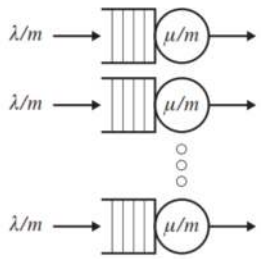
- *K independent rooms*
  Each room can be modeled as multiple $M/M/1$ queues with arrival rate λ and service rate μ. Average delay at any room $E(T) = 1/(\mu - \lambda)$.

- *Single large room*
  It can be modeled as a single $M/M/1$ queue with arrival rate $K\lambda$ and service rate $K\mu$. Average delay at the large room $E(T) = 1/(K\mu - K\lambda)$. That is the combined system is K time faster.

Example 2 Statistical Multiplexing (SM)
There are m independent Poisson data streams, each supplying packet at rate λ/m, arriving at a common concentrator where they are mixed into a single data stream of combined rate λ.
Packet lengths are independent and exponentially distributed with mean transmission time $1/\mu$.
The concentrator can be viewed as $M/M/1$ system which is statistically multiplexes the independent data streams into a single data stream.

Example 3 Time/Frequency Division Multiplexing (TDM/FDM)
In TDM and FDM, transmission capacity is divided equally over m data stream so that each data stream effectively sees a dedicated line with service rate μ/m.
TDM and FDM can be modeled as m $M/M/1$ systems operating in parallel. Each $M/M/1$ queue observes packet arrival rate of λ/m and service rate of μ/m.
It is easy to see that

$$E(T_{TDM}) = mE(T_{SM}).$$



Time/Frequency Division Multiplexing (TDM/FDM)

Question: Why would one ever use FDM?
Answer: Frequency-division multiplexing guarantees a specific service rate to each stream. Statistical multiplexing is unable to provide any such guarantee. More importantly, suppose the original m streams were very regular (i.e., the interarrival times were less variable than Exponential, say closer to Deterministic than Exponential). By merging the streams, we introduce lots of variability into the arrival stream. This leads to problems if an application requires a low variability in delay (e.g., voice or video).

Analysis of the busy period of the server

The system is said to be idle at time t if $N(t) = 0$ and busy at time t if $N(t) > 0$. A busy period begins at any instant in time at which the value of $N(t)$ increases from zero to one and ends at the first instant in time, following entry into a busy period, at which the value of $N(t)$ again reaches zero. An idle period begins when a given busy period ends and ends when the next busy period begins. From the perspective of the server, the $M/M/1$ queueing system alternates between two distinct types of periods: idle periods and busy periods. These types are descriptive; the busy periods are periods during which the server is busy servicing customers, and the idle periods are those during which the

server is not servicing customers. For the ordinary $M/M/1$ queueing system, the server is never idle when there is at least one customer in the system.

Because of the memoryless property of both the exponential distribution and the Poisson process, the length of an idle period is the same as the length of time between two successive arrivals from a Poisson process with parameter λ. The length of a busy period, on the other hand is dependent upon both the arrival and service processes. The busy period begins upon the arrival of its first customer. During the first service another customers arrive and the service and arrival processes continue until there are no longer any remaining customers, and at that point in time the system returns to an idle period. Thus, the length of a busy period is the total amount of time required to service all of the customers of all of the generations of the first customer of the busy period. Consequently, we can think of the busy period as being generated by its first customer. Alternatively, we can view the server as having to work until all of the first customers descendents die out. The distribution of the length of a busy period is of interest in its own right, but an understanding of the behavior of busy period processes is also extremely helpful in understanding waiting time and queue length behavior in both ordinary and priority queueing systems.

Before starting the investigations we need some additional knowledge about the properties of the exponential distribution. Let us see the following proof.

$X \in Exp(\lambda), Y \in Exp(\mu)$ independent, $Z = \min(X, Y)$. Find

$$P(Z < t \mid X < Y), P(Z < t \mid Y < X)$$

Solution:

$$\begin{aligned}
\frac{P(Z < t \mid X < Y)}{P(X < Y)} &= \frac{\int_0^\infty P(Z < t, X < Y) f_y dy}{P(X < Y)} \\
&= \frac{\int_0^t P(X < y)\mu e^{-\mu y} dy + \int_t^\infty P(X < t)\mu e^{-\mu y} dy}{P(X < Y)} \\
&= \frac{\lambda + \mu}{\lambda}\left[\int_0^t (1 - e^{-\lambda y}) \cdot \mu e^{-\mu y} dy + \int_t^\infty (1 - e^{-\lambda t}) \cdot \mu e^{-\mu y} dy\right] \\
&= \frac{\lambda + \mu}{\lambda}\left[(1 - e^{-\mu t}) - \frac{\mu}{\lambda + \mu}(1 - e^{-(\lambda+\mu)t}) + e^{-\mu t}(1 - e^{\lambda t})\right] \\
&= \frac{\lambda + \mu}{\lambda}\left[\frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t}\right] = 1 - e^{-(\lambda+\mu)t}.
\end{aligned}$$

$P(Z < t \mid Y < X) = 1 - e^{-(\lambda+\mu)t}$, can be proved exactly the same way.

Another Proof:

$$\begin{aligned}
P(Z < t \mid X < Y) &= 1 - P(Z > t \mid X < Y) = 1 - \frac{P(X > t, X < Y)}{P(X < Y)} \\
&= 1 - \frac{\int_t^\infty P(Y > x)\lambda e^{-\lambda x} dx}{P(X < Y)} = 1 - \frac{\int_t^\infty e^{-\mu x} \cdot \lambda e^{-\lambda x} dx}{P(X < Y)} \\
&= 1 - \left[\frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t}\right] \cdot \frac{\lambda + \mu}{\lambda} = 1 - e^{-(\lambda+\mu)t}.
\end{aligned}$$

An alternate and instructive way to view busy period process is to separate the busy period into two parts: the part occurring before the first customer arrival after the busy period has started, and the part occurring after the first customer arrival after the busy period has started, if such an arrival occurs. In the latter case we have two customers in the system and easy to see that the length of the busy period does not depend on the order of service. So the server will be idle of all the customers leave the system, that is we have two busy periods initiated by the generic customer and the first customer after the busy period started. These busy period are independent of each other because the arrival and service time are independent of each other.

Due to the memoryless property of the exponential distribution and taking into account the statements concerning of the minimum of independent exponentially distributed random variables it is not so difficult to see that for the Laplace-transform of the busy period δ we have

$$L_\delta(t) = \frac{\mu}{\lambda+\mu} \cdot \frac{\lambda+\mu}{\lambda+\mu+t} + \frac{\lambda}{\lambda+\mu} \cdot \frac{\lambda+\mu}{\lambda+\mu+t}(L_\delta(t))^2$$

$$L_\delta(t) = \frac{\mu}{\lambda+\mu+t} + \frac{\lambda}{\lambda+\mu+t}(L_\delta(t))^2$$

$$\Rightarrow \lambda(L_\delta(t))^2 - (\lambda+\mu+t)L_\delta(t) + \mu = 0$$

$$L_\delta(t) = \frac{\lambda+\mu+t \pm \sqrt{(\lambda+\mu+t)^2 - 4\lambda\mu}}{2\lambda}$$

$L_\delta(0) = 1, \quad$ that is why:

$$L_\delta(t) = \frac{\lambda+\mu+t - \sqrt{(\lambda+\mu+t)^2 - 4\lambda\mu}}{2\lambda} < 1.$$

We are interested in the mean and variance of the busy period, that is why we need

$$L'_\delta(t) = \frac{1}{2\lambda}\left[1 - \frac{1}{2}\left((\lambda+\mu+t)^2 - 4\lambda\mu\right)^{-\frac{1}{2}} .2(\lambda+\mu+t)\right]$$

$$L''_\delta(t) = \frac{1}{2\lambda}\left[\frac{1}{4}\left((\lambda+\mu+t)^2 - 4\lambda\mu\right)^{-\frac{3}{2}} \cdot 4(\lambda+\mu+t)^2 - \frac{1}{2}\left((\lambda+\mu+t)^2 - 4\lambda\mu\right)^{-\frac{1}{2}} \cdot 2\right]$$

$$L'_\delta(0) = \frac{1}{2\lambda}\left(1 - \frac{1}{2(\mu-\lambda)}2(\lambda+\mu)\right) = \frac{1}{2\lambda}\left(1 - \frac{\lambda+\mu}{\mu-\lambda}\right) = -\frac{1}{\mu-\lambda},$$

$$E(\delta) = \frac{1}{\mu-\lambda}.$$

To get the variance we proceed

$$L''_\delta(0) = \frac{1}{2\lambda}\left[\frac{(\lambda+\mu)^2}{(\mu-\lambda)^3} - \frac{1}{\mu-\lambda}\right] = \frac{1}{2\lambda}\frac{(\lambda+\mu)^2-(\mu-\lambda)^2}{(\mu-\lambda)^3} = \frac{1}{2\lambda}\frac{2\mu 2\lambda}{(\mu-\lambda)^3} = \frac{2\mu}{(\mu-\lambda)^3},$$

$$Var(\delta) = \frac{2\mu}{(\mu-\lambda)^3} - \left(\frac{1}{\mu-\lambda}\right)^2 = \frac{2\mu-\mu+\lambda}{(\mu-\lambda)^3} = \frac{\lambda+\mu}{(\mu-\lambda)^3} = \frac{1+\rho}{\mu^2(1-\rho)^3}.$$

Let us see another solution treated in Adan and Reising [1].

Let the random variable T_n be the time till the system is empty again if there are now n customers present in the system. Clearly, T_1 is the length of a busy period, since a busy period starts when the first customer after an idle period arrives and it ends when the system is empty again. The random variables T_n satisfy the following recursion relation. Suppose there are $n(>0)$ customers in the system. Then the next event occurs after an exponential time with parameter $\lambda+\mu$: with probability $\lambda/(\lambda+\mu)$ a new customer arrives, and with probability $\mu/(\lambda+\mu)$ service is completed and a customer leaves the system. Hence, for $n = 1, 2, \ldots$,

$$T_n = Z + \begin{cases} T_{n+1} & \text{with probability } \lambda/(\lambda+\mu), \\ T_{n-1} & \text{with probability } \mu/(\lambda+\mu), \end{cases} \tag{2.3}$$

where Z is an exponential random variable with parameter $\lambda+\mu$. From this relation we get for the Laplace-transform $\widetilde{T}_n(s)$ of T_n that

$$\widetilde{T}_n(s) = \frac{\lambda+\mu}{\lambda+\mu+s}\left(\widetilde{T}_{n+1}(s)\frac{\lambda}{\lambda+\mu} + \widetilde{T}_{n-1}(s)\frac{\mu}{\lambda+\mu}\right),$$

and thus, after rewriting,

$$(\lambda+\mu+s)\widetilde{T}_n(s) = \lambda\widetilde{T}_{n+1}(s) + \mu\widetilde{T}_{n-1}(s), \qquad n = 1, 2, \ldots$$

For *fixed* s this equation is a second order difference equation. Its general solution is

$$\widetilde{T}_n(s) = c_1 x_1^n(s) + c_2 x_2^n(s), \qquad n = 0, 1, 2, \ldots$$

where $x_1(s)$ and $x_2(s)$ are the roots of the quadratic equation

$$(\lambda+\mu+s)x = \lambda x^2 + \mu,$$

satisfying $0 < x_1(s) \le 1 < x_2(s)$. Since $0 \le \widetilde{T}_n(s) \le 1$ it follows that $c_2 = 0$. The coefficient c_1 follows from the fact that $T_0 = 0$ and hence $\widetilde{T}_0(s) = 1$, yielding $c_1 = 1$. Hence we obtain

$$\widetilde{T}_n(s) = x_1^n(s),$$

and in particular, for the Laplace-transform $L_\delta(s)$ of the busy period δ, we find

$$L_\delta(s) = \widetilde{T}_1(s) = x_1(s) = \frac{1}{2\lambda}\left(\lambda+\mu+s - \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}\right).$$

By inverting this transform we get for the density $f_\delta(t)$ of δ,

$$f_\delta(t) = \frac{1}{t\sqrt{\rho}} e^{-(\lambda+\mu)t} I_1(2t\sqrt{\lambda\mu}), \qquad t > 0,$$

where $I_1(\cdot)$ denotes the modified Bessel function of the first kind of order one, i.e.

$$I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!}.$$

As we will see later on for an $M/G/1$ system we have

$$L_\delta(t) = L_S(t + \lambda - \lambda L_\delta(t)), \tag{2.4}$$

where $L_S(t)$ denotes the Laplace-transform of the service time. For exponentially distributed service time we have

$$L_\delta(t) = \frac{\mu}{\mu + t + \lambda - \lambda L_\delta(t)}$$

from we we get the same equation as before, that is

$$\lambda(L_\delta(t))^2 - (\lambda + \mu + t)L_\delta(t) + \mu = 0.$$

Distribution of number of customers served during the busy period

Let $N_d(\delta)$ denote the number of departed/served customers during a busy period and let $G(z) = G_{N_d(\delta)}(z)$ its generating function.
Then similarly to above considerations it is not difficult to get

$$G(z) = z\frac{\mu}{\lambda+\mu} + G^2(z)\frac{\lambda}{\lambda+\mu} \Rightarrow \lambda G^2(z) - (\lambda+\mu)G(z) + z\mu = 0.$$

$$G(z) = \frac{\lambda + \mu \pm \sqrt{(\lambda+\mu)^2 - 4\lambda\mu z}}{2\lambda}$$

$$G(z) = \frac{1 + \rho - \sqrt{(1+\rho)^2 - 4\rho z}}{2\rho} = \frac{1+\rho}{2\rho}\left(1 - \sqrt{1 - \frac{4\rho z}{(1+\rho)^2}}\right).$$

$$G(1) = \frac{1 + \rho - \sqrt{(1-\rho)^2 - 4\rho}}{2\rho} = \frac{1 + \rho - 1 + \rho}{2\rho} = 1.$$

The mean and variance can be obtained in the following way

$$G(z) = \frac{1+\rho-\sqrt{(1-\rho)^2-4\rho z}}{2\rho}$$
$$G'(z) = \frac{-\frac{1}{2}\left((1+\rho)^2-4\rho z\right)^{-\frac{1}{2}}(-4\rho)}{2\rho} = \left((1+\rho)^2-4\rho z)\right)^{-\frac{1}{2}}.$$
$$G'(1) = \frac{1}{1-\rho}.$$
$$G''(z) = -\frac{1}{2}\left((1+\rho)^2-4\rho z\right)^{-\frac{3}{2}}(-4\rho)$$
$$G''(1) = \frac{2\rho}{(1-\rho)^3}.$$

Thus

$$E\left(N_d(\delta)\right) = G'(1) = \frac{1}{1-\rho},$$

and the variance is

$$Var\left(N_d(\delta)\right) = \frac{2\rho}{(1-\rho)^3} + \frac{1}{1-\rho} - \frac{1}{(1-\rho)^2} = \frac{2\rho+(1+\rho)^2-(1-\rho)}{(1-\rho)^3} = \frac{\rho+\rho^2}{(1-\rho)^3}.$$

Furthermore, the distribution of $N_d(\delta)$ can be obtained, too

$$G_{N_d(\delta)}(z) = \frac{1+\rho}{\rho}\left(1-\sqrt{1-\frac{4\rho z}{(1+\rho)^2}}\right)$$

$$P\left(N_d(\delta)=n\right) = \frac{1}{n}\binom{2n-2}{n-1}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}, \quad \text{n = 1,2,...}$$

Thus

$$E\left(N_d(\delta)\right) = \sum_{n=1}^{\infty}\binom{2n-2}{n-1}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}},$$

which is very difficult to calculate. It means that the generating function approach is very useful since we proved that

$$E\left(N_d(\delta)\right) = \frac{1}{1-\rho}.$$

As it will see later on for an $M/G/1$ system we have

$$G_{N_d(\delta)}(z) = zL_S(\lambda-\lambda G_{N_d(\delta)}(z)).$$

For exponentially distributed service time we have

$$G_(z) = z\frac{\mu}{\mu+\lambda-\lambda G_(z)}$$

from we we get the same equation as before, that is

$$\lambda(G_{N_d(\delta)}(z))^2 - (\lambda+\mu)G_{N_d(\delta)}(z) + \mu z = 0.$$

Also

$$Var(N_d(\delta)) = \frac{\rho(1-\rho)+\lambda^2 E(S^2)}{(1-\rho)^3} = \frac{\rho(1-\rho)+2\rho^2}{(1-\rho)^3} = \frac{\rho+\rho^2}{(1-\rho)^3}.$$

$M/M/1$ system with non-preemptive LCFS service discipline

In the following we show how the results concerning to the busy period analysis of a FCFS system can be used for the investigation of the waiting and response time of a system with non-preemptive LCFS (Last-Come- First-Served) service order. This means that the last customer does not interrupt the service of the current customer.

Since the service time are exponentially distributed due to the memoryless property the waiting time of the last customer will be the busy period length of the server. Thus for the Laplace-transform, mean and variance we have

$$\begin{aligned}
L_{W_{LCFS}}(t) &= (1-\rho) + \rho L_\delta(t) \\
L_{T_{LCFS}}(t) &= \frac{\mu}{\mu+t}\cdot(1-\rho+\rho L_\delta(t)) \\
E(W_{LCFS}^2) &= \rho E(\delta^2) = \rho\frac{2}{\mu^2(1-\rho)^3} \\
E(W_{LCFS}) &= \rho E(\delta) = \rho\frac{1}{\mu(1-\rho)}
\end{aligned}$$

$$Var(W_{LCFS}) = \rho\frac{2}{\mu^2(1-\rho)^3} - \left(\rho\frac{1}{\mu(1-\rho)}\right)^2 = \frac{2\rho-\rho^2(1-\rho)}{\mu^2(1-\rho)^3} = \frac{2\rho-\rho^2+\rho^3}{\mu^2(1-\rho)^3}.$$

$$Var(T_{LCFS}) = \frac{1}{\mu^2} + \frac{2\rho-\rho^2+\rho^3}{\mu^2(1-\rho)^3}.$$

As we will see later on for an $M/G/1$ system the Laplace-transform, mean, variance can be obtained by the following formula and hence we can check our result for exponentially distributed service time.

$$L_{W_{LCFS}}(t) = (1-\rho) + \rho\frac{1-L_\delta(t)}{(t+\lambda-\lambda L_\delta(t))E(S)},$$

$$L_{T_{LCFS}}(t) = L_{W_{LCFS}}(t)L_S(t),$$

$$
\begin{aligned}
Var(W_{LCFS}) &= \frac{\lambda E(S^3)}{3(1-\rho)^2} + \frac{\lambda^2(1+\rho)(E(S^2))^2}{4(1-\rho)^3} \\
&= \frac{6\lambda}{3\mu^3(1-\rho)^2} + \frac{4\lambda^2(1+\rho)}{4\mu^4(1-\rho)^3} = \frac{2\rho}{\mu^2(1-\rho)^2} + \frac{\rho^2(1+\rho)}{\mu^2(1-\rho)^3} \\
&= \frac{2\rho(1-\rho)+(1+\rho)\rho^2}{\mu^2(1-\rho)^3} = \frac{2\rho-2\rho^2+\rho^2+\rho^3}{\mu^2(1-\rho)^3} = \frac{2\rho-\rho^2+\rho^3}{\mu^2(1-\rho)^3}.
\end{aligned}
$$

Furthermore, it should be noted that the mean waiting and response time of an $M/M/1$ under any well-known service discipline will be the same due to the Little-formula and the fact that the service rate is always μ resulting the same distribution for the steady-state distribution of the number of customers in the system. However, the higher moment will be different depending on the service order. It can be proved that for $M/G/1$ systems we have

$$
Var(W_{SIRO}) = \frac{2\lambda E(S^3)}{3(1-\rho)(2-\rho)} + \frac{\lambda^2(2+\rho)(E(S^2))^2}{4(1-\rho)^2(2-\rho)}
$$

$$
Var(W_{LCFS}) = \frac{\lambda E(S^3)}{3(1-\rho)^2} + \frac{\lambda^2(1+\rho)(E(S^2))^2}{4(1-\rho)^3}
$$

$$
Var(W_{FCFS}) = \frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2(E(S^2))^2}{4(1-\rho)^2}
$$

Comparing the formulas term-by-term it is not difficult to prove that

$$
\begin{aligned}
Var(W_{FCFS}) < Var(W_{SIRO}) < Var(W_{LCFS}), \\
Var(T_{FCFS}) < Var(T_{SIRO}) < Var(T_{LCFS}).
\end{aligned}
$$

Analysis of the output process

Let us examine the states of an $M/M/1$ system at the departure instants of the customers. Our aim is to calculate the distribution of the departure times of the customers. As it was proved in (1.3) at departures the distribution is

$$
D_k = \frac{\lambda_k P_k}{\sum_{i=0}^{\infty} \lambda_i P_i}.
$$

In the case of Poisson arrivals $\lambda_k = \lambda, k = 0, 1, \ldots$, hence $D_k = P_k$.
Now we are able to calculate the Laplace-transform of the interdeparture time d. Conditioning on the state of the server at the departure instants, by using the theorem of total

Laplace-transform we have

$$L_d(s) = \varrho\frac{\mu}{\mu+s} + (1-\varrho)\frac{\lambda}{\lambda+s}\frac{\mu}{\mu+s},$$

since if the server is idle for the next departure a request should arrive first. Hence

$$L_d(s) = \frac{\mu\varrho(\lambda+s)+(1-\varrho)\lambda\mu}{(\lambda+s)(\mu+s)} = \frac{\lambda\mu\varrho+\lambda s+\lambda\mu-\lambda\mu\varrho}{(\lambda+s)(\mu+s)} = \frac{\lambda}{\lambda+s},$$

which shows that the distribution is exponential with parameter λ and not with μ as one might expect. The independence follows from the memoryless property of the exponential distributions and from their independence. This means that the departure process is a Poisson process with rate λ.

This observation is very important to investigate tandem queues, that is when several simple $M/M/1$ queueing systems as nodes are connected in serial to each other. Thus at each node the arrival process is a Poisson process with parameter λ and the nodes operate independently of each other. Hence if the service times have parameter μ_i at the ith node then introducing traffic intensity $\varrho_i = \dfrac{\lambda}{\mu_i}$ all the performance measures for a given node could be calculated. Consequently, the mean number of customers in the network is the sum of the mean number of customers in the nodes. Similarly, the mean waiting and response times for the network can be calculated as the sum of the related measures in the nodes.

Now, let us show how the density function d can be obtained directly without using the Laplace-transforms. By applying the theorem of total probability we have

$$\begin{aligned} f_d(x) &= \varrho\mu e^{-\mu x} + (1-\varrho)\left(\frac{\lambda\mu}{\lambda-\mu}e^{-\mu x} + \frac{\lambda\mu}{\mu-\lambda}e^{-\lambda x}\right) \\ &= \lambda e^{-\mu x} + \frac{\mu-\lambda}{\mu}\left(\frac{\lambda\mu}{\mu-\lambda}e^{-\lambda x} - \frac{\lambda\mu}{\mu-\lambda}e^{-\mu x}\right) \\ &= \lambda e^{-\mu x} + \lambda e^{-\lambda x} - \lambda e^{-\mu x} = \lambda e^{-\lambda x}. \end{aligned}$$

Let us see a more general method that works for any systems with Poisson arrivals and exponentially distributed service times.

We now proceed to verify the input-output identity with a constructive proof that utilizes a simple differential-difference argument (much like that used in the development of the birth-death process), which will show that, indeed, the inter-departure times are exponential with parameter λ.

Consider an $M/M/c/\infty$ system in steady state. Let $N(t)$ now represent the number of customers in the system at a time t after the last departure.
Since we are considering steady state, we have

$$Pr\{N(t) = n\} = p_n. \tag{2.5}$$

Furthermore, let d represent the random variable "time between successive departures" (inter-departure time), and

$$F_n(t) = Pr\{N(t) = n \quad \text{and} \quad d > t\}. \tag{2.6}$$

So $F_n(t)$ is the joint probability that there are n customers in the system at a time t after the last departure and that t is less than the inter-departure time d; that is, another departure has not as yet occurred. The cumulative distribution function of the random variable d, which will be denoted as $D(t)$, is given by

$$D(t) = P\{d \leq t\} = 1 - \sum_{n=0}^{\infty} F_n(t), \tag{2.7}$$

since

$$\sum_{n=0}^{\infty} = Pr\{d > t\} \tag{2.8}$$

is the marginal complementary cumulative distribution function of d. To find $D(t)$, it is necessary to first find $F_n(t)$.

As usual using the law of total probability we can write the following difference equations concerning $F_n(t)$:

$$\begin{aligned}
F_n(t + \Delta t) &= (1 - \lambda\Delta t)(1 - c\mu\Delta t)F_n(t) + \lambda\Delta t(1 - c\mu\Delta t)F_{n-1}(t) \\
&\quad + o(\Delta t), \qquad c \leq n, \\
F_n(t + \Delta t) &= (1 - \lambda\Delta t)(1 - n\mu\Delta t)F_n(t) + \lambda\Delta t(1 - n\mu\Delta t)F_{n-1}(t) \\
&\quad + o(\Delta t), \qquad 1 \leq n \leq c, \\
F_0(t + \Delta) &= (1 - \lambda\Delta t)F_0(t) + o(\Delta t).
\end{aligned}$$

Moving $F_n(t)$ from the right side of each of the above equations, dividing by Δt, and taking the limit as $\Delta \to 0$, we obtain the differential-difference equations as

$$\begin{aligned}
\frac{dF_n(t)}{dt} &= -(\lambda + c\mu)F_n(t) + \lambda F_{n-1}(t) \qquad c \leq n, \\
\frac{dF_n(t)}{dt} &= -(\lambda + n\mu)F_n(t) + \lambda F_{n-1}(t) \qquad 1 \leq n \leq c, \\
\frac{dF_0(t)}{dt} &= -\lambda F_0(t).
\end{aligned}$$

Using the boundary condition

$$F_n(t) \equiv Pr\{N(0) = n \quad \text{and} \quad d > 0\} = Pr\{N(0) = n\} = p_n.$$

Let us consider

$$F_n(t) = p_n e^{-\lambda t}. \tag{2.9}$$

The reader can easily verify that is the solution to the above system of differential equations by substitution, recalling that for $M/M/c/\infty$ models,

$$p_{n+1} = \begin{cases} \frac{\lambda}{(n+1)\mu} p_n, & 1 \leq n < c, \\ \frac{\lambda}{c\mu} p_n, & c \leq n. \end{cases}$$

To obtain $D(t)$, the cumulative distribution function of the inter-departure times, we use 2.9 in 2.7 to get

$$D(t) = 1 - \sum_{n=0}^{\infty} p_n e^{-\lambda t} = 1 - e^{-\lambda t} \sum_{n=0}^{\infty} p_n = 1 - e^{-\lambda t} \tag{2.10}$$

thus showing that the inter-departure times are exponential.

It is easy to see that this statement is valid for any state-dependent service intensities, that is instead of the service intensities of the $M/M/c$ system we can write $\mu_n, n = 0, 1, 2, ...$ Thus, the statement is valid for any $M/M/\infty$ system.

In the following we prove that the random variables $N(d)$ and d are independent and furthermore that successive inter-departure times are independent of each other. This result was first proved by Burke. So we see that the output distribution is identical to the input distribution and not at all affected by the exponential service mechanism.

$$P\left(N(d) = n, \;\; d > t\right) = \int_t^{\infty} F_{n+1}(x)\mu_{n+1} dx = p_{n+1} \cdot \mu_{n+1} \int_t^{\infty} e^{-\lambda x} dx$$

$$= \frac{p_{n+1}\mu_{n+1}}{\lambda} \cdot e^{-\lambda t} = p_n e^{-\lambda t} = P\left(N(d) = n\right) P(d > t).$$

Thus $N(d)$ and d are independent of each other.

$$\begin{aligned} P\left(d_1 > t_1 \mid N(d_1) = n, \;\; d_2 > t_2\right) &= P\left(d_1 > t_1 \mid N(d_1) = n\right) \\ &= P(d_1 > t_1) \end{aligned}$$

Therefore d_1 and d_2 are independent of each other.
In many problems, a customer requires service from several service stations before a task is completed. These problems require that we consider a *network* of queueing systems. In such networks, the departures from some queues become the arrivals to other queues. This is the reason why we are interested in the statistical properties of the departure process from a queue.

Consider two queues in tandem as shown in Fig. 2.2, where the departures from the first queue become the arrivals at the second queue. Assume that the arrivals to the first queue are Poisson with rate λ and that the service time at queue 1 is exponentially distributed with rate $\mu_1 > \lambda$ Assume that the service time in queue 2 is also exponentially distributed with rate $\mu_2 > \lambda$. The state of this system is specified by the number of customers in the two queues, $(N_1(t), N_2(t))$ This state vector forms a Markov process with the transition rate diagram shown in Fig. 2.3, and global balance equations are

$$\lambda P[N_1 = 0, N_2 = 0] = \mu_2 P[N_1 = 0, N_2 = 1] \tag{2.11}$$

$$(\lambda + \mu_1) P[N_1 = n, N_2 = 0] = \mu_2 P[N_1 = n, N_2 = 1] \tag{2.12}$$

$$+ \lambda P[N_1 = n - 1, N_2 = 0] \quad n > 0 \tag{2.13}$$

$$(\lambda + \mu_2) P[N_1 = 0, N_2 = m] = \mu_2 P[N_1 = 0, N_2 = m + 1] \tag{2.14}$$

$$+ \mu_1 P[N_1 = 1, N_2 = m - 1] \quad m > 0 \tag{2.15}$$

$$(\lambda + \mu_1 + \mu_2) P[N_1 = n, N_2 = m] = \mu_2 P[N_1 = n, N_2 = m + 1] \tag{2.16}$$

$$+ \mu_1 P[N_1 = n + 1, N_2 = m - 1] \tag{2.17}$$

$$+ \lambda P[N_1 = n - 1, N_2 = m] \tag{2.18}$$

$$n > 0, m > 0. \tag{2.19}$$

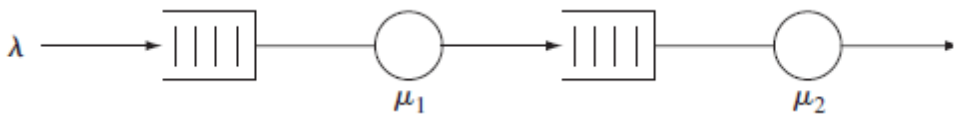


Figure 2.2: Two tandem exponential queues with Poisson input

It is easy to verify that the following joint probabilities satisfy Eqs. 2.11 through 2.19

$$P[N_1 = n, N_2 = m] = (1 - \rho_1)\rho_1^n (1 - \rho_2)\rho_2^m, \quad n \geq 0, m \geq 0, \tag{2.20}$$

where $\rho_i = \lambda/\mu_i$. We know that the first queue is an M/M/1 system, so

$$P[N_1 = n] = (1 - \rho_1)\rho_1^n, \quad n = 0, 1, \cdots \tag{2.21}$$

By summing Eq. 2.20 over all n, we obtain the marginal distribution of the second queue, that is

$$P[N_2 = m] = (1 - \rho_2)\rho_2^m, \quad m \geq 0. \tag{2.22}$$

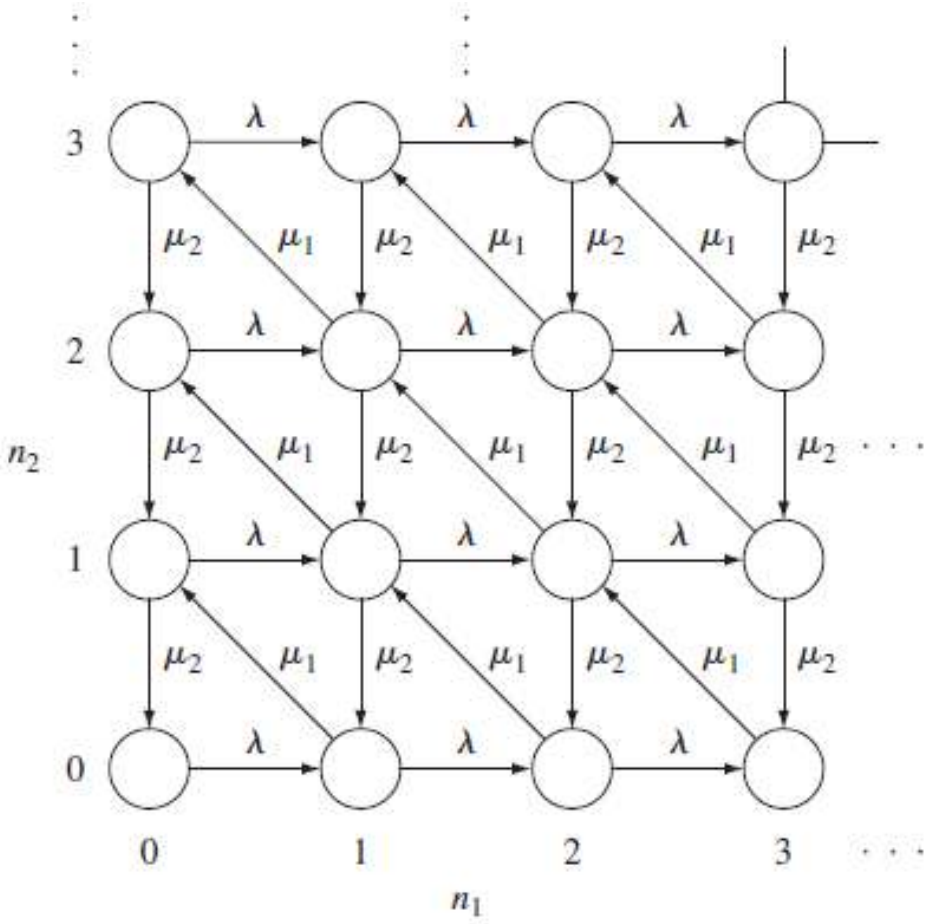


Figure 2.3: Transition rate diagram for two tandem exponential queues with Poisson input.

Equations 2.20 through 2.22 imply that

$$P[N_1 = n, N_2 = m] = P[N_1 = n]P[N_2 = m] \quad \text{for all } n, m. \tag{2.23}$$

In words, the number of customers at queue 1 and the number at queue 2 at the same time instant are independent random variables. Furthermore, the steady-state distribution at the second queue is that of an $M/M/1$ system with Poisson arrival rate λ and exponential service time μ_2.

We say that a network of queues has a **product-form solution** when the joint distribution of the vector of numbers of customers at the various queues is equal to the product of the marginal distribution of the number in the individual queues. We now discuss Burke's theorem, which states the fundamental result underlying the product-form solution in Eq. 2.23.

Burke's Theorem Consider an $M/M/1$, $M/M/c$, or $M/M/\infty$ queueing system at steady state with arrival rate λ then

1. The departure process is Poisson with rate λ
2. At each time t, the number of customers in the system $N(t)$ is independent of the sequence of departure times prior to t.

The product-form solution for the two tandem queues follows from Burke's theorem. Queue 1 is an $M/M/1$ queue, so from part 1 of the theorem the departures from queue 1 form a Poisson process. Thus the arrivals to queue 2 are a Poisson process, so the second queue is also an $M/M/1$ system with steady state pmf given by Eq. 2.22. It remains to show that the numbers of customers in the two queues at the same time instant are independent random variables.

The arrivals to queue 2 prior to time t are the departures from queue 1 prior to time t. By part 2 of Burke's theorem the departures from queue 1, and hence the arrivals to queue 2, prior to time t are independent of $N_1(t)$. Since $N_2(t)$ is determined by the sequence of arrivals from queue 1 prior to time t and the independent sequence of service times, it then follows that $N_1(t)$ and $N_2(t)$ are independent. Equation 2.23 then follows. Note that Burke's theorem does not state that $N_1(t)$ and $N_2(t)$ are independent random processes. This would require that $N_1(t)$ and $N_2(t)$ be independent random variables for all t_1 and t_2. This is clearly not the case.

Burke's theorem implies that the generalization of Eq. 2.23 holds for the tandem combination of any number of $M/M/1$, $M/M/c$, $M/M/\infty$ queues. Indeed, the result holds for any "feedforward" network of queues in which a customer cannot visit any queue more than once.

Example 4 *Find the joint distribution for the network of queues shown in Fig. 2.4, where queue 1 is driven by a Poisson process of rate λ_1, where the departures from queue 1 are randomly routed to queues 2 and 3, and where queue 3 also has an additional independent Poisson arrival stream of rate λ_2.*

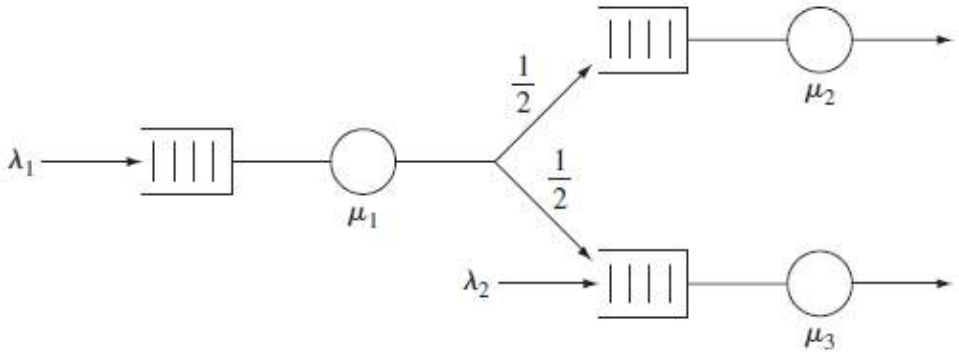


Figure 2.4: A feed-forward network of queues.

From Burke's theorem $N_1(t)$ and $N_2(t)$ are independent, as are $N_1(t)$ and $N_3(t)$. Since the random split of a Poisson process yields independent Poisson processes, we have that the inputs to queues 2 and 3 are independent. The input to queue 2 is Poisson with rate $\lambda_1/2$. The input to queue 3 is Poisson of rate $\lambda_1/2+\lambda_2$ since the merge of two independent Poisson processes is also Poisson. Thus

$$\begin{aligned} P[N_1(t) = k, N_2(t) = m, N_3(t) = n] \\ &= (1-\rho_1)\rho_1^k(1-\rho_2)\rho_2^m(1-\rho_3)\rho_3^n \quad k,m,n \geq 0, \end{aligned}$$

where $\rho_1 = \lambda_1/\mu_1$, $\rho_2 = \lambda_1/(2\mu_2)$, $\rho_3 = (\lambda_1/2 + \lambda_2)/\mu_3$, *and where we have assumed that all of the queues are stable.*

Now let us consider an $M/G/1$ system and we are interested in under which service time distribution the inter-departure time is exponentially distributed with parameterλ. First prove that the utilization of the system is $U_S = \varrho = \lambda\mathbb{E}(S)$. As it is understandable for any stationary stable $G/G/1$ queueing system the mean number of departures during the mean busy period length of the server is one more than the mean number of arrivals during the mean busy period length of the server. That is

$$\frac{\mathbb{E}(\delta)}{\mathbb{E}(S)} = 1 + \frac{\mathbb{E}(\delta)}{\mathbb{E}(\tau)},$$

where $\mathbb{E}(\tau)$ denotes the mean inter-arrival times. Hence

$$\mathbb{E}(\tau) + \mathbb{E}(\delta) = \mathbb{E}(\delta)\frac{\mathbb{E}(\tau)}{\mathbb{E}(S)}$$
$$\mathbb{E}(\delta) = \frac{\mathbb{E}(\tau)\mathbb{E}(S)}{\mathbb{E}(\tau) - \mathbb{E}(S)} = \mathbb{E}(S)\frac{1}{1-\varrho},$$

where $\varrho = \frac{\mathbb{E}(S)}{\mathbb{E}(\tau)}$. Clearly

$$U_S = \frac{\mathbb{E}(\delta)}{\mathbb{E}(\tau) + \mathbb{E}(\delta)} = \frac{\mathbb{E}(S)\frac{1}{1-\varrho}}{\mathbb{E}(\tau) + \frac{\mathbb{E}(S)}{1-\varrho}} = \frac{\frac{\varrho}{1-\varrho}}{1 + \frac{\varrho}{1-\varrho}} = \varrho < 1.$$

Thus the utilization for an $M/G/1$ system is ϱ. It should be noted that an $M/G/1$ system $D_k = P_k$, that is why our question can be formulated as

$$\frac{\lambda}{\lambda+s} = \varrho L_S(s) + (1-\varrho)\frac{\lambda}{\lambda+s}L_S(s) = L_S(s)\left(\varrho + \frac{\lambda(1-\varrho)}{\lambda+s}\right)$$
$$= L_S(s)\frac{\lambda^2\mathbb{E}(S) + s\lambda\mathbb{E}(S) + \lambda - \lambda^2\mathbb{E}(S)}{\lambda+s} = L_S(s)\frac{\lambda(1+s\mathbb{E}(S))}{\lambda+s},$$

thus

$$L_S(s) = \frac{1}{1+s\mathbb{E}(S)},$$

which is the Laplace-transform of an exponential distribution with mean $\mathbb{E}(S)$. In summary, only exponentially distributed service times assures that Poisson arrivals involves Poisson departures with the same parameters.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Example 5 *Let us consider a small post office in a village where on the average 70 customers arrive according to a Poisson process during a day. Let us assume that the service times are exponentially distributed with rate 10 clients per hour and the office operates 10 hours daily. Find the mean queue length, and the probability that the number of waiting customer is greater than 2. What is the mean waiting time and the probability that the waiting time is greater than 20 minutes ?*

Solution:
Let the time unit be an hour. Then $\lambda = 7$, $\mu = 10$, $\rho = \frac{7}{10}$

$$\begin{aligned}
\overline{N} &= \frac{\rho}{1-\rho} = \frac{7}{3} \\
\overline{Q} &= \overline{N} - \rho = \frac{7}{3} - \frac{7}{10} = \frac{70-21}{30} = \frac{49}{30} \\
P(n>3) &= 1 - P(n \leq 3) = 1 - P_0 - P_1 - P_2 - P_3 \\
&= 1 - 1 + \rho - (1-\rho)(\rho + \rho^2 + \rho^3) = \rho^4 = 0.343 \cdot 0.7 = 0.2401 \\
\overline{W} &= \frac{\overline{N}}{\mu} = \frac{7}{3 \cdot 10} = \frac{7}{30} \text{ hour } = 14 \text{ minutes} \\
P\left(W > \frac{1}{3}\right) &= 1 - F_W\left(\frac{1}{3}\right) = 0.7 \cdot e^{-10 \cdot \frac{1}{3} \cdot 0.3} = 0.7 \cdot e^{-1} = 0.257
\end{aligned}$$

■
The following 4 Examples are taken from Allen [3].

Example 6 *For a small batch computing system the processing time per job is exponentially distributed with an average time of 3 minutes. Jobs arrive randomly at an average rate of one job every 4 minutes and are processed on a first-come-first-served basis. The manager of the installation has the following concerns.*

(a) What is the probability that an arriving job will require more than 20 minutes to be processed (the job turn-around time exceeds 20 minutes)?

(b) A queue of jobs waiting to be processed will form, occasionally. What is the average number of jobs waiting in this queue?

(c) It is decided that, when the work load increases to the level such that the average time in the system reaches 30 minutes, the computer system capacity will be increased. What is the average arrival rate of jobs per hour at which this will occur? What is the percentage increase over the present job load? What is the average number of jobs in the system at this time?

(d) Suppose the criterion for upgrading the computer capacity is that not more than 10% of all jobs have a time in the system (turn-around time) exceeding 40 minutes. At the arrival rate at which this criterion is reached, what is the average number of jobs waiting to be processed?

Solution:

(a)

$$\begin{aligned}
&E(\tau) = 4 \text{ minutes, so} \\
&\quad \lambda = 1/E(\tau) = 0.25 \text{ jobs/minute} \\
&\text{and} \\
&\quad \rho = \lambda E(S) = 0.25 \times 3 = 0.75.
\end{aligned}$$

The average time in the system, $\bar{T} = E(S)/(1-\rho) = 12$ minutes, so

$$F_T(t) = P(T \le t) = 1 - e^{-t/12} \text{ or } P(T > t) = e^{-t/12}$$

Therefore, the probability that T exceeds 20 minutes is $e^{-20/12} = e^{-5/3} = 0.1889$

(b) If we assume a job queue has not formed unless there is a job in it, we use the formula

$$E[Q|Q > 0] = 1/(1-\rho) = 4 \text{ jobs}$$

If the question is interpreted to mean the average job queue length, including queues of length zero, then we calculate

$$\bar{Q} = E(Q) = \rho^2/(1-\rho) = (0.75)^2/0.25 = 2.25 \text{ jobs.}$$

The most reasonable answer to the question, as stated, is 4 jobs.

(c) When $\bar{T} = 30$ minutes the system is to be upgraded, assuming the current $E(S)$ is 3 minutes. We solve the equation

$$30 = \bar{T} = \frac{E(S)}{1-\lambda E(S)} = \frac{3}{1-3\lambda}$$

or

$$\lambda = 27/90 = 3/10 \text{ jobs/minute} = 18 \text{ jobs/hour}$$

The percentage increase is

$$100 \times (18-15)/15 = 100/5 = 20\%$$

When $\lambda = 18$ jobs/hour $= 3/10$ jobs/minute, the average number of jobs in the system

$$\bar{N} = \rho/(1-\rho) = 0.9/(1-0.9) = 9 \text{ jobs.}$$

(d) The criterion is that $\pi_T(90)$ reaches 40 minutes. We solve the equation

$$40 = \pi_T(90) = 2.8\bar{T} = \frac{2.3 \times E[s]}{1-\lambda E[s]} = \frac{2.3 \times 3}{1-3\lambda},$$

to obtain

$$\lambda = \frac{33.1}{120} \text{ jobs/minute} = 60 \times \frac{33.1}{120} = 16.55 \text{ jobs/hours.}$$

That is only a $[(16.55-15)/15] \times 100 = 10.3\%$ increase over the present arrival rate. At this arrival rate $\rho = \lambda E[s] = 0.8275$ and the average number of jobs in the queue is

$$\bar{Q} = E[Q] = \rho^2/(1-\rho) = 3.97$$

This is an increase over the current value of 2.25 jobs. The average time in the system at this increased arrival rate is 17.39 minutes; it is only 12 minutes at the current arrival rate.

In part (c) of the above example we see that increasing the arrival rate by 20% increased the average time a job would spend in the system from 12 minutes to 30 minutes a 150% increase! The curve of $E(T)/E(S)$ rises sharply as ρ approaches the value 1.

That is, the slope of the curve increases rapidly as ρ grows beyond about 0.8. Since

$$d\bar{T}/d\rho = E(S)(1 - \lambda E(S)^{-2}$$

a small change in ρ (due to a small change in λ, assuming $E[s]$ is fixed) causes a change in $\bar{T}$ given approximately by

$$(d\bar{T}/d\rho)\Delta\rho = (d\bar{T}/d\rho)E(S)\Delta\lambda = E(S)^2(1 - \lambda E(S))^{-2}\Delta\lambda.$$

Thus, if $\rho = 0.5$, a change $\Delta\lambda$ in λ will cause a change in $\bar{T}$ of about $4E[s]^2\Delta\lambda$, while, if $\rho = 0.9$, the change in $\bar{T}$ will be about $100E[s]^2\Delta\lambda$, or 2.5 times the size of the change that occurred for $\rho = 0.5$!
That is, when the system is operating at 90%server utilization, a small change in the system load (arrival rate) will cause 25 times as great an increase in the average system time as the same increase in load would cause if the system were operating at 50% utilization! This illustrates the danger of designing a system to operate at a high utilization level - a small increase in the load can have disastrous effects on the system performance. ■

Example 7 *A computing facility has a large computer dedicated to a certain type of on-line application for users who are scattered about the country. The arrival pattern of requests to the central machine is random (Poisson), and the service time provided is random (exponential) also, so the system is an $M/M/1$ queueing system. A proposal is made that the workload be divided equally among n smaller machines - each with $1/n$ times the processing power of the original machine. It is claimed that the response time (time a request is in the system) will not change but the users will have a local computer. Are these claims justified?*

Solution: Let λ, μ be the average arrival and service rates, respectively, of the current system so that $\rho = \lambda/\mu$ is the computer utilization. For each of the proposed new systems the average arrival rate is λ/n and the average service rate is μ/n, so the server utilization is $(\lambda/n)/(\mu/n) = \lambda/\mu = \rho$, the same value as the present system. If we assume the small computers also provide random service, then

$$\frac{\bar{T}_{proposed}}{\bar{T}_{current}} = \left(\frac{n/\mu}{(1-\rho)}\right) / \left(\frac{1/\mu}{1-\rho}\right) = n,$$

and

$$\frac{\bar{W}_{proposed}}{\bar{W}_{current}} = \left(\frac{\rho n/\mu}{(1-\rho)}\right) / \left(\frac{\rho/\mu}{1-\rho}\right) = n.$$

Thus, the average time in the system and the average time in the queue would increase n-fold rather than remain the same! Of course the n new computer systems, together, process the same number of requests per hour as before, but each individual request requires n times as long to be processed, on the average, as in the present system. Thus, if the present system has an average service time of 2 seconds with a utilization of 0.7, then it has an average response time of 6.67 seconds; a proposed system of 10 computers, each providing 20 seconds service time, would yield a response time of 66.7 seconds! The effect discussed in this example is called the "scaling effect". The result can be used to show that centralizing a computing facility can improve the response time while providing more computing capability for less money (economy of scale). ■

Example 8 *A branch office of a large engineering firm has one on-line terminal connected to a central computer system for 16 hours each day. Engineers, who work throughout the city, drive to the branch office to use the terminal for making routine calculations. The arrival pattern of engineers is random (Poisson) with an average of 20 persons per day using the terminal. The distribution of time spent by an engineer at the terminal is exponential with an average time of 30 minutes. Thus the terminal is 5/8 utilized (20 × 1/2 = 10 hours out of 16 hours available). The branch manager receives complaints from the staff about the length of time many of them have to wait to use the terminal. It does not seem reasonable to the manager to procure another terminal when the present one is only used five-eighths of the time, on the average. How can queueing theory help this manager?*

Solution:

The $M/M/1$ queueing system is a reasonable model with $\rho = 5/8$, as we computed above. The $M/M/1$ formulas give the following.

$\bar{T} = E(T) = E(S)/(1-\rho) = 80$ minutes.	Average time an engineer spends at the branch office.
$\bar{Q} = \rho^2/(1-\rho) = 1.0417.$	Average number of engineers waiting in the queue.
$E(Q\|Q>0) = 1/(1-\rho) = 8/3.$	Average number of engineers in nonempty queues.
$\bar{W} = E(W) = \rho E(S)/(1-\rho) = 50$ minutes.	Average waiting time in queue.
$E(W\|W>0) = E(T) = 80$ minutes.	Average waiting time of those who must wait.
$\pi_W(90) = \bar{T}\ln 10\rho = 146.61$ minutes.	90th percentile of time in the queue.
$\pi_T(90) \approx 2.3\bar{T} = 184$ minutes.	90th percentile time in the branch office.

Since $\rho = 5/8$, only three-eighths of the engineers who use the terminal need not wait. For those who must wait, the average wait for the terminal is 80 minutes - quite a long

wait, by most standards! Ten percent of the engineers spend over 3 hours (actually 184 minutes) in the office to do an average of 30 minutes of computing. The probability of waiting more than an hour to use the terminal is

$$P[W > 60] = \frac{5}{8}e^{-60/80} = 0.295229,$$

or almost 30%

These results may seem a little startling to those not acquainted with queueing theory. It might seem, intuitively, that adding another terminal would cut the average waiting time in half - from 50 minutes to 25 minutes (to 40 minutes for those who must wait). The queueing theory we have presented so far should suffice to convince the manager that an improvement is needed. ■

Example 9 *Traffic to a message switching center for one of the outgoing communication lines arrives in a random pattern at an average rate of 240 messages per minute. The line has a transmission rate of 800 characters per second. The message length distribution (including control characters) is approximately exponential with an average length of 176 characters. Calculate the principal statistical measures of system performance assuming that a very large number of message buffers are provided. What is the probability that 10 or more messages are waiting to be transmitted?*

Solution: The average service time is the average time to transmit a message or

$$\begin{aligned} E(S) &= \frac{\text{average message length}}{\text{line speed}} \\ &= \frac{176 \text{ characters}}{800 \text{ characters/second}} = 0.22 \text{ seconds.} \end{aligned}$$

Hence, since the average arrival rate

$$\lambda = 240 \text{ messages/minute} = 4 \text{ messages/second},$$

the server utilization

$$\rho = \lambda E(S) = 4 \times 0.22 = 0.88,$$

that is, the communication line is transmitting outgoing messages 88% of the time.

$\bar{N} = E(N) = \rho/(1-\rho) = 7.33$ messages)	Average number of messages in the system
$\bar{Q} = E(Q) = \rho^2/(1-\rho) = 6.45$ messages).	Average number of messages in the queue waiting to be transmitted.
$\bar{T} = E(T) = E(S)/(1-\rho) = 1.83$ seconds.	Average time a message spend in the system.
$\bar{W} = E(W) = \rho E(S)/(1-\rho) = 1.61$ seconds.	Average time a message wait for transmission.
$\pi_T(90) = 2.3\bar{T} = 4.209$ seconds.	90th percentile time in the system.
$\pi_W(90) = \bar{T}\ln 10\rho = 3.98$ seconds.	90th percentile waiting time in queue (90% of the messages wait no longer than 3.98 seconds.)

Since 10 or more messages are waiting if and only if 11 or more messages are in the system, the required probability is

$$P(\text{11 or more messages in the system}) = \rho^{11} = 0.245.$$

Our discussion of the $M/M/1$ model has been more complete than it will be for many queueing models because it is an important but simple model. It is also a pleasant model to study because the probability distributions of the random variables T, W, N and Q can be calculated; for some queueing models only the averages $\bar{T}$, $\bar{W}$, $\bar{N}$, and $\bar{Q}$ can be computed, and these only with difficulty. A number of systems can be modeled, at least in a limiting sense, as an $M/M/1$ queueing system.

■

2.2 The $M/M/1$ Queue with Balking Customers

Let us consider a modification of an $M/M/1$ system in which customers are discouraged when more and more requests are present at their arrivals. Let us denote by b_k the probability that a customers joints to the systems provided there are k customers in the system at the moment of his arrival.

It is easy to see, that the number of customers in the system is a birth-death process with birth rates

$$\lambda_k = \lambda \cdot b_k, \quad k = 0, 1, \ldots$$

Clearly, there are various candidates for b_k but we have to find such probabilities which result not too complicated formulas for the main performance measures. Keeping in mind this criteria let us consider the following

$$b_k = \frac{1}{k+1}, \quad k = 0, 1, \ldots$$

Thus

$$P_k = \frac{\rho^k}{k!} P_0, \qquad k = 0, 1, \ldots,$$

and then using the normalization condition we get

$$P_k = \frac{\rho^k}{k!} e^{-\rho}, \qquad k = 0, 1, \ldots$$

The stability condition is $\rho < \infty$, that is we do not need the condition $\rho < 1$ as in an $M/M/1$ system.
Notice that the number of customers follows a Poisson law with parameter ρ and we can expect that the performnace measures can be obtained in a simple way.

Performance measures

- $$U_S = 1 - P_0 = 1 - e^{-\rho},$$
  $$U_S = \frac{\mathbb{E}(\delta)}{\frac{1}{\lambda} + \mathbb{E}(\delta)},$$

  hence

  $$\mathbb{E}(\delta) = \frac{1}{\lambda} \cdot \frac{U_S}{1 - U_S} = \frac{1}{\lambda} \cdot \frac{1 - e^{-\rho}}{e^{-\rho}}.$$

- $$\overline{N} = \rho,$$
  $$Var(N) = \rho$$

- $$\overline{Q} = \overline{N} - U_S = \rho - (1 - e^{-\rho}) = \rho + e^{-\rho} - 1.$$
  $$\begin{aligned}\mathbb{E}(Q^2) &= \sum_{k=1}^{\infty} (k-1)^2 P_k = \sum_{k=1}^{\infty} k^2 P_k - 2\sum_{k=1}^{\infty} k P_k + \sum_{k=1}^{\infty} P_k \\ &= \mathbb{E}(N^2) - 2\overline{N} + U_S = \rho + \rho^2 - 2\rho + U_S = \rho^2 - \rho + 1 - e^{-\rho}.\end{aligned}$$

  Thus

  $$\begin{aligned}Var(Q) &= \mathbb{E}(Q^2) - (\mathbb{E}(Q))^2 = \rho^2 - \rho + 1 - e^{-\rho} - (\rho + e^{-\rho} - 1)^2 \\ &= \rho^2 - \rho + 1 - e^{-\rho} - \rho^2 - e^{-2\rho} - 1 - 2\rho e^{-\rho} + 2\rho + 2e^{-\rho} \\ &= \rho - e^{-2\rho} + e^{-\rho} - 2\rho e^{-\rho} = \rho - e^{-\rho}(e^{-\rho} + 2\rho - 1).\end{aligned}$$

- The probability that an arriving customer enters/joins into the system can be obtained with the help of the Bayes-formula, namely

$$P_J = \lim_{h \to 0} \frac{\sum_{j=0}^{\infty} (\lambda_j h + o(h)) P_j}{\sum_{k=0}^{\infty} (\lambda h + o(h)) P_k} = \frac{\sum_{j=0}^{\infty} \lambda_j P_j}{\sum_{k=0}^{\infty} \lambda P_k} = \frac{\mu(1 - e^{-\rho})}{\lambda} = \frac{1 - e^{-\rho}}{\rho}.$$

- To get the distribution of the response and waiting times we have to know the distribution of the system at the instant when an arriving customer joins to the system.

  By applying the Bayes's rule it is not difficult to see that

$$\Pi_k = \frac{\frac{\lambda}{k+1} \cdot P_k}{\sum_{i=0}^{\infty} \frac{\lambda}{i+1} \cdot P_i} = \frac{\frac{\rho^{k+1}}{(k+1)!} \cdot e^{-\rho}}{\sum_{i=0}^{\infty} \frac{\rho^{i+1}}{(i+1)!} e^{-\rho}} = \frac{P_{k+1}}{1-e^{-\rho}}.$$

  Notice, that this time

$$\Pi_k \neq P_k.$$

  Let us first determine $\overline{T}$ and then $\overline{W}$.

  By the law of total expectations we have

$$\overline{T} = \sum_{k=0}^{\infty} \frac{k+1}{\mu} \Pi_k = \frac{1}{\mu} \sum_{k=0}^{\infty} \frac{(k+1)P_{k+1}}{1-e^{-\rho}} = \frac{1}{\mu(1-e^{-\rho})} \cdot \overline{N} = \frac{\rho}{\mu(1-e^{-\rho})}.$$
$$\overline{W} = \overline{T} - \frac{1}{\mu} = \frac{1}{\mu} \left(\frac{\rho + e^{-\rho} - 1}{1-e^{-\rho}} \right).$$

  As we have proved in formula (1.5)

$$\overline{\lambda} = \sum_{k=0}^{\infty} \lambda_k P_k = \sum_{k=1}^{\infty} \mu_k P_k = \sum_{k=1}^{\infty} \mu P_k = \mu(1-e^{-\rho}),$$

  thus

$$\overline{\lambda} \cdot \overline{T} = \mu(1-e^{-\rho}) \cdot \frac{\rho}{\mu(1-e^{-\rho})} = \rho = \overline{N},$$
$$\overline{\lambda} \cdot \overline{W} = \mu(1-e^{-\rho}) \cdot \frac{\rho + e^{-\rho} - 1}{\mu(1-e^{-\rho})} = \rho + e^{-\rho} - 1 = \overline{Q}$$

  which is the **Little formula** for this system.

- To find the distribution of T and W we have to use the same approach as we did earlier, namely

$$f_T(x) = \sum_{k=0}^{\infty} f_T(x|k) \cdot \Pi_k = \sum_{k=0}^{\infty} \frac{\mu(\mu x)^k e^{-\mu x}}{k!} \cdot \frac{\rho^{k+1}}{(k+1)!} \frac{e^{-\rho}}{1-e^{-\rho}}$$
$$= \frac{\lambda e^{-(\rho+\mu x)}}{1-e^{-\rho}} \sum_{k=0}^{\infty} \frac{(\mu x \rho)^k}{k!(k+1)!},$$

  which is difficult to calculate. We have the same problems with $f_W(x)$, too.

However, the Laplace-transforms $L_T(s)$ and $L_W(s)$ can be obtained and the hence the higher moments can be derived.
Namely

$$\begin{aligned} L_T(s) &= \sum_{k=0}^{\infty} L_T(s|k)\Pi_k = \sum_{k=0}^{\infty} \left(\frac{\mu}{\mu+s}\right)^{k+1} \frac{\frac{\rho^{k+1}}{(k+1)!}e^{-\rho}}{1-e^{-\rho}} \\ &= \frac{e^{-\rho}}{1-e^{-\rho}} \sum_{k=0}^{\infty} \left(\frac{\mu\rho}{\mu+s}\right)^{k+1} \frac{1}{(k+1)!} = \frac{e^{-\rho}}{1-e^{-\rho}} \left(e^{\frac{\mu\rho}{\mu+s}} - 1\right). \\ L_W(s) &= L_T(s) \cdot \frac{\mu+s}{\mu}. \end{aligned}$$

Find $\overline{T}$ by the help of $L_T(s)$ to check the formula. It is easy to see that

$$\begin{aligned} L_T'(s) &= \frac{e^{-\rho}}{1-e^{-\rho}} \cdot e^{\frac{\mu\rho}{\mu+s}}(-\mu\rho(\mu+s)^{-2}) \\ L_T'(0) &= -\frac{e^{-\rho}}{1-e^{-\rho}} e^{\rho} \cdot \frac{\rho}{\mu} = -\frac{\rho}{\mu(1-e^{-\rho})}. \end{aligned}$$

Hence

$$\overline{T} = \frac{\rho}{\mu(1-e^{-\rho})},$$

as we have obtained earlier. $\overline{W}$ can be verified similarly.

To get $Var(T)$ and $Var(W)$ we can use the Laplace-transform method. As we have seen

$$L_T(s) = \frac{e^{-\rho}}{1-e^{-\rho}} \left(e^{\frac{\lambda}{\mu+s}} - 1\right).$$

Thus

$$L_T'(s) = \frac{e^{-\rho}}{1-e^{-\rho}} \cdot e^{\frac{\lambda}{\mu+s}}(-1)\lambda(\mu+s)^{-2},$$

therefore

$$L_T''(s) = \frac{e^{-\rho}}{1-e^{-\rho}} \cdot \left(e^{\frac{\lambda}{\mu+s}} \left((-1)\lambda(\mu+s)^{-2}\right)^2 + 2\lambda(\mu+s)^{-3} \cdot e^{\frac{\lambda}{\mu+s}}\right).$$

Hence

$$L_T''(0) = \frac{e^{-\rho}}{1-e^{-\rho}} \left(e^{\rho} \left(-\frac{\rho}{\mu}\right)^2 + \frac{2\rho}{\mu^2} e^{\rho}\right) = \frac{1}{\mu^2} \cdot \frac{\rho^2+2\rho}{1-e^{-\rho}}.$$

Consequently

$$\begin{aligned} Var(T) &= \frac{1}{\mu^2} \cdot \frac{\rho^2+2\rho}{1-e^{-\rho}} - \left(\frac{\rho}{\mu(1-e^{-\rho})}\right)^2 \\ &= \frac{(\rho^2+2\rho)(1-e^{-\rho}) - \rho^2}{\mu^2(1-e^{-\rho})^2} = \frac{\rho^2+2\rho-\rho^2e^{-\rho}-2\rho e^{-\rho}-\rho^2}{\mu^2(1-e^{-\rho})^2} \\ &= \frac{2\rho-\rho^2e^{-\rho}-2\rho e^{-\rho}}{\mu^2(1-e^{-\rho})^2} = \frac{\rho(2-(\rho+2)e^{-\rho})}{\mu^2(1-e^{-\rho})^2}. \end{aligned}$$

However, W and T can be considered as a random sum, too. That is

$$Var(W) = \mathbb{E}(N_a)\frac{1}{\mu^2} + Var(N_a)\left(\frac{1}{\mu}\right)^2 = \frac{1}{\mu^2}(\mathbb{E}(N_a) + Var(N_a)).$$

$$\begin{aligned}\mathbb{E}(N_a) &= \sum_{k=1}^{\infty} k\Pi_k = \sum_{k=1}^{\infty} \frac{kP_{k+1}}{1-e^{-\rho}} \\ &= \frac{1}{1-e^{-\rho}}\left(\sum_{k=0}^{\infty}(k+1)P_{k+1} - \sum_{k=0}^{\infty}P_{k+1}\right) \\ &= \frac{1}{1-e^{-\rho}}\left(\rho + e^{-\rho} - 1\right).\end{aligned}$$

Since

$$Var(N_a) = \mathbb{E}(N_a^2) - (\mathbb{E}(N_a))^2$$

first we have to calculate $\mathbb{E}(N_a^2)$, that is

$$\begin{aligned}\mathbb{E}(N_a^2) &= \sum_{k=1}^{\infty} k^2\Pi_k = \sum_{k=1}^{\infty} k^2\frac{P_{k+1}}{1-e^{-\rho}} \\ &= \frac{1}{1-e^{-\rho}}\sum_{k=0}^{\infty}\left((k+1)^2 - 2k - 1\right)P_{k+1} \\ &= \frac{1}{1-e^{-\rho}}\left(\sum_{k=0}^{\infty}(k+1)^2P_{k+1} - 2\sum_{k=0}^{\infty}kP_{k+1} - \sum_{k=0}^{\infty}P_{k+1}\right) \\ &= \frac{1}{1-e^{-\rho}}\left(\rho + \rho^2 - 2\left(\rho + e^{-\rho} - 1\right) - \left(1 - e^{-\rho}\right)\right) \\ &= \frac{1}{1-e^{-\rho}}\left(\rho^2 - \rho - e^{-\rho} + 1\right).\end{aligned}$$

Therefore

$$\begin{aligned}Var(N_a) &= \frac{1}{1-e^{-\rho}}\left(\rho^2 - \rho - e^{-\rho} + 1\right) - \left(\frac{1}{1-e^{-\rho}}(\rho + e^{-\rho} - 1)\right)^2 \\ &= \left(\frac{1}{1-e^{-\rho}}\right)^2\left(\left(1-e^{-\rho}\right)\left(\rho^2 - \rho - e^{-\rho} + 1\right) - \left(\rho + e^{-\rho} - 1\right)^2\right) \\ &= \left(\frac{1}{1-e^{-\rho}}\right)^2(\rho^2 - \rho - e^{-\rho} + 1 - \rho^2e^{-\rho} + \rho e^{-\rho} + e^{-2\rho} - e^{-\rho} \\ &- \rho^2 - e^{-2\rho} - 1 - 2\rho e^{-\rho} + 2\rho - 2e^{-\rho}) \\ &= \frac{\rho - e^{-\rho}(\rho^2 + \rho)}{(1-e^{-\rho})^2}.\end{aligned}$$

Finally

$$\begin{aligned}Var(W) &= \left(\frac{1}{\mu}\right)^2\left(\frac{1}{1-e^{-\rho}}(\rho + e^{-\rho} - 1) + \frac{\rho - e^{-\rho}(\rho^2+\rho)}{(1-e^{-\rho})^2}\right) \\ &= \frac{1}{(\mu(1-e^{-\rho}))^2}((\rho + e^{-\rho} - 1)(1-e^{-\rho}) + \rho - e^{-\rho}(\rho^2 + \rho)).\end{aligned}$$

Thus

$$
\begin{aligned}
Var(T) &= Var(W) + \frac{1}{\mu^2} \\
Var(T) &= \left(\frac{1}{\mu(1-e^{-\rho})}\right)^2 (\rho + e^{-\rho} - 1)(1-e^{-\rho}) + \rho - e^{-\rho}(\rho^2+\rho) + (1-e^{-\rho})^2) \\
&= \frac{(1-e^{-\rho})(\rho + e^{-\rho} - 1 + 1 - e^{-\rho}) + \rho - e^{-\rho}(\rho^2+\rho)}{(\mu(1-e^{-\rho}))^2} \\
&= \frac{2\rho - 2\rho e^{-\rho} - \rho^2 e^{-\rho}}{(\mu(1-e^{-\rho})^2}
\end{aligned}
$$

which is the same we have obtained earlier.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.3 The $M/M/1$ Priority Queues

In the following let us consider an $M/M/1$ systems with priorities. This means that we have two classes of customers. Each type of requests arrive according to a Poisson process with parameter λ_1, and λ_2, respectively and the processes are supposed to be independent of each other. The service times for each class are assumed to be exponentially distributed with parameter μ. The system is stable if

$$\rho_1 + \rho_2 < 1,$$

where $\rho_i = \lambda_i/\mu$, $i = 1, 2$.
Let us assume that class 1 has priority over class 2. This section is devoted to the investigation of preemptive and non-preemptive systems and some mean values are calculated.

Preemptive Priority

According to the discipline the service of a customer belonging to class 2 is never carried out if there is customer belonging to class 1 in the system. In other words it means that class 1 preempts class 2 that is if a class 2 customer is under service when a class 1 request arrives the service stops and the service of class 1 request starts. The interrupted service is continued only if there is no class 1 customer in the system.

Let N_i denote the number of class i customers in the system and let T_i stand for the response time of class i requests. Our aim is to calculate $\mathbb{E}(N_i)$ and $\mathbb{E}(T_i)$ for $i = 1, 2$. Since type 1 always preempts type 2 the service of class 1 customers is independent of the number of class 2 customers. Thus we have

$$\mathbb{E}(T_1) = \frac{1/\mu}{1-\rho_1}, \qquad \mathbb{E}(N_1) = \frac{\rho_1}{1-\rho_1}\,. \tag{2.24}$$

Since for all customers the service time is exponentially distributed with the same parameter, the number of customers does not depends on the order of service. Hence for the total number of customers in an $M/M/1$ we get

$$\mathbb{E}(N_1) + \mathbb{E}(N_2) = \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2}, \tag{2.25}$$

and then inserting (2.24) we obtain

$$\mathbb{E}(N_2) = \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2} - \frac{\rho_1}{1 - \rho_1} = \frac{\rho_2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)},$$

and using the Little's law we have

$$\mathbb{E}(T_2) = \frac{\mathbb{E}(N_2)}{\lambda_2} = \frac{1/\mu}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}.$$

Example 10 *Let us compare what is the difference if preemptive priority discipline is applied instead of FIFO.*

Let $\lambda_1 = 0.5$, $\lambda_2 = 0.25$ and $\mu = 1$. In FIFO case we get

$$\mathbb{E}(T) = 4.0, \quad \mathbb{E}(W) = 3.0, \quad \mathbb{E}(N) = 3.0$$

and in priority case we obtain

$$\mathbb{E}(T_1) = 2.0, \quad \mathbb{E}(W_1) = 1.0, \quad \mathbb{E}(N_1) = 1.0$$

$$\mathbb{E}(T_2) = 8.0, \quad \mathbb{E}(W_2) = 7.0, \quad \mathbb{E}(N_2) = 2.0$$

Non-preemptive Priority

The only difference between the two disciplines is that in the case the arrival of a class 1 customer does not interrupt the service of type 2 request. That is why sometimes this discipline is call HOL (Head Of the Line). Of course after finishing the service of class 1 starts.

By using the law of total expectations the mean response time for class 1 can be obtained as

$$\mathbb{E}(T_1) = \mathbb{E}(N_1)\frac{1}{\mu} + \frac{1}{\mu} + \rho_2\frac{1}{\mu}.$$

The last term shows the situation when an arriving class 1 customer find the server busy servicing a class 2 customer. Since the service time is exponentially distributed the residual service time has the same distribution as the original one. Furthermore, because of the Poisson arrivals the distribution at arrival moments is the same as at random

moments, that is the probability that the server is busy with class 2 customer is ρ_2. By using the Little's law

$$\mathbb{E}(N_1) = \lambda_1 \mathbb{E}(T_1),$$

after substitution we get

$$\mathbb{E}(T_1) = \frac{(1+\rho_2)/\mu}{1-\rho_1}, \qquad \mathbb{E}(N_1) = \frac{(1+\rho_2)\rho_1}{1-\rho_1}\,.$$

To get the means for class 2 the same procedure can be performed as in the previous case. That is using (2.25) after substitution we obtain

$$\mathbb{E}(N_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}\,,$$

and then applying the Little's law we have

$$\mathbb{E}(T_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}\,.$$

Example 11 *Now let us compare the difference between the two priority disciplines.* Let $\lambda_1 = 0.5$, $\lambda_2 = 0.25$ and $\mu = 1$, then

$$\mathbb{E}(T_1) = 2.5, \quad \mathbb{E}(W_1) = 1.5, \quad \mathbb{E}(N_1) = 1.25$$

$$\mathbb{E}(T_2) = 7.0, \quad \mathbb{E}(W_2) = 6.0, \quad \mathbb{E}(N_2) = 1.75$$

Of course knowing the mean response time and mean number of customers in the system the mean waiting time and the mean number of waiting customers can be obtained in the usual way.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.4 The $M/M/1/K$ Queue, Systems with Finite Capacity

Let K be the capacity of an $M/M/1$ system, that is the maximum number of customers in the system including the one under service. It is easy to see that the nu,ber of customers in the systems is a birth-death process with rates $\lambda_k = \lambda$, $k = 0, \ldots, K-1$ és $\mu_k = \mu$, $k = 1, \ldots, K$. For the steady-state distribution we have

$$P_k = \frac{\rho^k}{\sum_{i=0}^{K} \rho^i}, \qquad k = 0, \ldots, K,$$

that is

$$P_0 = \frac{1}{\sum_{i=0}^{K} \rho^i} = \begin{cases} \frac{1}{K+1}, & \rho = 1 \\ \\ \frac{1-\rho}{1-\rho^{K+1}}, & \rho \neq 1. \end{cases}$$

It should be noted that the system is stable for any $\rho > 0$ when K is fixed. However, if $K \to \infty$ the the stability condition is $\rho < 1$ since the distribution of $M/M/1/K$ converges to the distribution of $M/M/1$.
It can be verified analytically since $\rho^K \to 0$ then $P_0 \to 1 - \rho$.

Similarly to an $M/M/1$ systems after reasonable modifications the *performance measures* can be computed as

- $$U_S = 1 - P_0,$$
$$\mathbb{E}(\delta) = \frac{1}{\lambda} \frac{U_S}{1 - U_S}$$

- $$\begin{aligned} \overline{N} &= \sum_{k=1}^{K} k\rho^k P_0 = \rho P_0 \sum_{k=1}^{K} k\rho^{k-1} \\ &= \rho P_0 \left(\sum_{k=1}^{K} \rho^k \right)' = \rho P_0 \left(\rho \frac{1-\rho^K}{1-\rho} \right)' = \rho P_0 \left(\frac{\rho - \rho^{K+1}}{1-\rho} \right)' \\ &= \left(\left(1 - (K+1)\rho^K\right)(1-\rho) + \rho - \rho^{K+1} \right) \cdot \frac{\rho P_0}{(1-\rho)^2} \\ &= \frac{\rho P_0 \left(1 - (K+1)\rho^K - \rho + (K+1)\rho^{K+1} + \rho - \rho^{K+1}\right)}{(1-\rho)^2} \\ &= \frac{\rho P_0 \left(1 - (K+1)\rho^K + K\rho^{K+1}\right)}{(1-\rho)^2} \\ &= \frac{\rho \left(1 - (K+1)\rho^K + K\rho^{K+1}\right)}{(1-\rho)(1-\rho^{K+1})}. \end{aligned}$$

- $$E(N^2) = \sum_{k=1}^{K} k^2 P_k, \quad Var(N) = E(N^2) - (E(N))^2$$

- $$\overline{Q} = \sum_{k=1}^{K} (k-1) P_k = \sum_{k=1}^{K} k P_k - \sum_{k=1}^{K} P_k = \overline{N} - U_S$$

- $$E(Q^2) = \sum_{k=1}^{K} (k-1)^2 P_k, \quad Var(Q) = E(Q^2) - (E(Q))^2.$$

- To obtain the distribution of the response and waiting time we have to know the distribution of the system at the moment when the tagged customer enters into to system. It should be underlined that the customer should enter into the system and it is not the same as an arriving customer. An arriving customer can join the system or can be lost because the system is full. By using the Bayes' theorem it is easy to see that

$$\Pi_k = \frac{\lambda P_k}{\sum_{i=0}^{K-1} \lambda P_i} = \frac{P_k}{1 - P_K}.$$

Similarly to the investigations we carried out in an $M/M/1$ system the mean and the density function of the response time can be obtained by the help of the law of total means and law of total probability, respectively.

For the expectation we have

$$\begin{aligned} \overline{T} &= \sum_{k=0}^{K-1} \frac{k+1}{\mu} \Pi_k = \sum_{k=0}^{K-1} \frac{k+1}{\mu} \frac{\rho^k P_0}{1 - P_k} \\ &= \frac{1}{\lambda(1 - P_K)} \sum_{k=0}^{K-1} (k+1) P_{k+1} = \frac{\overline{N}}{\lambda(1 - P_K)}. \end{aligned}$$

Consequently

$$\overline{W} = \overline{T} - \frac{1}{\mu} = \frac{\overline{N}}{\lambda(1 - P_K)} - \frac{1}{\mu}.$$

We would like to show that the Little's law is valid in this case and the same time we can check the correctness of the formula.
It can easily be seen that the average arrival rate into the system is $\overline{\lambda} = \lambda(1 - P_K)$ and thus

$$\overline{\lambda} \cdot \overline{T} = \lambda(1 - P_K) \frac{\overline{N}}{\lambda(1 - P_K)} = \overline{N}.$$

Similarly

$$\begin{aligned} \overline{\lambda} \cdot \overline{W} &= \overline{\lambda} \left(\frac{\overline{N}}{\lambda(1 - P_K)} - \frac{1}{\mu} \right) = \overline{N} - \frac{\overline{\lambda}}{\mu} \\ &= \overline{N} - \rho(1 - P_K) = \overline{N} - U_S = \overline{Q}, \end{aligned}$$

since

$$\overline{\lambda} = \overline{\mu} = \mu U_S.$$

Since the conditional waiting time is Erlang distributed, it is easy to see that

$$E(W^2) = \sum_{k=1}^{K-1} \frac{(k+k^2)}{\mu^2} \Pi_k, \quad Var(W) = E(W^2) - (E(W))^2,$$

$$Var(T) = Var(W) + 1/\mu^2.$$

Now let us find the density function of the response and waiting times
By using the theorem of total probability we have

$$f_T(x) = \sum_{k=0}^{K-1} \mu \frac{(\mu x)^k}{k!} e^{-\mu x} \frac{P_k}{1-P_K},$$

and thus for the distribution function we get

$$\begin{aligned} F_T(x) &= \sum_{k=0}^{K-1} \left(\int_0^x \mu \frac{(\mu t)^k}{k!} e^{-\mu t} \mathrm{d}t \right) \frac{P_k}{1-P_K} \\ &= \sum_{k=0}^{K-1} \left(1 - \sum_{i=0}^{k} \frac{(\mu x)^i}{i!} e^{-\mu x} \right) \frac{P_k}{1-P_K} \\ &= 1 - \sum_{k=0}^{K-1} \left(\sum_{i=0}^{k} \frac{(\mu x)^i}{i!} e^{-\mu x} \right) \frac{P_k}{1-P_K}. \end{aligned}$$

These formulas are more complicated due to the finite summation as in the case of an $M/M/1$ system, but it is not difficult to see that in the limiting case as $K \to \infty$ we have

$$f_T(x) = \mu(1-\rho) e^{-\mu(1-\rho)x}.$$

For the density and distribution function of the waiting time we obtain

$$\begin{aligned} f_W(0) &= \frac{P_0}{1-P_K} \\ f_W(x) &= \sum_{k=1}^{K-1} \mu \frac{(\mu x)^{k-1}}{(k-1)!} e^{-\mu x} \frac{P_k}{1-P_K}, \qquad x > 0 \\ F_W(x) &= \frac{P_0}{1-P_K} + \sum_{k=1}^{K-1} \left(1 - \sum_{i=0}^{k-1} \frac{(\mu x)^i}{i!} e^{-\mu x} \right) \frac{P_k}{1-P_K} \\ &= 1 - \sum_{k=1}^{K-1} \left(\sum_{i=0}^{k-1} \frac{(\mu x)^i}{i!} e^{-\mu x} \right) \cdot \frac{P_k}{1-P_K}. \end{aligned}$$

These formulas can be calculated very easily by a computer.
As we can see the probability P_K plays an important role in the calculations. Notice that it is exactly the probability that an arriving customer find the system full that is it lost. It is called **blocking** or **lost probability** and denoted by P_B. Its correctness can be proved by the help of the Bayes's rule, namely

$$P_B = \lim_{h\to 0} \frac{(\lambda_K h + o(h))P_K}{\sum_{j=0}^{K}(\lambda_j h + o(h))P_j} = \frac{\lambda P_K}{\sum_{j=0}^{K} \lambda P_j} = P_K.$$

If we would like to show the dependence on K and ρ it can be denoted by

$$P_B(K,\rho) = \frac{\rho^K}{\sum_{k=0}^{K} \rho^k}.$$

Notice that

$$P_B(K,\rho) = \frac{\rho\rho^{K-1}}{\sum_{k=0}^{K-1} \rho^k + \rho\rho^{K-1}} = \frac{\rho P_B(K-1,\rho)}{1+\rho P_B(K-1,\rho)}.$$

Starting with the initial value $P_B(1,\rho) = \frac{\rho}{1+\rho}$ the probability of loss can be computed recursively. It is obvious that this sequence tends to 0 as $\rho < 1$. Consequently by using the recursion we can always find an K-t, for which

$$P_B(K,\rho) < P^*,$$

where P^* is a predefined limit value for the probability of loss.

To find the value of K without recursion we have to solve the inequality

$$\frac{\rho^K(1-\rho)}{1-\rho^{K+1}} < P^*$$

which is more complicated task.

Alternatively can can find an approximation method, too. Use the distribution of an $M/M/1$ system and find the probability that in the system there are at least K customers. It is easy to see that

$$P_B(K,\rho) = \frac{\rho^K(1-\rho)}{1-\rho^{K+1}} < \sum_{k=K}^{\infty} \rho^k(1-\rho) = \rho^K,$$

and thus if

$$\rho^K < P^*,$$

then $P_B^*(K,\rho) < P^*$. That is

$$K \ln \rho < \ln P^*$$
$$K > \frac{\ln P^*}{\ln \rho}.$$

Now let us turn our attention to the Laplace-transform of the response and waiting times. First let us compute it for the response time. Similarly to the previous arguments we have

$$\begin{aligned} L_T(s) &= \sum_{k=0}^{K-1} \left(\frac{\mu}{\mu+s}\right)^{k+1} \frac{\rho^k P_0}{1-P_K} \\ &= \frac{P_0}{\rho(1-P_K)} \sum_{l=1}^{K} \left(\frac{\mu\rho}{\mu+s}\right)^{l} \\ &= \frac{P_0}{\rho(1-P_K)} \frac{\lambda}{\mu+s} \frac{1-\left(\frac{\lambda}{\mu+s}\right)^K}{1-\frac{\lambda}{\mu+s}} \\ &= \frac{\mu P_0}{(1-P_K)} \frac{1-\left(\frac{\lambda}{\mu+s}\right)^K}{\mu-\lambda+s}. \end{aligned}$$

The Laplace-transform of the waiting time can be obtained as

$$\begin{aligned} L_W(s) &= \sum_{k=0}^{K-1} \left(\frac{\mu}{\mu+s}\right)^{k} \frac{\rho^k P_0}{1-P_K} \\ &= \frac{P_0}{1-P_K} \sum_{k=0}^{K-1} \left(\frac{\mu\rho}{\mu+s}\right)^{k} \\ &= \frac{P_0}{1-P_K} \frac{1-\left(\frac{\lambda}{\mu+s}\right)^K}{1-\frac{\lambda}{\mu+s}} \\ &= \frac{P_0}{1-P_K} \frac{(\mu+s)\left(1-\left(\frac{\lambda}{\mu+s}\right)^K\right)}{\mu-\lambda+s}, \end{aligned}$$

which also follows from relation

$$L_T(s) = L_W(s) \cdot \frac{\mu}{\mu+s}.$$

By the help of the Laplace-transforms the higher moments of the involved random variables can be computed, too.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Example 12 *Consider the queue at an output port of router. The transmission link is a T1 line (1.544Mbps), packets arrive according to a Poisson process with mean rate* $\lambda = 659.67$ *packets/sec, the packet lengths are exponentially distributed with a mean length of 2048 bits/packet. If the system size is 16 packets what is the packet loss rate?*

Solution:
$\lambda = 659.67$, $\mu = 1.544\ Mbps/2048\ bits/packet = 753.9\ packets/sec$, $\rho = 0.875$.
Thus the packet loss rate = blocking probability x $\lambda = 0.0165\ x\ 659.67 = 10.88$. ■

Example 13 *A data concentrator has 40 terminals connected to it. During the busiest time of day each terminal is occupied and produces packets which are exponentially distributed with a mean of 1000 bits. The link connecting the concentrator to the campus network carries traffic at 1.552 Mbps. The arrival process of packets to the concentrator forms a Poisson process with ten of the terminals producing on average 1 packet per 10 msec, twenty of the terminals producing on average 1 packet per 50 msec, and ten of the terminals producing on average 1 packet per 0.5 second.*
(a) Determine the utilization of the concentrator.
(b) Assuming the buffer at the concentrator is infinite, determine the average delay in the queue.
(c) If the concentrator has a system capacity of 20 packets, determine the packet loss rate.

Solution:
(a) Determine the utilization of the concentrator.
mean service rate $\mu = 1.552 * 10\ bps/1000\ bits/packet\ = 1552\ packets/sec$
mean arrival rate $\lambda = 10 * (1\ packet/10\ msec) + 20 * (1\ packet/50\ msec)$
$+ 10 * (1\ packet/0.5sec)\ =\ 1420\ packets/sec$. Thus $\rho = 1420/1552 = 0.9149$.
(b) Assuming the buffer at the concentrator is infinite, determine the average delay in the queue. $E(W) = 6.93\ msec$
(c) If concentrator has a system capacity of 20 packets, find the packet loss rate.
The system is now modeled as a $M/M/1/K$ queue.
Packet loss rate = the blocking probability * λ = 0.017*1420 = 24.14 packets/sec.
■

2.5 The $M/M/\infty$ Queue

Similarly to the previous systems it is easy to see that the number of customers in the system, that is the process $(N(t), t \geq 0)$ is a birth-death process with rates

$$\lambda_k = \lambda, \quad k = 0, 1, \ldots$$
$$\mu_k = k\mu, \quad k = 1, 2, \ldots.$$

Hence the steady-state distribution can be obtained as

$$P_k = \frac{\varrho^k}{k!} P_0, \text{ where } P_0^{-1} = \sum_{k=0}^{\infty} \frac{\varrho^k}{k!} = e^{\varrho},$$

That is

$$P_k = \frac{\varrho^k}{k!} e^{-\varrho},$$

showing that N follows a Poisson law with parameter ϱ.

It is easy to see that the *performance measures* can be computed as

$$\overline{N} = \varrho, \quad \overline{\lambda} = \lambda, \quad \overline{T} = \frac{1}{\mu}, \quad \overline{W} = 0, \quad \overline{r} = \overline{N}, \quad \overline{\mu} = \overline{r}\mu$$
$$U_r = 1 - e^{-\varrho}, \quad \frac{\mathbb{E}(\delta_r)}{\frac{1}{\lambda}} = \frac{1 - e^{-\varrho}}{e^{-\varrho}}, \quad \mathbb{E}(\delta_r) = \frac{1}{\lambda} \frac{1 - e^{-\varrho}}{e^{-\varrho}}.$$

It can be proved that these formulas remain valid for an $M/G/\infty$ system as well where $\mathbb{E}(S) = \frac{1}{\mu}$.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Agner Krarup Erlang, 1878-1929

2.6 The $M/M/n/n$ Queue, Erlang-Loss System

This system is the oldest and thus the most famous system in queueing theory. The origin of the traffic theory or congestion theory started by the investigation of this system and Erlang was the first who obtained his well-reputed formulas, see for example Erlang [29, 30].
By assumptions customers arrive according to a Poisson process and the service times are exponentially distributed. However, if n servers all busy when a new customer arrives it will be lost because the system is full. The most important question is what proportion of the customers is lost.

The process $(N(t), t \geq 0)$ is said to be in state k if k servers are busy, which is the same as k customers are in the system. It is easy to see that $(N(t), t \geq 0)$is a birth-death process with rates

$$\lambda_k = \begin{cases} \lambda, & \text{if } k < n, \\ 0, & \text{if } k \geq n, \end{cases}$$

$$\mu_k = k\mu, \quad k = 1, 2, ..., n.$$

Clearly the steady-state distribution exists since the process has a finite state space. The

stationary distribution can be obtained as

$$P_k = \begin{cases} P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} & \text{, if } k \leq n, \\ 0 & \text{, if } k > 0. \end{cases}$$

Due to the normalizing condition we have

$$P_0 = \left(\sum_{k=0}^{n} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \right)^{-1},$$

and thus the distribution is

$$P_k = \frac{\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}{\sum_{i=0}^{n} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}} = \frac{\frac{\varrho^k}{k!}}{\sum_{i=0}^{n} \frac{\varrho^i}{i!}} = \frac{\frac{\varrho^k}{k!} e^{-\rho}}{\sum_{i=0}^{n} \frac{\varrho^i}{i!} e^{-\rho}}, \qquad k \leq n.$$

which is called as a **truncated Poisson** distribution with parameter ρ.

The most important measure of the system is

$$P_n = \frac{\frac{\varrho^n}{n!}}{\sum_{k=0}^{n} \frac{\varrho^k}{k!}} = B(n, \rho)$$

which was introduced by Erlang and it is referred to as **Erlang's B-formula, or loss formula** and generally denoted by $B(n, \lambda/\mu)$.

By using the Bayes's rule it is easy to see that

$$B(n, \rho) = \lim_{h \to 0} \frac{(\lambda_n h + o(h)) P_n}{\sum_{j=0}^{n} (\lambda_j h + o(h)) P_j} = \frac{\lambda P_n}{\sum_{j=0}^{n} \lambda P_j} = P_n.$$

For moderate n the probability P_0 can easily be computed. For large n and small ϱ $P_0 \approx e^{-\varrho}$, and thus

$$P_k \approx \frac{\varrho^k}{k!} e^{-\varrho},$$

that is the Poisson distribution. For large n and large ϱ

$$\sum_{j=0}^{n} \frac{\varrho^j}{j!} \neq e^{\varrho}.$$

However, in this case the central limit theorem can be used, since the denominator is the sum of the first $(n+1)$ terms of a Poisson distribution with mean ϱ. Thus by the central

limit theorem this Poisson distribution can be approximated by a normal law with mean ϱ and dispersion $\sqrt{\varrho}$ that is

$$P_n \approx \frac{\Phi(\frac{n+\frac{1}{2}-\varrho}{\sqrt{\varrho}}) - \Phi(\frac{n-1+\frac{1}{2}-\varrho}{\sqrt{\varrho}})}{\Phi(\frac{n+\frac{1}{2}-\varrho}{\sqrt{\varrho}})} = 1 - \frac{\Phi(\frac{n-\frac{1}{2}-\varrho}{\sqrt{\varrho}})}{\Phi(\frac{n+\frac{1}{2}-\varrho}{\sqrt{\varrho}})},$$

where

$$\Phi(s) = \int_{-\infty}^{s} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

is the distribution function of the standard normal distribution.

Another way to calculate $B(n,\rho)$ is to find a recursion. This can be obtained as follows

$$\begin{aligned} B(n,\rho) &= \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^{n} \frac{\rho^i}{i!}} = \frac{\frac{\rho}{n}\frac{\rho^{n-1}}{(n-1)!}}{\sum_{i=0}^{n-1} \frac{\rho^i}{i!} + \frac{\rho}{n}\frac{\rho^{n-1}}{(n-1)!}} \\ &= \frac{\frac{\rho}{n}B(n-1,\rho)}{1+\frac{\rho}{n}B(n-1,\rho)} = \frac{\rho B(n-1,\rho)}{n+\rho B(n-1,\rho)}. \end{aligned}$$

Using $B(1,\rho) = \dfrac{\rho}{1+\rho}$ as an initial value the probabilities $B(n,\rho)$ can be computed for any n. It is important since the direct calculation can cause a problem due to the value of the factorial.
For example for $n = 1000, \rho = 1000$ the exact formula cannot be computed but the approximation and the recursion gives the value 0.024.

Due to the great importance of $B(n,\rho)$ in practical problems so-called calculators have been developed which can be found at

`http://www.erlang.com/calculator/`

To compare the approximations and the exact values please use

`https://qsa.inf.unideb.hu`

Now determine the *main performance measures* of this $M/M/n/n$ system

- *Mean number of customers in the systems, mean number of busy servers*

$$\overline{N} = \overline{n} = \sum_{j=0}^{n} jP_j = \sum_{j=0}^{n} j\frac{\varrho^j}{j!}P_0 = \varrho\sum_{j=0}^{n-1}\frac{\varrho^i}{i!}P_0 = \varrho(1-P_n),$$

  thus the mean number of requests for a given server is

$$\frac{\varrho}{n}(1-P_n).$$

- *Utilization of a server*

  As we have seen
  $$U_s = \sum_{i=1}^{n} \frac{i}{n} P_i = \frac{\bar{n}}{n}.$$
  This case
  $$U_s = \frac{\varrho}{n}(1 - P_n).$$

- *The mean idle period for a given server*

  By applying the well-known relation
  $$P(\text{the server is busy }) = \frac{1/\mu}{\overline{e} + 1/\mu},$$
  where $\overline{e}$ is the mean idle time of the server. Thus
  $$\frac{\varrho}{n}(1 - P_n) = \frac{1/\mu}{\overline{e} + 1/\mu},$$
  hence
  $$\overline{e} = \frac{n}{\lambda(1 - P_n)} - \frac{1}{\mu}.$$

- *The mean busy period of the system*

  Clearly
  $$U_r = 1 - P_0 = \frac{E\delta_r}{\frac{1}{\lambda} + E\delta_r},$$
  thus
  $$E\delta_r = \frac{1 - P_0}{\lambda P_0} = \frac{\sum_{i=1}^{n} \frac{\varrho^i}{i!}}{\lambda\left(1 + \sum_{i=1}^{n} \frac{\varrho^i}{i!}\right)}.$$

It can be proved that these formulas **remain valid** for an $M/G/n/n$ system as well where $\mathbb{E}(S) = \dfrac{1}{\mu}$.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Example 14 *In busy parking lot cars arrive according to a Poisson process one in* 20 *seconds and stay there in the average of* 10 *minutes.*
How many parking places are required if the probability of a loss is no to exceed 1% ?

Solution:

$$\rho = \frac{\lambda}{\mu} = \frac{10}{\frac{1}{3}} = 30, P_n = 0.01.$$

Following a normal approximation

$$P_n = 0.01 = \frac{\frac{\rho^n}{n!}e^{-\rho}}{\Phi\left(\frac{n+\frac{1}{2}-\rho}{\sqrt{\rho}}\right)} = \frac{\Phi\left(\frac{n+\frac{1}{2}-\rho}{\sqrt{\rho}}\right) - \Phi\left(\frac{n-\frac{1}{2}-\rho}{\sqrt{\rho}}\right)}{\Phi\left(\frac{n+\frac{1}{2}-\rho}{\sqrt{\rho}}\right)}.$$

Thus

$$0.99\Phi\left(\frac{n+\frac{1}{2}-\rho}{\sqrt{\rho}}\right) = \Phi\left(\frac{n-\frac{1}{2}-\rho}{\sqrt{\rho}}\right).$$

It is not difficult to verify by using the Table for the standard normal distribution that $n = 41$.

Thus the approximation value of P_{41} is 0.009917321712214377,
and the exact value is 0.01043318100246811.

■

Example 15 *A telephone exchange consists of* 50 *lines and calls arrive according to a Poisson process, the mean interarrival time is* 10 *minutes. The mean service time is* 5 *minutes.*
Find the main performance measures.

Solution:
Using Poisson approximation where $\rho = \frac{\lambda}{\mu} = 0.5$

$P_{50} = 0.00000$, event for $n = 6$

$P_6 = 0,00001$. This means that a call is almost never lost.

Mean number of busy lines can be obtain as

$$\overline{n} = \rho(1 - P_n) = \rho = 0.5 \ ,$$

The utilization of a line is

$$\frac{0.5}{50} = \frac{5 \times 10^{-1}}{5 \times 10} = 10^{-2}$$

The utilization of the system is

$$U_r = 1 - 0.606 = 0.394$$

The mean busy period of the system can be obtained as

$$E\delta_r = \frac{(1 - P_0)}{(\lambda P_0)} = \frac{0.394}{2 \times 0.606} = \frac{0.394}{1.212} = 0.32 \text{ minutes}$$

Mean idle period of a line is

$$\overline{e} = \frac{n}{\lambda(1 - P_n)} - \frac{\rho}{\lambda} = \frac{50}{2(1-0)} - \frac{0,5}{2} = 25 - \frac{1}{4} = 24.75 \text{ minutes}$$

■

Heterogeneous Servers

In the case of an $M/\overrightarrow{M}/n/n$ system the service time distribution depends on the index of the server. That is the service time is exponentially distributed with parameter μ_i for server i. An arriving customer choose randomly among the idle servers, that is each idle server is chosen with the same probability. Since the servers are heterogeneous it is not enough to to the number of busy servers but we have to identify them by their index. It means that we have to deal with general Markov-processes.

Let $(i_1, \ldots, i_k)$ denote the indexes of the busy servers, which are the combinations of n objects taken k at a time without replacement. Thus the state space of the Markov-chain is the set of these combinations, that is $(0, (i_1, \ldots, i_k) \in C_k^n, k = 1, \ldots, n)$.

Let us denote by

$$\begin{aligned} P_0 &= P(0), \\ P(i_1, \ldots, i_k) &= P((i_1, \ldots, i_k)), (i_1, \ldots, i_k) \in C_k^n, \quad k = 1, \ldots, n \end{aligned}$$

the steady-state distribution of the chain which exists since the chain has a finite state space and it is irreducible. The set of steady-state balance equations can be written as

$$\lambda P_0 = \sum_{j=1}^{n} \mu_j P(j) \tag{2.26}$$

$$\begin{aligned} (\lambda + \sum_{j=1}^{k} \mu_{i_j}) P(i_1, \ldots, i_k) = \frac{\lambda}{n-k+1} \sum_{j=1}^{k} P(i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_k) \\ + \sum_{j \neq i_1, \ldots, i_k} \mu_j P(i'_1, \ldots, i'_k, j') \end{aligned} \tag{2.27}$$

$$\left(\sum_{j=1}^{n} \mu_j \right) P(1, \ldots, n) = \lambda \sum_{j=1}^{n} P(1, \ldots, j-1, j+1, \ldots, n) \tag{2.28}$$

where $(i'_1, \ldots, i'_k, j')$ denotes the ordered set $i_1, \ldots, i_k, j$, i_{-1} and i_{n+1} are not defined. Despite of the large number of unknowns, which is 2^n, the solution is quite simple, namely

$$P(i_1, \ldots, i_k) = (n-k)! \prod_{j=1}^{k} \varrho_{i_j} C, \tag{2.29}$$

where $\varrho_j = \dfrac{\lambda}{\mu_i}, \quad j = 1, \ldots, n, \quad P_0 = n!C$, which can be determined by the help of the normalizing condition

$$P_0 + \sum_{k=1}^{n} \sum_{(i_1, \ldots, i_k) \in C_k^n} P(i_1, \ldots, i_k) = 1.$$

Let us check the first equation (2.26). By substitution we have

$$\lambda n!C = \sum_{j=1}^{n} \mu_j \frac{\lambda}{\mu_j}(n-1)!C = n!\lambda C.$$

Lets us check now the third equation (2.28)

$$\left(\sum_{j=1}^{n} \mu_j\right)\frac{\lambda^n}{\mu_1 \cdots \mu_n}C = \lambda \sum_{j=1}^{n} \frac{\lambda^{n-1}C}{\mu_1 \cdots \mu_{j-1}\mu_{j+1} \cdots \mu_n} = \frac{\lambda^n}{\mu_1 \cdots \mu_n}\left(\sum_{j=1}^{n} \mu_j\right)C.$$

Finally, let us check the most complicated one, the second set of equations (2.27), namely

$$\begin{aligned}
&(\lambda + \sum_{j=1}^{k} \mu_{i_j})(n-k)! \prod_{j=1}^{k} \varrho_{i_j} C \\
&= \frac{\lambda}{n-k+1}(n-k+1)! \sum_{j=1}^{k} \frac{\lambda^{k-1}C}{\mu_{i_1} \cdots \mu_{i_{j-1}}\mu_{i_{j+1}} \cdots \mu_{i_k}} \\
&+ \sum_{j \neq i_1,\ldots,i_k} (n-k-1)! \frac{\lambda^{k+1}\mu_j C}{\mu_{i_1} \cdots \mu_{i_k}\mu_j} \\
&= (n-k)! \sum_{j=1}^{k} \frac{\mu_{i_j}\lambda^k C}{\mu_{i_1} \cdots \mu_{i_k}} + \lambda \sum_{j \neq i_1,\ldots,i_k} (n-k-1)! \frac{\lambda^k C}{\mu_{i_1} \cdots \mu_{i_k}} \\
&= (n-k)! \left(\sum_{j=1}^{k} \mu_{i_j}\right) \frac{\lambda^k C}{\mu_{i_1} \cdots \mu_{i_k}} + \lambda(n-k)! \frac{\lambda^k C}{\mu_{i_1} \cdots \mu_{i_k}},
\end{aligned}$$

which shows the equality.

Thus the usual *performance measures* can be obtained as

- the utilization of the jth server U_j can be calculated as

$$U_j = \sum_{k=1}^{n} \sum_{j \in (i_1,\ldots,i_k)} P(i_1, \ldots, i_k),$$

  and thus

$$U_j = \frac{\frac{1}{\mu_j}}{\frac{1}{\mu_j} + \mathbb{E}(e_j)},$$

  where $\mathbb{E}(e_j)$ is the mean idle period of the jth server. Hence

$$\mathbb{E}(e_j) = \frac{1}{\mu_j}\frac{1-U_j}{U_j}.$$

- $\overline{N} = \sum_{j=1}^{n} U_j$
- The probability of loss is $P_B = P(1, \ldots, n)$.

It should be noted that in this case the following relation also holds

$$\lambda(1 - P_B) = \sum_{j=1}^{n} U_j \mu_j.$$

In homogeneous case, that is when $\mu_j = \mu, j = 1, \ldots, n$, after substitution we have

$$P_k = \sum_{(i_1,\ldots,i_k)\in C_k^n} P(i_1, \ldots, i_k) = \binom{n}{k}(n-k)!\varrho^k C = \frac{\varrho^k}{k!} n! C = \frac{\varrho^k}{k!} P_0 = \frac{\frac{\varrho^k}{k!}}{\sum_{j=0}^{n} \frac{\varrho^j}{j!}},$$

that is it reduces to the Erlang's formula derived earlier.

It should be noted that these formulas remains valid under generally distributed service times with finite means with $\rho_i = \lambda E(S_i)$. In other words the Erlang's loss formula is robust to the distribution of the service time, it does not depend on the distribution itself but only on its mean.

2.7 The $M/M/n$ Queue

It is a variation of the classical queue assuming that the service is provided by n servers operating independently of each other. This modification is natural since if the mean arrival rate is greater than the service rate the system will not be stable, that is why the number of servers should be increased. However, in this situation we have parallel services and we are interested in the distribution of first service completion.
That is why we need the following observation.
Let X_i be exponentially distributed random variables with parameter μ_i, $(i = 1, 2, ..., r)$ and denote by Y their minimum. It is not difficult to see that Y is also exponentially distributed with parameter $\sum_{i=1}^{r} \mu_i$ since

$$P(Y < x) = 1 - P(Y \geq x) = 1 - P(X_i \geq x,\ i = 1, ..., r) =$$

$$= 1 - \prod_{i=1}^{r} P(X_i \geq x) = 1 - e^{-\left(\sum_{i=1}^{r} \mu_i\right)x}.$$

Similarly to the earlier investigations, it can easily be verified that the number of customers in the system is a birth-death process with the following transition probabilities

$$P_{k,k-1}(h) = (1 - (\lambda h + o(h)))\,(\mu_k h + o(h)) + o(h) = \mu_k h + o(h),$$
$$P_{k,k+1}(h) = (\lambda h + o(h))\,(1 - (\mu_k h + o(h))) + o(h) = \lambda h + o(h),$$

where

$$\mu_k = \min(k\mu, n\mu) = \begin{cases} k\mu & \text{, for } 0 \le k \le n, \\ \\ n\mu & \text{, for } n < k. \end{cases}$$

It is understandable that the stability condition is $\lambda/n\mu < 1$.

To obtain the distribution P_k we have to distinguish two cases according to as μ_k depends on k. Thus if $k < n$, then we get

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}.$$

Similarly, if $k \ge n$, then we have

$$P_k = P_0 \prod_{i=0}^{n-1} \frac{\lambda}{(i+1)\mu} \prod_{j=n}^{k-1} \frac{\lambda}{n\mu} = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{n! n^{k-n}}.$$

In summary

$$P_k = \begin{cases} P_0 \dfrac{\rho^k}{k!} & \text{, for } k \le n, \\ \\ P_0 \dfrac{a^k n^n}{n!} & \text{, for } k > n, \end{cases}$$

where

$$a = \frac{\lambda}{n\mu} = \frac{\rho}{n} < 1.$$

This a is exactly the $u$tilization of a given server . Furthermore

$$P_0 = \left(1 + \sum_{k=1}^{n-1} \frac{\rho^k}{k!} + \sum_{k=n}^{\infty} \frac{\rho^k}{n!} \frac{1}{n^{k-n}}\right)^{-1},$$

and thus

$$P_0 = \left(\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n}{n!} \frac{1}{1-a}\right)^{-1}.$$

Since the arrivals follow a Poisson law the the distribution of the system at arrival instants equals to the distribution at random moments, hence the probability that an arriving customer has to wait is

$$P(\text{waiting}) = \sum_{k=n}^{\infty} P_k = \sum_{k=n}^{\infty} P_0 \frac{\rho^k}{n!} \frac{1}{n^{k-n}}.$$

that is it can be written as

$$P(\text{waiting}) = \frac{\dfrac{\rho^n}{n!} \dfrac{1}{1-a}}{\displaystyle\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n}{n!} \frac{1}{1-a}} = \frac{\frac{\rho^n}{n!} \frac{n}{n-\rho}}{\displaystyle\sum_{k=0}^{n-1} \frac{\rho^k}{k!} + \frac{\rho^n n}{n!(n-\rho)}} = C(n, \rho).$$

This probability is frequently used in different practical problems, for example in telephone systems, call centers, just to mention some of them. It is also a very famous formula which is referred to as **Erlang's C formula**,or **Erlang's delay formula** and it is denoted by $C(n, \lambda/\mu)$.

The main *performance measures* of the systems can be obtained as follows

- *For the mean queue length we have*

$$\overline{Q} = \sum_{k=n}^{\infty}(k-n)P_k = \sum_{j=0}^{\infty} jP_{n+j} = \sum_{j=0}^{\infty} \frac{j\left(\frac{\lambda}{\mu}\right)^{n+j}}{n!n^j}P_0 =$$

$$= \sum_{j=0}^{\infty} j\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}a^j P_0 = P_0\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}a\sum_{j=0}^{\infty}\frac{da^j}{da} = P_0\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}a\frac{d}{da}\sum_{j=0}^{\infty}a^j =$$

$$= P_0\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}\frac{a}{(1-a)^2} = \frac{\rho}{n-\rho}C(n,\rho).$$

- For the *m*ean number of busy servers we obtain

$$\overline{n} = \sum_{k=0}^{n-1} kP_k + \sum_{k=n}^{\infty} nP_k = P_0\left(\rho\sum_{k=0}^{n-2}\frac{\rho^k}{k!} + \frac{\rho^n}{(n-1)!}\frac{1}{1-a}\right) =$$

$$= \rho\left(\sum_{k=0}^{n-2}\frac{\rho^k}{k!} + \frac{\rho^{n-1}}{(n-1)!} + \frac{\rho^{n-1}}{(n-1)!}\left(\frac{1}{1-a}-1\right)\right)P_0 =$$

$$= \rho\left(\sum_{k=0}^{n-1}\frac{\rho^k}{k!} + \frac{\rho^n}{n!}\frac{1}{1-a}\right)P_0 = \rho\frac{1}{p_0}P_0 = \rho.$$

- *For the mean number of customers in the system we get*

$$\overline{N} = \sum_{k=0}^{\infty} kP_k = \sum_{k=0}^{n-1} kP_k + \sum_{k=n}^{\infty}(k-n)P_k + \sum_{k=n}^{\infty} nP_k = \overline{n} + \overline{Q}$$
$$= \rho + \frac{\rho}{n-\rho}C(n,\rho),$$

  which is understandable since a customer is either in the queue or in service. Let us denote by $\overline{S}$-gal the mean number of idle servers. Then it is easy to see that

$$\overline{n} = n - \overline{S},$$
$$\overline{S} = n - \frac{\lambda}{\mu},$$

thus

$$\overline{N} = n - \overline{S} + \overline{Q},$$

hence

$$\overline{N} - n = \overline{Q} - \overline{S}.$$

- *Distribution of the waiting time*

An arriving customer has to wait if at his arrival the number of customers in the system is at least n. In this case the time while a customer is serviced is exponentially distributed with parameter $n\mu$, consequently if there $n+j$ customers in the system the waiting time is Erlang distributed with parameters $(j+1, n\mu)$. By applying the theorem of total probability for the density function of the waiting time we have

$$f_W(x) = \sum_{j=0}^{\infty} P_{n+j} (n\mu)^{j+1} \frac{x^j}{j!} e^{-n\mu x}.$$

Substituting the distribution we get

$$\begin{aligned}
f_W(x) &= \sum_{j=0}^{\infty} P_0 \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} a^j (n\mu)^{j+1} \frac{x^j}{j!} e^{-n\mu x} \\
&= \frac{P_0 \left(\frac{\lambda}{\mu}\right)^n}{n!} n\mu e^{-n\mu x} \sum_{j=0}^{\infty} \frac{(an\mu x)^j}{j!} \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 n\mu e^{-(n\mu - \lambda)x} \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 n\mu e^{-n\mu(1-a)x} \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 \frac{1}{1-a} n\mu(1-a) e^{-n\mu(1-a)x} \\
&= P(\text{waiting}) n\mu(1-a) e^{-n\mu(1-a)x}.
\end{aligned}$$

Hence for the complement of the distribution function we obtain

$$\begin{aligned}
P(W > x) &= \int_x^{\infty} f_W(u) du = P(\text{waiting}) e^{-n\mu(1-a)x} \\
&= C(n, \rho) \cdot e^{-\mu(n-\rho)x}.
\end{aligned}$$

Therefore the distribution function can be written as

$$\begin{aligned}
F_W(x) &= 1 - P(\text{waiting}) + P(\text{waiting}) \left(1 - e^{-n\mu(1-a)x}\right) \\
&= 1 - P(\text{waiting}) e^{-n\mu(1-a)x} = 1 - C(n, \rho) \cdot e^{-\mu(n-\rho)x}.
\end{aligned}$$

Consequently the mean waiting time can be calculated as

$$\overline{W} = \int_0^{\infty} x f_W(x) dx = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 \frac{1}{(1-a)^2 n\mu} = \frac{1}{\mu(n-\rho)} C(n, \rho).$$

It is not difficult to see that

$$Var(W) = \frac{C(n,\rho)(2 - C(n,\rho))}{(\mu(n-\rho))^2}.$$

- *Distribution of the response time*

The service immediately starts if at arrival the number of customer in the system is than n. However, if the arriving customer has to wait then the response time is the sum of this waiting and service times. By applying the law of total probability for the density function of the response time we get

$$f_T(x) = P(\text{no waiting})\mu e^{-\mu x} + f_{W+S}(x)$$

As we have proved

$$f_W(x) = P(\text{waiting})e^{-n\mu(1-a)x} n\mu(1-a).$$

Thus

$$f_{W+S}(z) = \int_0^z f_W(x)\mu e^{-\mu(z-x)}dx =$$

$$= P(\text{waiting})n\mu(1-a)\mu \int_0^z e^{-n\mu(1-a)x} e^{-\mu(z-x)}dx =$$

$$= \frac{\rho^n}{n!}P_0 \frac{1}{(1-a)} n\mu(1-a)\mu e^{-z\mu} \int_0^z e^{-\mu(n-1-\lambda/\mu)x}dx =$$

$$= \frac{\rho^n}{n!}P_0 n\mu \frac{1}{n-1-\lambda/\mu} e^{-\mu z}\left(1 - e^{-\mu(n-1-\lambda/\mu)z}\right).$$

Therefore

$$f_T(x) = \left(1 - \left(\frac{\lambda}{\mu}\right)^n \frac{P_0}{n!(1-a)}\right)\mu e^{-\mu x} +$$

$$+ \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} n\mu P_0 \frac{1}{n-1-\lambda/\mu} e^{-\mu x}\left(1 - e^{-\mu(n-1-\lambda/\mu)x}\right) =$$

$$= \mu e^{-\mu x}\left(1 - \frac{\left(\frac{\lambda}{\mu}\right)^n P_0}{n!(1-a)} + \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} nP_0 \frac{1}{n-1-\lambda/\mu}\left(1 - e^{-\mu(n-1-\lambda/\mu)x}\right)\right) =$$

$$= \mu e^{-\mu x}\left(1 + \frac{\left(\frac{\lambda}{\mu}\right)^n P_0}{n!(1-a)} \frac{1 - (n-\lambda/\mu)e^{-\mu(n-1-\lambda/\mu)x}}{n-1-\lambda/\mu}\right).$$

Consequently for the complement of the distribution function of the response time we have

$$P(T > x) = \int_x^\infty f_T(y)dy =$$

$$= \int_x^\infty \mu e^{-\mu y} + \frac{\left(\frac{\lambda}{\mu}\right)^n P_0}{n!(1-a)} \frac{1}{n-1-\lambda/\mu} \left(\mu e^{-\mu y} - \mu(n-\lambda/\mu) e^{-\mu(n-\lambda/\mu)y} \right) dy =$$

$$= e^{-\mu x} + \left(\frac{\lambda}{\mu}\right)^n P_0 \frac{1}{n!(1-a)(n-1-\lambda/\mu)} \left(e^{-\mu x} - e^{-\mu(n-\lambda/\mu)x} \right) =$$

$$= e^{-\mu x} \left(1 + \frac{\left(\frac{\lambda}{\mu}\right)^n P_0}{n!(1-a)} \frac{1 - e^{-\mu(n-1-\lambda/\mu)x}}{n-1-\lambda/\mu} \right).$$

Thus the distribution function can be written as

$$F_T(x) = 1 - P(T > x).$$

In addition for the mean response time we obtain

$$\overline{T} = \int_0^\infty x f_T(x) dx = \frac{1}{\mu} + \frac{1}{n\mu} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 \frac{1}{(1-a)^2} = \frac{1}{\mu} + \overline{W},$$

as it was expected.

In stationary case the mean number of arriving customer should be equal to the mean number of departing customers, so the mean number of customer in the system is equal to the number of customers arrived during a mean response time. That is

$$\lambda \overline{T} = \overline{N} = \overline{Q} + \overline{n},$$

in addition

$$\lambda \overline{W} = \overline{Q}.$$

These are the **Little's formulas**, that can be proved by simple calculations. As we have seen

$$\overline{N} = \rho + P_0 \frac{\rho^n}{n!(1-a)^2} a.$$

Since

$$\overline{T} = \frac{1}{\mu} + \frac{1}{n\mu} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 \frac{1}{(1-a)^2},$$

thus

$$\lambda \overline{T} = \frac{\lambda}{\mu} + \frac{\rho^n}{n!} P_0 \frac{a}{(1-a)^2},$$

that is

$$\overline{N} = \lambda \overline{T},$$

because $\dfrac{\lambda}{\mu} = \rho$.

Furthermore

$$\overline{Q} = \lambda \overline{W},$$

since

$$\overline{n} = \rho.$$

- *Overall utilization of the servers* can be obtained as

  The utilization of a single server is

$$U_s = \sum_{k=1}^{n-1} \frac{k}{n} P_k + \sum_{k=n}^{\infty} P_k = \frac{\bar{n}}{n} = a.$$

  Hence the overall utilization can be written as

$$U_n = nU_s = \bar{n}.$$

- *The mean busy period of the system* can be computed as

  The system is said to be idle if the is no customer in the system, otherwise the system is busy. Let $E\delta_r$ denote the mean busy period of the system. Then the utilization of the system is

$$U_r = 1 - P_0 = \frac{E\delta_r}{\frac{1}{\lambda} + E\delta_r},$$

  thus

$$E\delta_r = \frac{1 - P_0}{\lambda P_0}.$$

  If the individual servers are considered then we assume that a given server becomes busy earlier if it became idle earlier. Hence if $j < n$ customers are in the system then the number of idle servers is $n - j$.

  Let as consider a given server. On the condition that at the instant when it became idle the number of customers in the system was j its mean idle time is

$$\overline{e}_j = \frac{n - j}{\lambda}.$$

  The probability of this situation is

$$a_j = \frac{P_j}{\sum_{i=0}^{n-1} P_i}.$$

  Then applying the law of total expectations for its mean idle period we have

$$\overline{e} = \sum_{j=0}^{n-1} a_j \overline{e}_j = \sum_{j=0}^{n-1} \frac{(n-j)P_j}{\lambda \sum_{i=0}^{n-1} P_i} = \frac{\overline{S}}{\lambda P(e)},$$

where $P(e) = 1 - C(n, \rho)$ denotes the probability that an arriving customer find an idle server.

Since

$$U_s = a = \frac{E\delta}{\overline{e} + E\delta},$$

thus

$$a\overline{e} = (1 - a)E\delta,$$

where $E\delta$ denotes its mean busy period.

Hence

$$E\delta = \frac{a}{1 - a}\frac{\overline{S}}{\lambda P(e)}.$$

In the case of $n = 1$ it reduces to

$$\overline{S} = 1 - a, \qquad P(e) = P_0 = 1 - a, \qquad a = \frac{\lambda}{\mu},$$

thus

$$E\delta = \frac{1}{\mu - \lambda},$$

which was obtained earlier.

In the following we are going to show what is the connection between these two famous Erlang's formulas. Namely, first we prove how the delay formula can be expressed by the help of loss formula, that is

$$C\left(m, \frac{\lambda}{\mu}\right) = \frac{(\frac{\lambda}{\mu})^m}{m!}\frac{1}{1 - \frac{\lambda}{m\mu}}\frac{1}{\sum_{k=0}^{m-1}\frac{(\frac{\lambda}{\mu})^k}{k!} + \frac{(\frac{\lambda}{\mu})^m}{m!}\frac{1}{1-\frac{\lambda}{m\mu}}} = \frac{\frac{(\frac{\lambda}{\mu})^m}{m!}}{\sum_{k=0}^{m-1}\frac{(\frac{\lambda}{\mu})^m}{m!}(1 - \frac{\lambda}{m\mu}) + \frac{(\frac{\lambda}{\mu})^m}{m!}}$$
$$= \frac{B(m, \frac{\lambda}{\mu})}{(1 - B(m, \frac{\lambda}{\mu}))(1 - \frac{\lambda}{m\mu}) + B(m, \frac{\lambda}{\mu})} = \frac{B(m, \frac{\lambda}{\mu})}{1 - \frac{\lambda}{m\mu}(1 - B(m, \frac{\lambda}{\mu}))}.$$

As we have seen in the previous investigations the delay probability $C(n, \rho)$, plays an important role in determining the main performance measures. Notice that the above formula can be rewritten as

$$C(n, \rho) = \frac{nB(n, \rho)}{n - \rho + \rho B(n, \rho)} > B(n, \rho),$$

moreover it can be proved that there exists a recursion for it, namely

$$C(n, \rho) = \frac{\rho(n - 1 - \rho) \cdot C(n - 1, \rho)}{(n - 1)(n - \rho) - \rho C(n - 1, \rho)},$$

starting with the value $C(1,\rho)=\rho$.

If the quality of service parameter is $C(n,\rho)$ then it is easy to see that there exists an olyan n_α^*, for which $C(n_\alpha^*,\rho)<\alpha$. This n_α^* can easily be calculated by a computer using the above recursion.

Let us show another method for calculating this value. As we have seen earlier the probability of loss can be approximated as

$$B(n,\rho)\approx\frac{\varphi\left(\frac{n-\rho}{\sqrt{\rho}}\right)}{\sqrt{\rho}\phi\left(\frac{n-\rho}{\sqrt{\rho}}\right)}.$$

Let $k=\frac{n-\rho}{\sqrt{\rho}}$, thus $n=\rho+\sqrt{\rho}k$. Hence

$$\begin{aligned}C(n,\rho)=\frac{nB(n,\rho)}{n-\rho+\rho B(n,\rho)}&\approx\frac{(\rho+k\sqrt{\rho})\frac{\varphi(k)}{\sqrt{\rho}\phi(k)}}{\rho+k\sqrt{k}-\rho+\rho\frac{\varphi(k)}{\sqrt{\rho}\phi(k)}}\\&\approx\frac{\sqrt{\rho}\frac{\varphi(k)}{\phi(k)}}{\sqrt{\rho}\left(k+\frac{\varphi(k)}{\phi(k)}\right)}=\left(1+k\frac{\phi(k)}{\varphi(k)}\right)^{-1}.\end{aligned}$$

That is if we would like to find such an n_α^* for which $C(n_\alpha^*,\rho)<\alpha$, then we have to solve the following equation

$$\left(1+k_\alpha\frac{\phi(k_\alpha)}{\varphi(k_\alpha)}\right)^{-1}\approx\alpha$$

which can be rewritten as

$$k_\alpha\frac{\phi(k_\alpha)}{\varphi(k_\alpha)}=\frac{1-\alpha}{\alpha}$$

If k_α is given then

$$n_\alpha^*=\rho+k_\alpha\sqrt{\rho}.$$

It should be noted that the search for k_α is independent of the value of ρ and n thus it can be calculated for various values of α.

For example, if $\alpha=0.8,0.5,0.2,0.1$,
then the corresponding k_α-as are $0.1728,0.5061,1.062,1.420$.

The formula $n_\alpha^*=\rho+k_\alpha\sqrt{\rho}$ is called as **square-root staffing rule**. As we can see in the following Table it gives a very good approximation, see Tijms [117].
Let us see an example for illustration.
Let us consider two service centers which operate separately. Then using this rule overall we have to use $2(\rho+k_\alpha\sqrt{\rho})$ servers. However, if we have a joint queue to get the same service level we should use $2\rho+k_\alpha\sqrt{2\rho}$ servers. The reduction is $(2-\sqrt{2})k_\alpha\sqrt{\rho}$, that is

Table 2.1: Exact and approximated values of n^*

	$\alpha = 0.5$		$\alpha = 0.2$		$\alpha = 0.1$	
	exact	approximation	exact	approximation	exact	approximation
$\rho = 1$	2	2	3	3	3	3
$\rho = 5$	7	7	8	8	9	9
$\rho = 10$	12	12	14	14	16	15
$\rho = 50$	54	54	58	58	61	61
$\rho = 100$	106	106	111	111	115	115
$\rho = 250$	259	259	268	267	274	273
$\rho = 500$	512	512	525	524	533	532
$\rho = 1000$	1017	1017	1034	1034	1046	1045

the reason that the joint queue is used in practice.

$C(n, \rho)$ is of great importance in practical problems hence so-called calculators have been developed and can be used at the link

`http://www.erlang.com/calculator/`

Separated $M/M/1$ and common queue $M/M/2$ systems

$$C(2,\rho) = \frac{\rho^2}{2-\rho}\frac{1}{1+\rho+\frac{\rho^2}{2-\rho}} == \frac{\rho^2}{2-\rho+2\rho-\rho^2+\rho^2} = \frac{\rho^2}{2+\rho}.$$

Thus

$$\overline{Q} = \frac{\rho}{2-\rho}\frac{\rho^2}{2+\rho}, \qquad \overline{W} = \frac{\rho^2}{\mu(4-\rho^2)}.$$

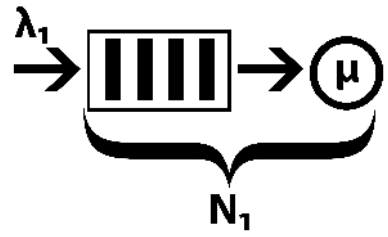
Therefore

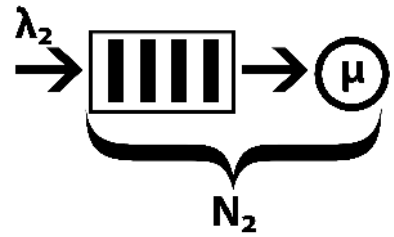
$$\overline{N} = \rho + \overline{Q} = \frac{\rho(4-\rho^2)+\rho^3}{4-\rho^2} = \frac{4\rho}{4-\rho^2} =$$
$$= \frac{\rho}{1-\left(\frac{\rho}{2}\right)^2}.$$

Thus by using the Little formula we have

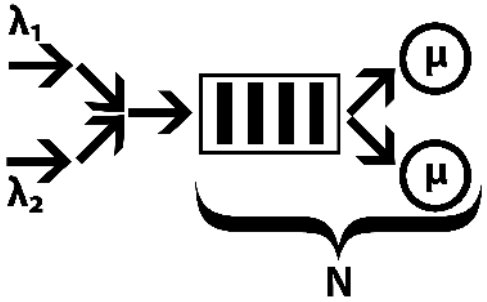
$$\overline{T} = \frac{1}{\mu}\frac{1}{1-\left(\frac{\rho}{2}\right)^2} = \frac{4}{\mu(4-\rho^2)}.$$

Example 16 *Let us consider 2 separated $M/M/1$ queues with λ_1, λ_2 arrival intensities and with the same service intensity μ. Of course $\lambda_1 < \mu$, $\lambda_2 < \mu$. Aggregate the arrival processes and consider a 2 server system with service μ intensities at each server. Assume that $\lambda_1 \geq \lambda_2$.*





Total number of customers in the aggregated system is $N_1 + N_2$.



1. *Show that*
   $\overline{T} < \overline{T_1}$

2. *Find the condition that implies*
   $\overline{T} < \overline{T_2}$,

where $\overline{T_1}, \overline{T_2}$ are the mean response times for the separated queues and $\overline{T}$ denotes the mean response time for the $M/M/2$ system.

Solution:

Obvious that

$$\overline{T_2} = \frac{1}{\mu - \lambda_2} \leq \overline{T_1} = \frac{1}{\mu - \lambda_1},$$

$$\overline{T} = \frac{1}{\mu\left(1 - \left(\frac{\lambda_1+\lambda_2}{2\mu}\right)^2\right)}.$$

First prove that

$$\frac{1}{\mu\left(1 - \left(\frac{\lambda_1+\lambda_2}{2\mu}\right)^2\right)} < \frac{1}{\mu - \lambda_1},$$

$$\frac{4\mu^2}{\mu\left(4\mu^2 - (\lambda_1+\lambda_2)^2\right)} < \frac{1}{\mu - \lambda_1},$$

$$4(\mu - \lambda_1)\mu < 4\mu^2 - (\lambda_1+\lambda_2)^2,$$

$$(\lambda_1+\lambda_2)^2 < 4\lambda_1\mu,$$

$$(\lambda_1+\lambda_2)^2 < (2\lambda_1)^2 < 4\lambda_1\lambda_1 < 4\lambda_1\mu.$$

since $\lambda_1 \geq \lambda_2, \lambda_1 < \mu$.

Similarly,

$$\overline{T} < \frac{1}{\mu - \lambda_2} \qquad iff$$

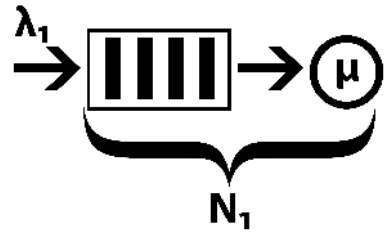
$$(\lambda_1+\lambda_2)^2 < 4\lambda_2\mu.$$

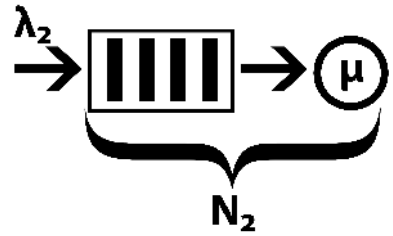
If $\lambda_1 = \lambda_2 = \lambda$ then $(2\lambda)^2 < 4\lambda\mu$.

which is valid since$\lambda < \mu$. ■

Separated and common queues

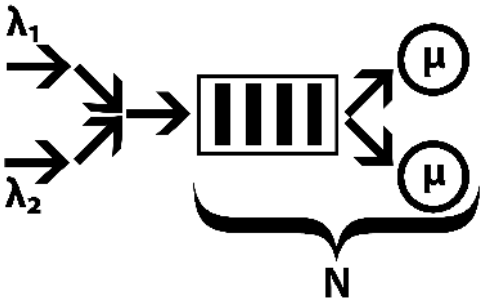
Separated queues





Total number of customers in the aggregated system is $N_1 + N_2$.

Common queue



Show that $\overline{N} < \overline{N_1} + \overline{N_2}$.

We have to prove that

$$\frac{\rho_1 + \rho_2}{1 - \left(\frac{\rho_1+\rho_2}{2}\right)^2} < \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2},$$

$$\frac{4(\rho_1 + \rho_2)}{4 - (\rho_1 + \rho_2)^2} < \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2},$$

$$4\Big(\rho_1^2(\rho_2 - 1) + \rho_2^2(\rho_1 - 1)\Big) <$$
$$(\rho_1 + \rho_2)\Big(\rho_1^2(2\rho_2 - 1) + \rho_2^2(2\rho_1 - 1)\Big).$$

After arrangement we get

$$\rho_1^2\Big[(2\rho_2-1)(\rho_1+\rho_2)+4(1-\rho_2)\Big]+$$
$$+\rho_2^2\Big[(2\rho_1-1)(\rho_1+\rho_2)+4(1-\rho_1)\Big]>0.$$

We show that

$$(2\rho_i-1)(\rho_1+\rho_2)+4(1-\rho_i)>0, \qquad i=1,2.$$

That is

$$(\rho_1+\rho_2)(1-2\rho_i)<4(1-\rho_i).$$

It is easy to see that

$$(\rho_1+\rho_2)(1-2\rho_i)<(\rho_1+\rho_2)(1-\rho_i)$$
$$<2(1-\rho_i)<4(1-\rho_i), \qquad i=1,2.$$

From this the statement follows.

If $\lambda_1=\lambda_2=\lambda$ then $\rho<1$, furthermore for the aggregated $M/M/2$ and the combined separated $M/M/1$ systems we get

$$\overline{N}=\frac{2\rho}{1-\rho^2}, \quad \overline{N_1+N_2}=\frac{2\rho}{1-\rho}.$$

That is

$$\frac{2\rho}{1-\rho^2}=\frac{2\rho}{(1-\rho)(1+\rho)}<\frac{2\rho}{1-\rho}.$$

Hence

$$\frac{\overline{N}}{\overline{N_1+N_2}}=\frac{\frac{2\rho}{1-\rho^2}}{\frac{2\rho}{1-\rho}}=\frac{1}{1+\rho}>\frac{1}{2}.$$

In other form and by using the Little-formula we get

$$\overline{N}=\frac{1}{1+\rho}\overline{N_1+N_2}=2\lambda\overline{T}.$$

Thus

$$\overline{T}=\frac{1}{\mu(1-\rho^2)}=\frac{1}{\mu(1-\rho)(1+\rho)}.$$

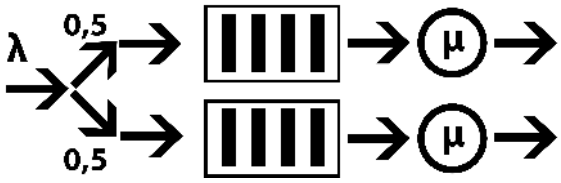
It is easy to see that

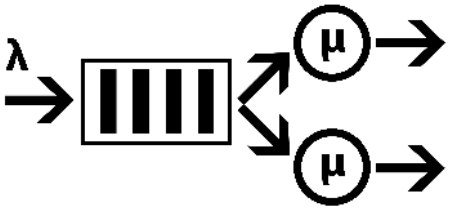
$$\overline{T_1}=\overline{T_2}=\frac{1}{\mu-\lambda}=\frac{1}{\mu(1-\rho)}.$$

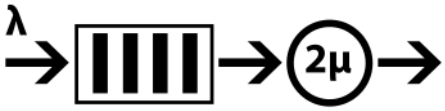
Consequently

$$\frac{\overline{T}}{\overline{T_1}}=\frac{\frac{1}{\overline{\mu(1-\rho)(1+\rho)}}}{\frac{1}{\overline{\mu(1-\rho)}}}=\frac{1}{1+\rho}>\frac{1}{2}, \quad \overline{T}=\frac{1}{1+\rho}\overline{T_1}.$$

Separated versus common $M/M/2$ queues

1. 

2. 

3. 

Compare the queues with respect to mean response time with the same traffic intensity
Solution:

$$\overline{T}_1 = \frac{1}{\mu(1-\frac{\lambda}{2\mu})} = \frac{1}{\mu(1-\frac{\rho}{2})} = \frac{2}{\mu(2-\rho)},$$

$$\overline{T}_2 = \frac{4}{\mu(4-\rho^2)} = \frac{4}{\mu(2-\rho)(2+\rho)}.$$

$$\overline{T}_3 = \frac{1}{2\mu(1-\frac{\rho}{2})} = \frac{2}{2\mu(2-\rho)} = \frac{1}{\mu(2-\rho)}.$$

Thus for the comparison we have

$$\frac{1}{\mu(2-\rho)} < \frac{1}{\mu(2-\rho)}\frac{4}{2+\rho} = \frac{2}{\mu(2-\rho)}\frac{2}{2+\rho} < \frac{2}{\mu(2-\rho)},$$

since $\rho < 2$, thus

$$\overline{T}_3 < \overline{T}_2 < \overline{T}_1.$$

■

Server Farms and Distributed Server Systems

In the server farm shown in Figure 2.5 jobs arrive according to a Poisson process with rate λ and are probabilistically split between two servers, with p fraction of the jobs going to server 1, which has service rate μ_1 , and $q = 1 - p$ fraction going to server 2, which has service rate μ_2 . Assume that job sizes are exponentially distributed.
It is easy to see that the response time of an arbitrary job is hiper-exponential with parameters $p, 1-p, \quad and \quad \mu_1, \mu_2$. The number of customers in the system and in the queues are the sum of the corresponding numbers in the separated $M/M/1$ systems and the distribution of the waiting and response times can be calculated with the help of law of total probability.

We can formulate the following two optimization problems:

- If we have a **total service capacity of** μ for the two servers, how should we optimally split μ between the two servers, into μ_1 and μ_2 , where $\mu = \mu_1 + \mu_2$, so as to minimize mean response time $E(T)$? We assume that $p \geq 1/2$.
- How can we **choose the probability** p so as to minimize $E(T)$?

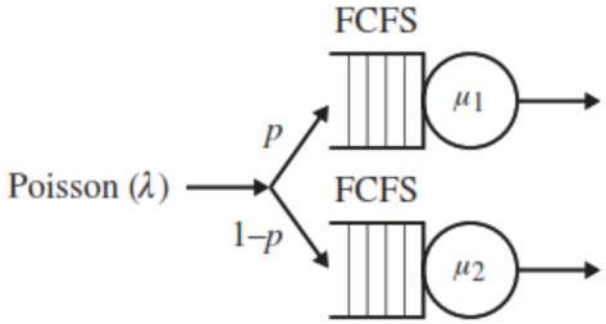


Figure 2.5: Server farms and distributed server systems

Let us see the solutions, first the **total service capacity** case.
Consider a frequently used option called **balanced load**. In this situation

$$\frac{\lambda p}{\mu_1} = \frac{\lambda(1-p)}{\mu_2} = \frac{\lambda(1-p)}{\mu - \mu_1}, \qquad p(\mu - \mu_1) = (1-p)\mu_1, \qquad \mu_1 = p\mu.$$

In this case

$$E(T_b) = \frac{p}{\mu_1 - \lambda p} + \frac{1-p}{\mu - \mu_1 - \lambda(1-p)} = \frac{p}{p(\mu - \lambda)} + \frac{1-p}{(1-p)(\mu - \lambda)} = \frac{2}{\mu - \lambda}.$$

The optimization problem can be formulates as follows

$$E(T) = \frac{p}{\mu_1 - \lambda p} + \frac{1-p}{\mu_2 - (1-p)\lambda}$$

subject to

$$\lambda p < \mu_1, \quad \lambda(1-p) < \mu_2, \quad \mu_1 + \mu_2 = \mu, \quad p \geq 1/2.$$

To find the optimal value we have to find the roots of the following equation

$$\frac{dE(T)}{d\mu_1} = \frac{p(-1)}{(\mu_1 - \lambda p)^2} + \frac{(1-p)(-1)(-1)}{(\mu - \mu_1 - (1-p)\lambda)^2} = 0$$
$$\frac{(1-p)}{(\mu - \lambda - (\mu_1 - \lambda p))^2} = \frac{p}{(\mu_1 - \lambda p)^2}$$

Let us introduce the notation

$$x = \mu_1 - \lambda p$$

then we can rewrite the equation as

$$(1-p)x^2 = p((\mu-\lambda)^2 + x^2 - 2(\mu-\lambda)x)$$
$$(2p-1)x^2 - 2p(\mu-\lambda)x + p(\mu-\lambda)^2 = 0$$

The solution

$$x_{1,2} = \frac{2p(\mu-\lambda) \pm \sqrt{4p^2(\mu-\lambda)^2 - 4(\mu-\lambda)p^2(2p-1)}}{2(2p-1)}$$
$$x_{1,2} = \frac{2p(\mu-\lambda) \pm (\mu-\lambda)2\sqrt{p^2 - 2p^2 + p}}{2(2p-1)}$$
$$x_{1,2} = \frac{(\mu-\lambda)(p \pm \sqrt{(1-p)p})}{2p-1} = \frac{\sqrt{p}(\mu-\lambda)(\sqrt{p} \pm \sqrt{1-p})}{2p-1}$$
$$= \frac{(\mu-\lambda)}{2p-1} \cdot \frac{\sqrt{p}(p-(1-p))}{p \mp \sqrt{1-p}} = (\mu-\lambda)\frac{\sqrt{p}}{\sqrt{p} \mp \sqrt{1-p}}.$$

Since

$$\frac{\sqrt{p}}{\sqrt{p} - \sqrt{1-p}} > 1$$

the solution is

$$x^* = (\mu-\lambda)\frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}}$$

thus the optimal value

$$\mu_1^* = \lambda p + (\mu-\lambda)\frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}}.$$

Hence the extra service intensity

$$(\mu-\lambda)\frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}}.$$

Let us show that

$$\frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}} \le p$$

that is

$$1 \le p(p + 1 - p + 2\sqrt{p(1-p)}) = p(1 + 2\sqrt{p(1-p)}) \le p(1 + 2 \times 1/2) = 2p$$

Hence $p \geq \frac{1}{2}$ which is true.

Thus the optimal mean response time

$$\begin{aligned}
E(T_m) &= \frac{p}{\mu_1 - \lambda p} + \frac{1-p}{\mu_2 - \lambda(1-p)} \\
&= \frac{p}{(\mu-\lambda)\frac{\sqrt{p}}{\sqrt{p}+\sqrt{1-p}}} + \frac{1-p}{\mu - \left[\lambda p + \frac{(\mu-\lambda)\sqrt{p}}{\sqrt{p}+\sqrt{1-p}}\right] - (1-p)\lambda} \\
&= \frac{\sqrt{p}(\sqrt{p}+\sqrt{1-p})}{\mu-\lambda} + \frac{\sqrt{1-p}(\sqrt{p}+\sqrt{1-p})}{\mu-\lambda} \\
&= \frac{(\sqrt{p}+\sqrt{1-p})^2}{\mu-\lambda} = \frac{p+1-p+2\sqrt{p(1-p)}}{\mu-\lambda} \\
&= \frac{1+2\sqrt{p(1-p)}}{\mu-\lambda} \leq \frac{1+2\cdot\frac{1}{2}}{\mu-\lambda} = \frac{2}{\mu-\lambda} = E(T_b).
\end{aligned}$$

which was expected since it is the minimal value. At the same time we can see that the distribution of the total capacity is not proportional to p, which was the balanced load. However, if $p = 1/2$ then the balanced load value and the minimal value are the same. In other words, if we choose the servers with the same probability $1/2$ which many times happens because we have no information about the speed of the servers, that we have to give the half of the total service capacity because this minimizes the mean response time.

Let us see the solution to the minimization problem with **respect to** p.
We have the same expected response time function, namely

$$E(T) = \frac{p}{\mu_1 - p\lambda} + \frac{1-p}{\mu_2 - (1-p)\lambda}, \quad \mu_1 = \alpha \cdot \mu_2, \;\; \alpha \geq 1, \;\; p \leq$$

We have to find the solution to derivative function

$$\begin{gathered}
\frac{dE(T)}{dp} = \frac{1(\alpha\mu_2 - p\lambda) + \lambda p}{(\alpha\mu_2 - p\lambda)^2} + \frac{-(\alpha\mu_2 - p\lambda) + \lambda p}{(\mu_2 - (1-p)\lambda)^2} = 0 \\
\frac{(\alpha\mu_2 - p\lambda) + \lambda p}{(\alpha\mu_2 - p\lambda)^2} = \frac{\mu_2}{(\mu_2 - (1-p)\lambda)^2} \\
\alpha(\mu_2 - (1-p)\lambda)^2 = (\alpha\mu_2 - p\lambda)^2
\end{gathered}$$

Since

$$\mu_2 > (1-p)\lambda, \quad \alpha\mu_2 > p\lambda, \quad p \leq 1$$

then

$$\begin{aligned}
\sqrt{\alpha}(\mu_2 - (1-p)\lambda) &= \alpha\mu_2 - p\lambda \\
\sqrt{\alpha}(\mu_2 - \lambda + \lambda p) &= \alpha\mu_2 - \lambda p \\
(\lambda\sqrt{\alpha} + \lambda)p &= \mu_2(\alpha - \sqrt{\alpha}) + \lambda\sqrt{\alpha}
\end{aligned}$$

Thus

$$p = \frac{\mu_2(\alpha - \sqrt{\alpha}) + \lambda\sqrt{\alpha}}{\lambda(1 + \sqrt{\alpha})}.$$

In addition, since p is a probability its value should not be greater than 1, that is we have another condition

$$\frac{\mu_2(\alpha - \sqrt{\alpha}) + \lambda\sqrt{\alpha}}{\lambda(1 + \sqrt{\alpha})} \leq 1$$

which results

$$\mu_2(\alpha - \sqrt{\alpha}) \leq \lambda.$$

It is easy to see that

$$\frac{\mu_2(\alpha - \sqrt{\alpha}) + \lambda\sqrt{\alpha}}{\lambda(1 + \sqrt{\alpha})} \geq \frac{1}{2}$$

$$2(\mu_2(\alpha - \sqrt{\alpha}) + \lambda\sqrt{\alpha}) \geq \lambda(1 + \sqrt{\alpha})$$

$$2\mu_2(\alpha - \sqrt{\alpha}) + \lambda(\sqrt{\alpha} - 1) \geq 0.$$

Thus the conditions are

$$\mu_2(\alpha - \sqrt{\alpha}) \leq \lambda, \quad \lambda p < \alpha\mu_2, \quad (1 - p)\lambda < \mu_2.$$

In other words, if μ_1, μ_2 are fixed $p = 1/2$ does not minimize the expected response times except they are equal. If $\alpha = 1$, then $\mu_1 = \mu_2 = \mu/2$, and $p = \lambda/(\lambda \cdot 2) = 1/2$.

In the case of **balanced load** we have

$$\frac{p}{\mu} = \frac{1 - p}{\mu}$$

thus $p = 1/2$, that is the optimal value for p and the value obtain by using the balanced load principle are the same, thus the minimum expected response times are the same, too.

Let us calculate the minimal value, that is

$$E(T) = \frac{p}{\mu_1 - \lambda p} + \frac{1 - p}{\mu_2 - (1 - p)\lambda}$$

$$p = \frac{\mu_2(\alpha - \sqrt{\alpha}) + \sqrt{\alpha}\lambda}{\lambda(1 + \sqrt{\alpha})}, \quad \mu_1 = \alpha\mu_2$$

$$\begin{aligned} 1 - p &= 1 - \frac{\mu_2(\alpha - \sqrt{\alpha}) + \sqrt{\alpha}\lambda}{\lambda(1 + \sqrt{\alpha})}\mu_2 \\ &= \frac{\lambda + \lambda\sqrt{\alpha} - \mu_2(\alpha - \sqrt{\alpha}) - \sqrt{\alpha}\lambda}{\lambda(1 + \sqrt{\alpha})} \\ &= \frac{\lambda - \mu_2(\alpha - \sqrt{\alpha})}{\lambda(1 + \sqrt{\alpha})} \end{aligned}$$

$$
\begin{aligned}
E(T) &= \frac{\frac{(\mu_2(\alpha-\sqrt{\alpha})+\sqrt{\alpha}\lambda}{\lambda(1+\sqrt{\alpha})}}{\alpha\mu_2 - \frac{\lambda\left(\mu_2(\alpha-\sqrt{\alpha})+\sqrt{\alpha}\lambda\right)}{\lambda(1+\sqrt{\alpha})}} + \frac{\frac{\lambda-\mu_2(\alpha-\sqrt{\alpha})}{\lambda(1+\sqrt{\alpha})}}{\mu_2 - \frac{\lambda-\mu_2(\alpha-\sqrt{\alpha})}{\lambda(1+\sqrt{\alpha})}\cdot\lambda} \\
&= \frac{\mu_2(\alpha-\sqrt{\alpha})+\sqrt{\alpha}\lambda}{\lambda\left(\alpha\mu_2(1+\sqrt{\alpha}) - \mu_2(\alpha-\sqrt{\alpha}) - \sqrt{\alpha}\lambda\right)} + \frac{\lambda-\mu_2(\alpha-\sqrt{\alpha})}{\lambda\left(\mu_2(1+\sqrt{\alpha}) - \lambda + \mu_2(\alpha-\sqrt{\alpha})\right)} \\
&= \frac{\mu_2(\alpha-\sqrt{\alpha})+\sqrt{\alpha}\lambda}{\lambda\left(\alpha\mu_2\sqrt{\alpha} + \mu_2\sqrt{\alpha} - \sqrt{\alpha}\lambda\right)} + \frac{\lambda-\mu_2(\alpha+\sqrt{\alpha})}{\lambda(\mu_2 - \lambda + \mu_2\alpha)} \\
&= \frac{\sqrt{\alpha}(\mu_2\sqrt{\alpha} - \mu_2 + \lambda)}{\lambda\sqrt{\alpha}(\alpha\mu_2 + \mu_2 - \lambda)} + \frac{\lambda-\mu_2(\alpha-\sqrt{\alpha})}{\lambda(\mu_2 + \mu_2\alpha - \lambda)} \\
&= \frac{\mu_2\sqrt{\alpha} - \mu_2 + \lambda + \lambda - \mu_2(\alpha-\sqrt{\alpha})}{\lambda(\mu_2 + \mu_2\alpha - \lambda)} \\
&= \frac{2\mu_2\sqrt{\alpha} - \mu_2 + 2\lambda - \mu_2\alpha}{\lambda(\mu_2 + \mu_2\alpha - \lambda)} \\
&= \frac{\mu_2(2\sqrt{\alpha} - \alpha - 1) + 2\lambda}{\lambda\left(\mu_2(1+\alpha) - \lambda\right)}
\end{aligned}
$$

Thus the minimum response time

$$
E(T_m) = \frac{\mu_2(2\sqrt{\alpha} - \alpha - 1) + 2\lambda}{\lambda\left(\mu_2(1+\alpha) - \lambda\right)}
$$

If $\alpha = 1$ then $\mu_1 = \mu_2 = \mu$

$$
E(T_m) = \frac{2\lambda}{\lambda(2\mu - \lambda)} = \frac{2}{2\mu - \lambda}
$$

In the case of **balanced load** we have

$$
\frac{p}{\mu_1} = \frac{1-p}{\mu_2}, \quad \frac{p}{\alpha\mu_2} = \frac{1-p}{\mu_2}, \quad p = \frac{\alpha}{\alpha+1}
$$

Thus in this case the mean value is

$$
E(T_b) = \frac{\frac{\alpha}{\alpha+1}}{\alpha\mu_2 - \frac{\alpha}{\alpha+1}\lambda} + \frac{\frac{1}{\alpha+1}}{\mu_2 - \frac{1}{1+\alpha}\lambda}
$$

$$
\begin{aligned}
E(T_b) &= \frac{\alpha}{\alpha\mu_2(1+\alpha) - \alpha\lambda} + \frac{1}{\mu_2(1+\alpha) - \lambda} \\
&= \frac{1}{\mu_2(1+\alpha) - \alpha} + \frac{1}{\mu_2(1+\alpha) - \lambda} = \frac{2}{\mu_2(1+\alpha) - \lambda}
\end{aligned}
$$

Hence we have the final result for the two cases, namely
balanced load

$$E(T_b) = \frac{2}{\mu_2(1+\alpha) - \lambda}$$

the minimal value

$$E(T_m) = \frac{\mu_2(2\sqrt{\alpha} - \alpha - 1) + 2\lambda}{\lambda\left(\mu_2(1+\alpha) - \lambda\right)}$$

Comparing them we have

$$\begin{aligned}\frac{E(T_m)}{E(T_b)} &= \frac{\mu_2(2\sqrt{\alpha} - \alpha - 1) + 2\lambda}{2\lambda} \\ &= \frac{2\lambda - \mu_2(\sqrt{\alpha} - 1)^2}{2\lambda} \leq 1\end{aligned}$$

If $\alpha = 1$ then $\mu_1 = \mu_2 = \mu$ and

$$E(T_m) = E(T_b) = \frac{2}{2\mu - \lambda}$$

.

Usually the customer has no information about the service speed in advance and that is why $p = 1/2$, that is the expected response time is not minimal.

Let us compare numerically the mean of the total number of customers in the heterogeneous system $M/M/2$, that is two separated queues with different service intensities $\mu_1 = 10,\ \mu_2 = 2$ and $\lambda = 1,\ p = 0.9$, with the corresponding homogeneous system $M/M/2$, when the service rate is $(\mu_1 + \mu_2)/2 = 6$.
Using QSA we get $\bar{N}_1 + \bar{N}_2 = 0.0989 + 0.0526 = 0.1515$ and $\bar{N} = 0.168$ which means that the combined separated system is preferable.

However, if $p = 0.5$ than the $M/M/2$ common queue systems is always better with respect to the mean total number of customers in the system. The proof is the following.
We show that

$$\frac{\frac{\lambda}{2\mu_1}}{1 - \frac{\lambda}{2\mu_1}} + \frac{\frac{\lambda}{2\mu_2}}{1 - \frac{\lambda}{2\mu_2}} > \frac{\frac{2\lambda}{\mu_1+\mu_2}}{1 - \left(\frac{2\lambda}{2(\mu_1+\mu_2)}\right)^2}.$$

Thus,

$$\begin{aligned}\frac{1}{2\mu_1 - \lambda} + \frac{1}{2\mu_2 - \lambda} &> \frac{2}{(\mu_1 + \mu_2) - \frac{\lambda^2}{\mu_1+\mu_2}} \\ \frac{2\mu_1 - \lambda + 2\mu_2 - \lambda}{(2\mu_1 - \lambda)(2\mu_2 - \lambda)} &> \frac{2(\mu_1 + \mu_2)}{(\mu_1 + \mu_2)^2 - \lambda^2} \\ \frac{\mu_1 + \mu_2 - \lambda}{(2\mu_1 - \lambda)(2\mu_2 - \lambda)} &> \frac{\mu_1 + \mu_2}{(\mu_1 + \mu_2 + \lambda)(\mu_1 + \mu_2 - \lambda)}\end{aligned}$$

$$\frac{(\mu_1+\mu_2-\lambda)^2}{(2\mu_1-\lambda)(2\mu_2-\lambda)} > \frac{\mu_1+\mu_2}{\mu_1+\mu_2+\lambda}$$

We show that

$$\frac{(\mu_1+\mu_2-\lambda)^2}{(2\mu_1-\lambda)(2\mu_2-\lambda)} \geq 1$$

from which the inequality follows.
That is

$$(\mu_1+\mu_2-\lambda)^2 \geq (2\mu_1-\lambda)(2\mu_2-\lambda)$$
$$(\mu_1+\mu_2)^2+\lambda^2-2\lambda(\mu_1+\mu_2) \geq 4\mu_1\mu_2-2\mu_1\lambda-2\mu_2\lambda+\lambda^2$$
$$(\mu_1+\mu_2)^2 \geq 4\mu_1\mu_2$$
$$(\mu_1-\mu_2)^2 \geq 0$$

which is true.
Since

$$\frac{\mu_1+\mu_2}{\mu_1+\mu_2+\lambda} < 1$$

we are ready with the proof.
It is not difficult to show that the probability of waiting of an arbitrary customer after selecting a server is

$$P(W>0) = P_W = p\frac{p\lambda}{\mu_1} + (1-p)\frac{(1-p)\lambda}{\mu_2}.$$

Let us consider the following example dealing with closed and open systems, see the Figure below

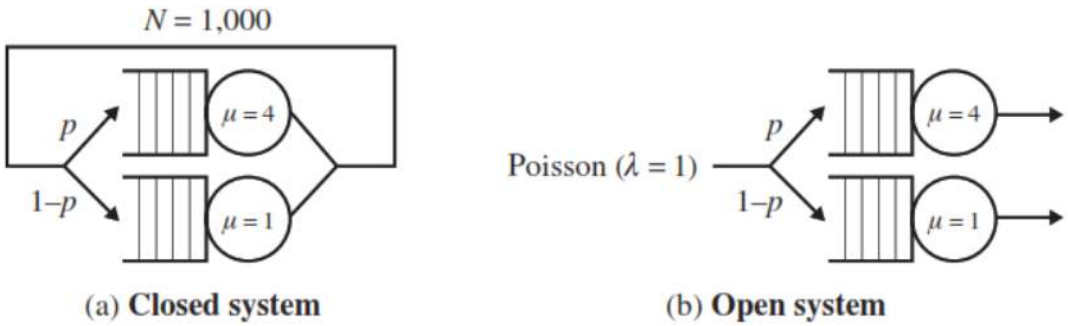


Figure 2.6: Closed and open systems

It is proved that in any closed system where the number of circulated jobs N (level of multiprogramming) is high enough the expected response/waiting time is minimized if the balanced load principle is applied, see Harchol-Balter [46] from where the example is taken. The optimal value is

$$E(T_m) = E(T_b) = N/(\mu_1+\mu_2).$$

$M/M/2$ with heterogeneous servers and fastest free server service discipline

Consider a variant of the $M/M/2$ queue where the service rates of the two servers are not identical. Denote the service rate of the first server by μ_1 and the service rate of the second server by μ_2, where $\mu_1 \geq \mu_2$. In the case of heterogeneous servers, the rule is that when both servers are idle, the faster server is scheduled for service before the slower one, that is called **FFS - Fastest Free Server** But if there is only one server free when an arrival occurs, it enters service with the free server regardless of the service rate. If both servers are busy, the arriving customer waits in common line for service in the order of arrival. Define the utilization, a, for this system to be $a = \lambda/(\mu_1 + \mu_2)$.
Let us determine the mean number of jobs in the system $E(N)$, mean response and waiting time $E(T), E(W)$, respectively.
It was proved, for example in Harchol-Balter [46] and Trivedi [120] that

$$E(N) = \frac{1}{A(1-a)^2}, \quad A = \frac{\mu_1\mu_2(1+2a)}{\lambda(\lambda+\mu_2)} + \frac{1}{1-a}.$$

Using Little law $E(T) = E(N)/\lambda$.

It can be shown that $P(Q = i) = a^{i+1}/A,\ i = 0, 1, 2, ...\infty$ and thus

$$E(Q) = \frac{a^2}{A(1-a)^2}, \quad E(W) = \frac{E(Q)}{\lambda}, \quad E(S) = \frac{E(N)-E(Q)}{\lambda}, P(W > 0) = \frac{a}{(1-a)A}.$$

It is easy to see if $\mu_1 = \mu_2 = \mu$ then $a = \rho/2$ and there is no difference between the servers thus the corresponding measures are the same as in the homogeneous $M/M/2$ system, that is

$$E(N) = \frac{4\rho}{4-\rho^2}, \quad E(Q) = \frac{\rho^3}{4-\rho^2}, \quad E(S) = \frac{1}{\mu}.$$

$M/M/2$ with heterogeneous servers and random free server service discipline

Let us see the previous system with the exception that and arriving customer select between the free servers with the same probability, that is the selection probability is 0.5. Let us call this discipline as **RFS - Random Free Server** . However, this small modification makes the calculations rather complicated, but it can be treated numerically. To do so, let us introduce the following notations: Let (c_1, c_2, k) be the state of the system, where k is the number of customers in the queue, and c_i is 1 if server i is busy and 0 otherwise. Let us denote by $\Pi_{0,0,0},\ \Pi_{0,1,0},\ \Pi_{1,0,0},\ \Pi_{1,1,k},\ k = 0, 1, 2, ... \infty$ the steady-state distribution of the system which exists if $a < 1$. These probabilities can be computed in the following way

$$\Pi_{0,1,0} = \left(\mu_1 + \frac{\mu_1\mu_2}{2\lambda+\mu_1}\right) \Bigg/ \left(\lambda + \frac{\mu_2}{2} - \frac{\mu_1\mu_2}{4\lambda+2\mu_1}\right)\Pi_{1,1,0},$$

$$\Pi_{1,0,0} = \frac{2\mu_2}{2\lambda+\mu_1}\left(\Pi_{0,1,0}\frac{1}{2} + \Pi_{1,1,0}\right), \quad \Pi_{0,0,0} = (\Pi_{1,0,0}\mu_1 + \Pi_{0,1,0}\mu_2)/\lambda,$$

$$\Pi_{1,1,k} = P(Q = k) = a^k\Pi_{1,1,0}, \quad k = 0, 1, ..., \infty.$$

Of course $\Pi_{1,1,0}$ should satisfy the normalizing condition, that is

$$\Pi_{0,0,0} + \Pi_{0,1,0} + \Pi_{1,0,0} + \Pi_{1,1,0}/(1-a) = 1.$$

It can be obtained very easily, starting the calculation by an initial value then calculate the sum, and then divide each term by this sum.

If we have the distribution the expectations can be calculated in the standard way, that is

$$E(N) = \Pi_{0,1,0} + \Pi_{1,0,0} + \sum_{k=0}^{\infty}(2+k)a^k\Pi_{1,1,0} = \Pi_{0,1,0} + \Pi_{1,0,0} + \frac{2-a}{(1-a)^2}\Pi_{1,1,0},$$

$$E(Q) = \frac{a}{(1-a)^2}\Pi_{1,1,0},$$

$$E(T) = E(N)/\lambda, \quad E(W) = E(Q)/\lambda, \quad E(S) = (E(N) - E(Q))/\lambda, \quad P(W>0) = \frac{\Pi_{1,1,0}}{1-a}.$$

It is not difficult to see that in FFS and RFS cases

$$P(W > x) = P_W e^{-(\mu_1+\mu_2-\lambda)x}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Example 17 *Compare numerically the mean response, waiting time and probability of waiting in the heterogeneous $M/M/2$ systems with separated and common queues. Use FFS and RFS discipline in the heterogeneous case, in the homogeneous $M/M/2$ system the service rate is $(\mu_1+\mu_2)/2$. The input parameters are $\lambda = 1, \quad \mu_2 = 2$.*

Table 2.2: Comparison of systems

Separated	E(T) $\mu_1 = 10$	E(T) $\mu_1 = 3.5$	E(W) $\mu_1 = 10$	E(W) $\mu_1 = 3.5$	P(W>0) $\mu_1 = 10$	P(W>0) $\mu_1 = 3.5$
balanced	0.1818	0.4444	0.0151	0.0808	0.0833	0.1818
p=0.9	0.1515	0.3987	0.0115	0.0916	0.086	0.2361
p=0.5	0.3859	0.5	0.0859	0.1071	0.15	0.1964
optimal		0.3981		0.0989		0.2529
Common						
FFS	0.1341	0.3391	0.0009	0.0112	0.0102	0.0504
RFS	0.2689	0.3964	0.0018	0.0131	0.0205	0.0589
Hom.	0.1678	0.376	0.0011	0.0124	0.0128	0.0559

Example 18 *Consider a service center with* 4 *servers where* $\lambda = 6$, $\mu = 2$.
Find the performance measures of the system.

Solution:

$$P_0 = 0.0377, \quad \overline{Q} = 1.528, \quad \overline{N} = 4.528, \quad \overline{S} = 1, \quad \overline{n} = 3,$$
$$P(W > 0) = P(n \geq 4) = C(4,3) = 0.509, \overline{W} = 0.255 \text{ time unit}, \quad \overline{T} = 0.755 \text{ time unit},$$
$$U_n = \frac{3}{4}, \overline{e} = 0.35 \text{ time unit}, \quad E\delta = 1.05 \text{ time unit},$$
$$E\delta_r = 4.2 \text{ time unit}, \quad U_r = 0.9623.$$

■

Example 19 *Find the number of runways in an airport such a way the the probability of waiting of an airplane should not exceed* 0.1. *The arrivals are supposed to be Poisson distributed with rate* $\lambda = 27$ *per hour and the service times are exponentially distributed with a mean of* 2 *minutes.*

Solution:
First use the same time unit for the rates, let us compute in hours. Hence $\mu = 30$ and for stability we need $\frac{\lambda}{n\mu} < 1$ which results $n > 1$.

Denote by $P_i(W > 0)$ the probability of waiting for i runways. By applying the corresponding formulas we get

$$P_2(W > 0) = 0.278, \quad P_3(W > 0) = 0.070, \quad P_4(W > 0) = 0.014.$$

Hence the solution is $n = 3$. In this case $P_0 = 0.403$ and $\overline{W} = 0.0665 hour, \quad \overline{Q} = 0.03$.

■

Example 20 *Consider a fast food shop where to the customers arrive according to a Poisson law one customer in* 6 *seconds on the average. The service time is exponentially distributed with* 20 *seconds mean. Assuming that the maintenance cost of a server is* 100 *Hungarian Forint and the waiting cost is the same find the optimal value of the server which minimizes the mean cost per hour.*

Solution:

$$\overline{Q} = \lambda \overline{W} = \frac{3600}{6} \overline{W}$$
$$E(TC) = 100 \times n + 100 \times 600 \times \overline{W}$$
$$\frac{\lambda}{\mu} = \frac{\frac{1}{6}}{\frac{1}{20}} = \frac{20}{6} \text{ thus } n \geq 4 .$$

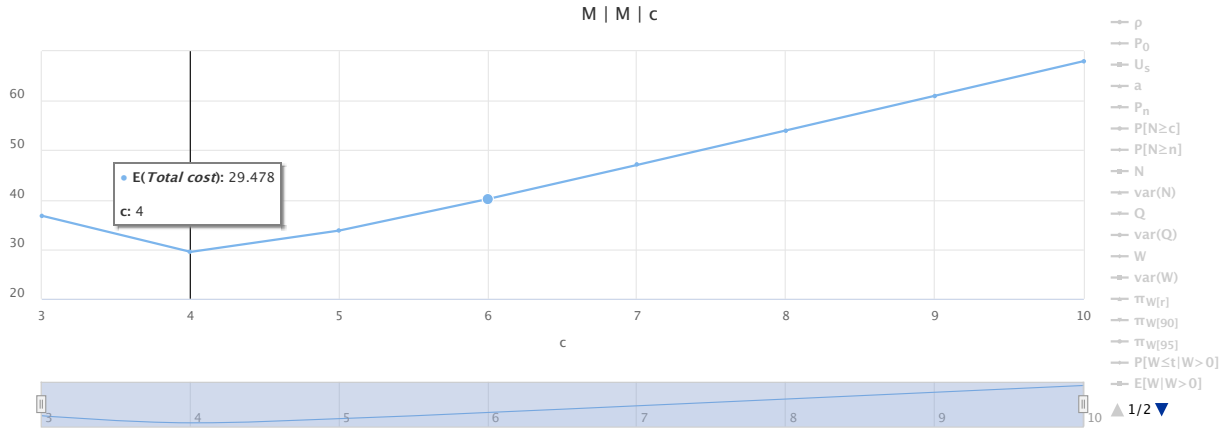
Computing for the values $n = 4,\ 5,\ 6,\ 7,\ 8$ we have found that the minimum is achieved at $n = 5$. This case the performance measures are

$$\overline{W} = 3.9 second, \quad P(e) = 0.66, \quad P(W) = 0.34 ,$$
$$E\delta = 29.7 second, \quad \overline{e} = 14.9 second, \quad \overline{Q} = 0.65 ,$$
$$\overline{n} = 2.5, \quad \overline{N} = 3.15, \quad \overline{S} = 2.5, \quad E(TC) = 565 \text{ HUF/hour}.$$

■

Example 21 *Let us consider the following optimization problem. Find the number of servers which minimizes the total expected cost per unit time with the following input parameters and costs:*

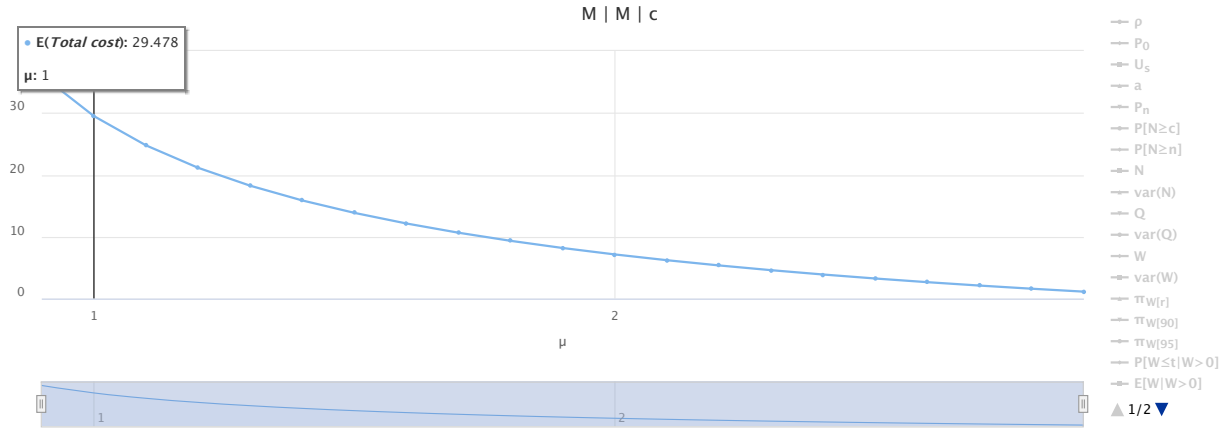
$$\lambda = 2, \quad \mu = 1, \quad CS = 1, \quad CWS = 20, \quad CI = 1, \quad CSR = 5, \quad R = 20.$$



Expected total cost function with respect to number of servers

Example 22 *Let us consider the following optimization problem. Find the intensity of service which minimizes the total expected cost per unit time with the following input parameters and costs:*

$$\lambda = 2, \quad n = 4, \quad CS = 1, \quad CWS = 20, \quad CI = 1, \quad CSR = 5, \quad R = 20.$$



Expected total cost function with respect to service rate

2.8 The $M/M/c$ Non-preemptive Priority Queue (HOL)

There are n priority classes with each class having a Poisson arrival pattern with mean arrival rate λ_i. Each customer has the same exponential service time requirement. Then the overall arrival pattern is Poisson with mean $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$. The server utilization

$$a = \frac{\lambda \overline{S}}{c} = \frac{\lambda}{c\mu} = \frac{\rho}{c}, \qquad \overline{W}_1 = \frac{C[c,\rho]\overline{S}}{c(1-\lambda_1 \overline{S}/c)},$$

and the following equations are also true:

$$\overline{W}_j = \frac{C[c,\rho]\overline{S}}{c\left[1-\left(\overline{S}\sum_{i=1}^{j-1}\lambda_i\right)/c\right]\left[1-\left(\overline{S}\sum_{i=1}^{j}\lambda_i\right)/c\right]}, \quad j = 2,\ldots,n$$

$$\overline{T}_j = \overline{W}_j + \overline{S}, \quad \overline{Q}_j = \overline{\lambda}_j \cdot \overline{W}_j, \quad \overline{N}_j = \overline{\lambda}_j \cdot \overline{T}_j, \quad j = 1,2,\ldots,n$$

$$\overline{W} = \frac{\lambda_1}{\lambda}\overline{W}_1 + \frac{\lambda_2}{\lambda}\overline{W}_2 + \ldots + \frac{\lambda_n}{\lambda}\overline{W}_n$$

$$\overline{Q} = \overline{\lambda} \cdot \overline{W}, \quad \overline{T} = \overline{W} + \overline{S}, \quad \overline{N} = \overline{\lambda} \cdot \overline{T}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.9 The $M/M/c/K$ Queue - Multiserver, Finite-Capacity Systems

This queue is a variation of a multiserver system and only maximum K customers are allowed to stay in the system. As earlier the number of customers in the system is a birth-death process with appropriate rates and for the steady-state distribution we have

$$P_n = \begin{cases} \frac{\lambda^n}{n!\mu^n}P_0, & \text{for } 0 \le n \le c \\ \\ \frac{\lambda^n}{c^{n-c}c!\mu^n}P_0, & \text{for } c \le n \le K. \end{cases}$$

From the normalizing condition for P_0 we have

$$P_0 = \left(\sum_{n=0}^{c-1}\frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{K}\frac{\lambda^n}{c^{n-c}c!\mu^n}\right)^{-1}.$$

To simplify this expression let $\rho = \dfrac{\lambda}{\mu}, \quad a = \dfrac{\rho}{c}.$

Then

$$\sum_{n=c}^{K}\frac{\rho^n}{c^{n-c}c!} = \frac{\rho^c}{c!}\sum_{n=c}^{K}a^{n-c} = \begin{cases} \frac{\rho^c}{c!}\frac{1-a^{K-c+1}}{1-a}, & \text{if } a \ne 1 \\ \\ \frac{\rho^c}{c!}(K-c+1), & \text{if } a = 1. \end{cases}$$

Thus

$$P_0 = \begin{cases} \left[\frac{\rho^c}{c!}\frac{1-a^{K-c+1}}{1-a} + \sum_{n=0}^{c-1}\frac{\rho^n}{n!}\right]^{-1}, & \text{if } a \neq 1 \\ \left[\frac{\rho^c}{c!}(K-c+1) + \sum_{n=0}^{c-1}\frac{\rho^n}{n!}\right]^{-1}, & \text{if } a = 1. \end{cases}$$

The main performance measures can be obtained as follows

- Mean and variance of queue length

$$\begin{aligned} \overline{Q} &= \sum_{n=c+1}^{K}(n-c)P_n = \sum_{n=c+1}^{K}(n-c)\frac{\lambda^n}{c^{n-c}c!\mu^n}P_0 = \frac{P_0\rho^c}{c!}\sum_{n=c+1}^{K}(n-c)\frac{\rho^{n-c}}{c^{n-c}} \\ &= \frac{P_0\rho^c a}{c!}\sum_{n=c+1}^{K}(n-c)a^{n-c-1} = \frac{P_0\rho^c a}{c!}\sum_{i=1}^{K-c}ia^{i-1} = \frac{P_0\rho^c a}{c!}\frac{d}{da}\left(\sum_{i=0}^{K-c}a^i\right) \\ &= \frac{P_0\rho^c a}{c!}\frac{d}{da}\left(\frac{1-a^{K-c+1}}{1-a}\right) \end{aligned}$$

  which results

$$\overline{Q} = \frac{P_0\rho^c a}{c!(1-a)^2}[1 - a^{K-c+1} - (1-a)(K-c+1)a^{K-c}]$$

  In particular, if $a = 1$ then the L'Hopital's rule should be applied twice.

$$\begin{aligned} &\bar{Q} = \frac{P_0\rho^c}{c!}\left[\frac{a - a^{K-c+2} - (1-a)(K-c+1)a^{K-c+1}}{(1-a)^2}\right] \\ &\lim_{a\to 1}\bar{Q}: \quad \text{L'Hospital rule} \\ &= \frac{P_0\rho^c}{c!}\cdot\frac{\left(1 - (K-c+2)a^{(K-c+1)} - (-1)(K-c+1)a^{(K-c+1)} - (1-a)(K-c+1)^2a^{(K-c)}\right)}{-2(1-a)} \\ &= \frac{P_0\rho^c}{c!}\cdot\frac{1 - a^{(K-c+1)} - (1-a)(K-c+1)^2a^{(K-c)}}{-2(1-a)}. \end{aligned}$$

  Applying again the L'Hospital rule

$$= \frac{P_0\rho^c}{2c!}\left[-(K-c+1)a^{(K-c)} + (K-c+1)^2a^{(K-c)} - (1-a)(K+c+1)^2(K-c)a^{(K-c-1)}\right].$$

  Then

$$\lim_{a\to 1}\bar{Q}(a) = \frac{P_0\rho^c}{2c!}\left[(K-c+1)^2 - (K-c+1)\right] = \frac{P_0\rho^c}{2c!}(K-c)(K-c+1).$$

If $c = 1$ and $\rho = 1$ we get

$$\bar{Q} = \frac{K(K-1)}{2(K+1)}.$$

$$E(Q^2) = \sum_{k=c}^{K} (k-c)^2 P_k, \quad Var(Q) = E(Q^2) - (E(Q))^2.$$

- *Mean and variance of number of customers in the system*
  It is easy to see that

$$\overline{\lambda} = \lambda(1 - P_K) = \mu\overline{c}$$

and since

$$\overline{N} = \overline{Q} + \overline{c}$$

we get

$$\overline{N} = \overline{Q} + \rho(1 - P_K).$$

$$E(N^2) = \sum_{k=1}^{K} k^2 P_k, \quad Var(N) = E(N^2) - (E(N))^2$$

- *Distribution at the moment of arrival into the system*
  By applying the Bayes's rule we have

$$\begin{aligned}\Pi_n &= P(\text{ there are } n \text{ customers in the system} \\ &\quad \text{given a customer is about to enter into the system }) \\ &= \lim_{\Delta t \to 0} \left\{ \frac{[\lambda \Delta t + o(\Delta t)] P_n}{\sum_{n=0}^{K-1} [\lambda \Delta t + o(\Delta t)] P_n} \right\} = \lim_{\Delta t \to 0} \left\{ \frac{[\lambda + o(\Delta t)/\Delta t] P_n}{\sum_{n=0}^{K-1} [\lambda + o(\Delta t)/\Delta t] P_n} \right\} \\ &= \frac{\lambda P_n}{\lambda \sum_{n=0}^{K-1} P_n} = \frac{P_n}{1 - P_K}, \qquad (n \leq K-1).\end{aligned}$$

Obviously in the case of an $M/M/c/\infty$ system $\Pi_n = P_n$ since P_K tends to 0.

- *Mean and variance of response and waiting times*
  The mean times can be obtained by applying the **Little's law**, that is

$$\overline{W} = \frac{\overline{Q}}{\lambda(1 - P_K)} = \sum_{k=c}^{K-1} \frac{(k-c+1)}{(c\mu)} \Pi_k$$

$$\overline{T} = \frac{\overline{N}}{\lambda(1 - P_K)} = \overline{W} + 1/\mu$$

Since the conditional waiting time is Erlang distributed, it is easy to see that

$$E(W^2) = \sum_{k=c}^{K-1} \frac{(k-c+1)+(k-c+1)^2}{(c\mu)^2}\Pi_k, \quad Var(W) = E(W^2) - (E(W))^2,$$

$$Var(T) = Var(W) + 1/\mu^2.$$

- *Distribution of the waiting time*
  As in the previous parts for $F_W(t)$ the theorem of total probability is applied resulting

$$\begin{aligned} F_W(t) &= F_W(0) + \sum_{n=c}^{K-1} \Pi_n \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx \\ &= F_W(0) + \sum_{n=c}^{K-1} \Pi_n \left(1 - \int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx\right). \end{aligned}$$

  Since

$$\int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x}\, dx = \sum_{i=0}^{m} \frac{(\lambda t)^i e^{-\lambda t}}{i!}$$

  applying substitutions $m = n - c$, $\lambda = c\mu$ we have

$$\int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx = \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!},$$

  thus

$$\begin{aligned} F_W(t) &= F_W(0) + \sum_{n=c}^{K-1} \Pi_n - \sum_{n=c}^{K-1} \Pi_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!} \\ &= 1 - \sum_{n=c}^{K-1} \Pi_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}. \end{aligned}$$

The Laplace-transform of the waiting and response times can be derived similarly, by using the law of total Laplace-transforms.

- *Overall utilization of the servers* can be obtained as

  The utilization of a single server is

$$U_s = \sum_{k=1}^{n-1} \frac{k}{c} P_k + \sum_{k=c}^{K} P_k = \frac{\bar{c}}{c} = \frac{\rho(1-P_K)}{c}.$$

  Hence the overall utilization can be written as

$$U_n = cU_s = \bar{c}.$$

- *The mean busy period of the system* can be computed as follows

  The system is said to be idle if the is no customer in the system, otherwise the system is busy. Let $E\delta_r$ denote the mean busy period of the system. Then the utilization of the system is

$$U_r = 1 - P_0 = \frac{E\delta_r}{\frac{1}{\lambda} + E\delta_r},$$

  thus

$$E\delta_r = \frac{1 - P_0}{\lambda P_0}.$$

  If the individual servers are considered then we assume that a given server becomes busy earlier if it became idle earlier. Hence if $j < c$ customers are in the system then the number of idle servers is $c - j$.

  Let as consider a given server. On the condition that at the instant when it became idle the number of customers in the system was j its mean idle time is

$$\overline{e}_j = \frac{c - j}{\lambda}.$$

  The probability of this situation is

$$a_j = \frac{\Pi_j}{\sum_{i=0}^{c-1} \Pi_i} = \frac{P_j}{\sum_{i=0}^{c-1} P_i}.$$

  Then applying the law of total expectations for its mean idle period we have

$$\overline{e} = \sum_{j=0}^{c-1} a_j \overline{e}_j = \sum_{j=0}^{c-1} \frac{(c-j)\Pi_j}{\lambda \sum_{i=0}^{c-1} \Pi_i} = \sum_{j=0}^{c-1} (c-j)\Pi_j \frac{1}{\lambda P(e)} = \frac{c - \rho(1 - P_K)}{\lambda \sum_{i=0}^{c-1} P_i},$$

  where $P(e) = \sum_{j=0}^{c-1} \Pi_j$ denotes the probability that an arriving customer finds an idle server.

  Since

$$U_s = \frac{E\delta}{\overline{e} + E\delta},$$

  thus

$$E\delta = \frac{U_s}{1 - U_s} \overline{e}$$

  where $E\delta$ denotes its mean busy period.

It is easy to see if $P_K = 0$ than the performance measures of the $M/M/c$ and the $M/M/c/K$ systems are the same which is reasonable.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.10 The $M/M/c/K$ Queue with Balking and Reneging

In real practice, it often happens that arrivals become **discouraged or balked** when the queue is long and do not wish to wait. One such model is the $M/M/c/K$ that is, if people see K ahead of them in the system, they do not join. Generally, unless K is the result of a physical restriction such as no more places to park or room to wait, people will not act quite like that voluntarily. Rarely do all customers have exactly the same discouragement limit all the time. Another approach to balking is to employ a series of monotonically decreasing functions of the system size multiplying the rate λ. Let b_k be this function, so that $\lambda_k = b_n\lambda$ and $b_{k+1} \leq b_k \leq 1, k > 0, b_0 = 1$, that is the probability of joining the system provided it is in state k.

Possible examples that may be useful for the $b_k = 1/(k+1), k = 1, ..., K$. People are not always discouraged because of queue size, but may attempt to estimate how long they would have to wait. If the queue moving quickly, then the person may join a long one. On the other hand, if the queue is slow-moving, a customer may become discouraged even if the line is short. Now if k customers are in the system, an estimate for the average waiting time might be $k/c\mu$, if the customer had an idea of μ. In this case $b_k = e^{-\frac{k\alpha}{c\mu}}$. The $M/M/c/K$ system can be obtained as $b_k = 1, k = 0, ..., K$.

Customers who tend to be impatient may not always be discouraged by excessive queue size, but may instead join the queue to see how long their wait may become. However, they **renege, abandon** if their estimate of their total wait is intolerable and they leave the system without service.

Let $r_k h + o(h) =$ probability of reneging during h given k customers in the system, that is the reneging intensity is r_k. A good possibility for the reneging function r_k is
$r_k = 0,\ k = 0, \cdots, K$ classical system, $r_k = (k-r)\theta$, $r_k = e^{\frac{k\alpha}{c\mu}}, k = c, ...K$, and zero otherwise, where θ is the parameter of the exponentially distributed impatience time of a customer.
It is not so difficult to see, that the number of customers in the systems is a birth-death process with

$$\lambda_k = \lambda b_k,\ k = 0, \cdots, K-1$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \cdots, c \\ c\mu + r_k, & k = c, \cdots, K. \end{cases}$$

As usual, the steady-state distribution can be obtained as

$$P_k = \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} P_0, \qquad P_0 = \left(1 + \sum_{j=1}^{K} \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j}\right)^{-1}.$$

The **main performance measures** can be calculated as follows

$$U_r = 1 - P_0,\ E(\delta_r) = \frac{1}{\lambda} \cdot \frac{U_s}{1 - U_s}$$
$$\bar{N} = \sum_{k=1}^{K} kP_k,\ \bar{Q} = \sum_{k=c}^{K} (k-c)P_k$$
$$\bar{N^2} = \sum_{k=1}^{K} k^2 P_k,\ \bar{Q^2} = \sum_{k=c}^{K} (k-c)^2 P_k$$

$$Var(N) = \bar{N^2} - (\bar{N})^2,\ Var(Q) = \bar{Q^2} - (\bar{Q})^2$$
$$\bar{c} = \sum_{k=1}^{c-1} kP_k + \sum_{k=c}^{K} cP_k,\ \bar{N} = \bar{Q} + \bar{c},\ U_c = \bar{c}/c$$
$$\bar{\lambda} = \sum_{k=0}^{K-1} \lambda_k P_k,\ \bar{\mu} = \sum_{k=1}^{K} \mu_k P_k,\ \bar{\lambda} = \bar{\mu}$$
$$\bar{T} = \bar{N}/\bar{\lambda},\ \bar{W} = \bar{Q}/\bar{\lambda}$$
$$\bar{r} = \sum_{k=c}^{K} r_k P_k,\ \text{mean reneging rate}$$

The probability that an entering customer finds k customers in the system is

$$\Pi_k = \frac{\lambda_k P_k}{\bar{\lambda}}, \qquad k = 0, \ldots K-1.$$

$$P(\text{an arriving customer enters the system}) = \frac{\bar{\lambda}}{\lambda},$$

$$P(\text{a departing customer leaves the system without service}) = \frac{\bar{r}}{\bar{\mu}}$$

$$P(\text{waiting}) = \sum_{k=c}^{K-1} \Pi_k, \qquad P(\text{blocking}) = \frac{\lambda_K P_K}{\sum_{k=0}^{K} \lambda_k P_k}.$$

In the case of a balking system we can calculate the variance of waiting and response time and the distribution function of the waiting time, too.
Namely, we have

$$\overline{W} = \frac{\overline{Q}}{\overline{\lambda}} = \sum_{k=c}^{K-1} \frac{(k-c+1)}{(c\mu)} \Pi_k$$
$$\overline{T} = \frac{\overline{N}}{\overline{\lambda}} = \overline{W} + 1/\mu$$

Since the conditional waiting time is Erlang distributed, it is easy to see that

$$E(W^2) = \sum_{k=c}^{K-1} \frac{(k-c+1)+(k-c+1)^2}{(c\mu)^2}\Pi_k, \quad Var(W) = E(W^2) - (E(W))^2,$$

$$Var(T) = Var(W) + 1/\mu^2.$$

Distribution function of the waiting time
As in the previous parts for $F_W(t)$ the theorem of total probability is applied resulting

$$\begin{aligned} F_W(t) &= F_W(0) + \sum_{n=c}^{K-1} \Pi_n \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx \\ &= F_W(0) + \sum_{n=c}^{K-1} \Pi_n \left(1 - \int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx\right). \end{aligned}$$

Similarly to the previous section we have

$$\begin{aligned} F_W(t) &= F_W(0) + \sum_{n=c}^{K-1} \Pi_n - \sum_{n=c}^{K-1} \Pi_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!} \\ &= 1 - \sum_{n=c}^{K-1} \Pi_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}. \end{aligned}$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.11 The $M/G/1$ Queue

So far systems with exponentially distributed serviced times have been treated. We must admit that it is a restriction since in many practical problems these times are not exponentially distributed. It means that the investigation of queueing systems with generally distributed service times is natural. It is not the aim of this book to give a detailed analysis of this important system I concentrate only on the mean value approach and some practice oriented theorems are stated without proofs. A simple proof for the Little's law is also given.

Little's Law

As a first step for the investigations let us give a simple proof for the **Little's theorem, Little's law, Little's formula**, which states a relation between the mean number of customers in the systems, mean arrival rate and the mean response time. Similar version can be stated for the mean queue length, mean arrival rate and mean waiting time.

Let $\alpha(t)$ denote the number of customers arrived into the system in a time interval $(0,t)$, and let $\delta(t)$ denote the number of departed customers in $(0,t)$. Supposing that $N(0)=0$, the number of customers in the system at time t is $N(t)=\alpha(t)-\delta(t)$.
Let the mean arrival rate into the system during $(0,t)$ be defined as

$$\bar{\lambda}_t := \frac{\alpha(t)}{t}.$$

Let $\gamma(t)$ denote the overall sojourn times of the customers until t and let $\overline{T}_t$ be defined as the mean sojourn time for a request. Clearly

$$\overline{T}_t = \frac{\gamma(t)}{\alpha(t)}.$$

Finally, let $\bar{N}_t$ denote the mean number of customers in the system during in the interval $(0,t)$, that is

$$\bar{N}_t = \frac{\gamma(t)}{t}.$$

From these relations we have

$$\bar{N}_t = \bar{\lambda}_t \bar{T}_t.$$

Supposing that the following limits exist

$$\bar{\lambda} = \lim_{t\to\infty} \bar{\lambda}_t, \qquad \bar{T} = \lim_{t\to\infty} \bar{T}_t.$$

we get

$$\bar{N} = \bar{\lambda}\bar{T},$$

which is called **Little's law** .

Similar version is

$$\bar{Q} = \bar{\lambda}\bar{W}.$$

The Embedded Markov Chain

As before let $N(t)$ denote the number of customers in the system at time t. As time evolves the state changes and we can see that changes to neighboring states occur, up and down, that is from state k either to $k+1$ or to $k-1$. Since we have a single server the number of $k \to k+1$ type transitions may differ by at most one from the number of $k+1 \to k$ type transitions. So if the system operate for a long time the relative frequencies should be the same. It means that in stationary case the distributions at the arrival instants and the departure instants should be the same. More formally, $\Pi_k = D_k$.

For further purposes we need the following statements
Statement 1 For Poisson arrivals

$$P\left(N\left(t\right)=k\right)=P\left(\text{an arrival at time } t \text{ finds } \ k \text{ customers in the system }\right).$$

Statement 2 If in any system $N\left(t\right)$ changes its states by one then if either one of the following limiting distribution exists, so does the other and they are equal.

$$\Pi_k := \lim_{t\to\infty}\left(\text{an arrival at time } t \text{ finds } \ k \text{ customers in the system }\right),$$

$$D_k := \lim_{t\to\infty}\left(\text{a departure at time } t \text{ leaves } \ k \text{ customers behind }\right),$$

$$\Pi_k = D_k.$$

Thus for an $M/G/1$ system

$$\Pi_k = P_k = D_k,$$

that is in stationary case these 3 types of distributions are the same.

Due to their importance we prove them. Les us consider first Statement 1.
Introduce the following notation

$$P_k(t) := P\left(N\left(t\right)=k\right),$$

$$\Pi_k(t) := P\left(\text{an arriving customer at instant } \ t \text{ finds } \ k \text{ customers in the system }\right).$$

Let $A(t,t+\Delta t)$ be the event that one arrival occurs in the interval $(t,t+\Delta t)$. Then

$$\Pi_k(t) = \lim_{\Delta t\to 0} P\left(N\left(t\right)=k \mid A(t,t+\Delta t)\right).$$

By the definition of the conditional probability we have

$$\Pi_k(t) = \lim_{\Delta t\to 0} \frac{P\left(N\left(t\right)=k, A(t,t+\Delta t)\right)}{P\left(A(t,t+\Delta t)\right)} =$$

$$= \lim_{\Delta t\to 0} \frac{P\left(A(t,t+\Delta t) \mid N\left(t\right)=k\right) P\left(N(t)=k\right)}{P\left(A(t,t+\Delta t)\right)}.$$

Due to the memoryless property of the exponential distribution event $A(t,t+\Delta t)$ does not depend on the number of customers in the systems and even on t itself thus

$$P\left(A(t,t+\Delta t) \mid N\left(t\right)=k\right) = P\left(A(t,t+\Delta t)\right),$$

hence

$$\Pi_k(t) = \lim_{\Delta t \to 0} P\left(N(t) = k\right),$$

that is

$$\Pi_k(t) = P_k(t).$$

This holds for the limiting distribution as well, namely

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

Let us prove Statement 2 by the help of Statement 1 .

Let $\hat{R}_k\left(t\right)$ denote the number of arrivals into the system when it is in state k during the time interval $(0,t)$ and let $\hat{D}_k\left(t\right)$ denote the number of departures that leave the system behind in state k during $(0,t)$. Clearly

$$|\hat{R}_k\left(t\right) - \hat{D}_k\left(t\right)| \leq 1. \tag{2.30}$$

Furthermore if the total number of departures is denoted by $D\left(t\right)$, and the total number of arrivals is denoted by $R\left(t\right)$ then

$$D\left(t\right) = R\left(t\right) + N\left(0\right) - N\left(t\right).$$

The distribution at the departure instants can be written as

$$D_k = \lim_{t\to\infty} \frac{\hat{D}_k\left(t\right)}{D\left(t\right)}.$$

It is easy to see that the after simple algebra we have

$$\frac{\hat{D}_k\left(t\right)}{D\left(t\right)} = \frac{\hat{R}_k\left(t\right) + \hat{D}_k\left(t\right) - \hat{R}_k\left(t\right)}{R\left(t\right) + N\left(0\right) - N\left(t\right)}.$$

Since $N\left(0\right)$ is finite and $N\left(t\right)$ is also finite due to the stationarity from (2.30) and $\hat{R}\left(t\right) \to \infty$, with probability one follows that

$$D_k = \lim_{t\to\infty} \frac{\hat{D}_k\left(t\right)}{D\left(t\right)} = \lim_{t\to\infty} \frac{\hat{R}_k\left(t\right)}{R\left(t\right)} = \Pi_k.$$

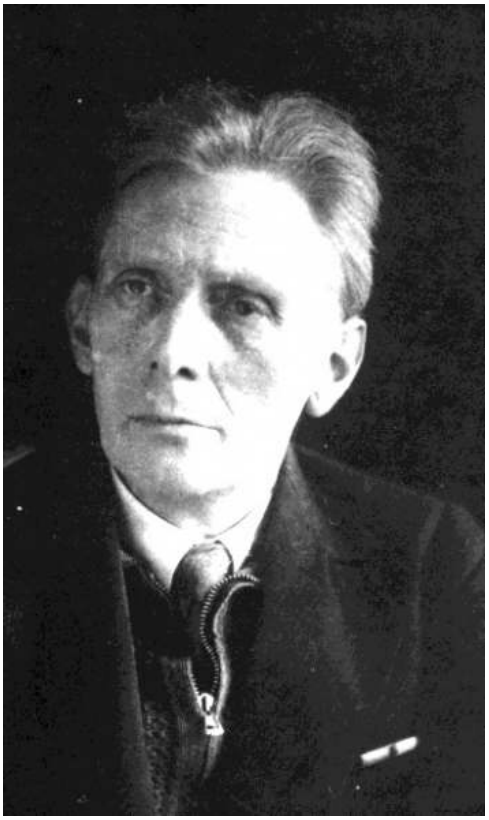
Consequently, by using Statement 1 the equality of the three probabilities follows.

Mean Value Approach

Let S denote the service time and let R denote the residual (remaining) service time. It can be proved that the systems is stable if $\rho = \lambda E(S) < 1$, furthermore $P_0 = 1 - \rho$. Then it can easily be seen that

$$\begin{aligned}\mathbb{E}(W) &= \sum_{k=1}^{\infty} \left(\mathbb{E}(R) + (k-1)\mathbb{E}(S)\right)\Pi_k \\ &= \sum_{k=1}^{\infty} \mathbb{E}(R)P_k + \left(\sum_{k=1}^{\infty}(k-1)P_k\right)\mathbb{E}(S) = \mathbb{E}(R)\rho + \mathbb{E}(Q)\mathbb{E}(S),\end{aligned}$$

Felix Pollaczek, 1892-1981



Alexander Y. Khintchine, 1894-1959

where $\mathbb{E}(R)$ denotes the mean residual time.
By applying the Little's law we have

$$\mathbb{E}(Q) = \lambda \mathbb{E}(W),$$

and thus

$$\mathbb{E}(W) = \frac{\rho \mathbb{E}(R)}{1-\rho} \tag{2.31}$$

known as **Pollaczek-Khintchine mean value formula**.

In subsection 2.11 we will show that

$$\mathbb{E}(R) = \frac{\mathbb{E}(S^2)}{2\mathbb{E}(S)}\,, \tag{2.32}$$

which can be written as

$$\mathbb{E}(R) = \frac{\mathbb{E}(S^2)}{2\mathbb{E}(S)} = \frac{Var(S) + \mathbb{E}^2(S)}{2\mathbb{E}(S)} = \frac{1}{2}(C_S^2 + 1)\mathbb{E}(S), \tag{2.33}$$

where C_S^2 is the squared coefficient of the service time S. It should be noted that mean residual service time depends on the first two moments of the service time.

Thus for the mean waiting time we have

$$\mathbb{E}(W) = \frac{\rho\mathbb{E}(R)}{1-\rho} = \frac{\rho}{2(1-\rho)}(C_S^2+1)\mathbb{E}(S).$$

By using the Little's law for the mean queue length we get

$$\mathbb{E}(Q) = \frac{\rho^2}{1-\rho}\frac{C_S^2+1}{2}.$$

Clearly, the mean response time and the mean number of customers in the systems can be expressed as

$$\mathbb{E}(T) = \frac{\rho}{1-\rho}\frac{C_S^2+1}{2}\mathbb{E}(S) + \mathbb{E}(S),$$

$$\mathbb{E}(N) = \rho + \frac{\rho^2}{1-\rho}\frac{C_S^2+1}{2},$$

which are also referred to as **Pollaczek-Khintchine mean value formulas**.

Example 23 *For an exponential distribution $C_S^2 = 1$, and thus $\mathbb{E}(R) = \mathbb{E}(S)$ which is evident from the memoryless property of the exponential distribution. In this case we get*

$$\mathbb{E}(W) = \frac{\rho}{1-\rho}\mathbb{E}(S), \quad \mathbb{E}(Q) = \frac{\rho^2}{1-\rho}, \quad \mathbb{E}(T) = \frac{1}{1-\rho}\mathbb{E}(S), \quad \mathbb{E}(N) = \frac{\rho}{1-\rho}.$$

Example 24 *In the case of deterministic service time $C_S^2 = 0$, thus $\mathbb{E}(R) = \mathbb{E}(S)/2$. Consequently we have*

$$\mathbb{E}(W) = \frac{\rho}{1-\rho}\frac{\mathbb{E}(S)}{2}, \quad \mathbb{E}(Q) = \frac{\rho^2}{2(1-\rho)}$$

$$\mathbb{E}(T) = \frac{1}{1-\rho}\frac{\mathbb{E}(S)}{2} + \mathbb{E}(S), \quad \mathbb{E}(N) = \rho + \frac{\rho^2}{2(1-\rho)}.$$

For an $M/G/1$ system we have proved that

$$\Pi_k = D_k = P_k, \qquad k = 0, 1, \ldots$$

therefore the generating function of the number of customers in the system is equal to the generating function of the number of customers at departure instant. Furthermore, it is clear that the number of customers at departure instants is equal the number customers arrived during the response time. In summary we have

$$D_k = P_k = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} f_T(x)\mathrm{d}x.$$

Thus the corresponding generating function can be obtained as

$$\begin{aligned} G_N(z) &= \sum_{k=0}^{\infty} z^k \int_0^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda x} f_T(x) \mathrm{d}x \\ &= \int_0^{\infty} \sum_{k=0}^{\infty} \frac{(\lambda x z)^k}{k!} e^{-\lambda x} f_T(x) \mathrm{d}x \\ &= \int_0^{\infty} e^{-\lambda(1-z)x} f_T(x) \mathrm{d}x = L_T(\lambda(1-z)), \end{aligned}$$

that is it can be expressed by the help of the Laplace-transform of the response time T. By applying the properties of the generating function and the Laplace-transform we have

$$G_N^{(k)}(1) = \mathbb{E}(N(N-1)\dots(N(-k+1))) = (-1)^k L_T^{(k)}(0)\lambda^k = \lambda^k \mathbb{E}(T^k).$$

In particular, the first derivative results to the Little's law, that is

$$\overline{N} = \lambda \overline{T},$$

and hence this formula can be considered as the generalization of the Little's law for an $M/G/1$ queueing systems.
By the help of this relation the higher moments of N can be obtained, thus the variance can be calculated if the second moment of T is known.

Residual Service Time

Let us suppose that the tagged customer arrives when the server is busy and denote the total service time of the request in service by X, that is a special interval. Let $f_X(x)$ denote the density function of X. The key observation to find $f_X(x)$ is that it is more likely that the tagged customer arrives in a longer service time than in a short one. Thus the probability that X is of length x should be proportional to the length x as well as the frequency of such service times, which is $f_S(x)\,dx$. Thus we may write

$$P(x \le X \le x + dx) = f_X(x)dx = Cx f_S(x)dx,$$

where C is a constant to normalize this density. That is

$$C^{-1} = \int_{x=0}^{\infty} x f_S(x)dx = E(S),$$

thus

$$f_X(x) = \frac{x f_S(x)}{E(S)}\ .$$

$$\mathbb{E}(X) = \int_0^{\infty} x f_X(x)\,dx = \frac{1}{\mathbb{E}(S)} \int_0^{\infty} x^2 f_S(x)\,dx = \frac{\mathbb{E}(S^2)}{\mathbb{E}(S)}.$$

Since the tagged customers arrives randomly in service time S hence the mean residual can be obtained as

$$\mathbb{E}(R) = \frac{\mathbb{E}(X)}{2} = \frac{\mathbb{E}(S^2)}{2\mathbb{E}(S)}$$

Example 25 *Let the service time be Erlang distributed with parameters* (n, μ) *then*

$$\mathbb{E}(S) = \frac{n}{\mu}, \quad Var(S) = \frac{n}{\mu^2},$$

thus

$$\mathbb{E}(S^2) = Var(S) + \mathbb{E}^2(S) = \frac{n(1+n)}{\mu^2}$$

hence

$$\mathbb{E}(R) = \frac{1+n}{2\mu}.$$

It is easy to see that using this approach the the density function the residual service time can be calculated. Given that the tagged customer arrives in a service time of length x, the arrival moment will be a random point within this service time, that is it will be uniformly distributed within the service time interval $(0, x)$. Thus we have

$$P(x \leq X \leq x + dx, y \leq R \leq y + dy) = \frac{dy}{x} f_X(x)dx, qquad0 \leq y \leq x.$$

After substitution for $f_X(x)$ and integrating over x we get the desired density function of the residual service time, that is

$$f_R(y) = \frac{1 - F_S(y)}{\mathbb{E}(S)}.$$

Hence

$$\mathbb{E}(R) = \int_0^\infty x f_R(x)dx = \int_0^\infty x \frac{1 - F_S(x)}{\mathbb{E}(S)} dx,$$

Thus

$$\mathbb{E}(R) = \frac{\mathbb{E}(S^2)}{2\mathbb{E}(S)}.$$

Now let us show how to calculate this type of integrals.
Let X be a non-negative random variable with finite nth moment. Then

$$\int_0^\infty x^n f(x)dx = \int_0^y x^n f(x)dx + \int_y^\infty x^n f(x)dx,$$

thus

$$\int_y^\infty x^n f(x)dx = \int_0^\infty x^n f(x)dx - \int_0^y x^n f(x)dx.$$

Since

$$\int_y^\infty x^n f(x)dx \geq y^n \int_y^\infty f(x)dx = y^n \left(1 - F\left(y\right)\right),$$

hence

$$0 \leq y^n \left(1 - F\left(y\right)\right) \leq \int_0^\infty x^n f(x)dx - \int_0^y x^n f(x)dx,$$

therefore

$$0 \leq \lim_{y\to\infty} y^n \left(1 - F\left(y\right)\right) \leq \int_0^\infty x^n f(x)dx - \lim_{y\to\infty} \int_0^y x^n f(x)dx,$$

that is

$$\lim_{y\to\infty} y^n \left(1 - F\left(y\right)\right) = 0.$$

Then using integration by parts keeping in mind the above relation we get

$$\int_0^\infty x^{n-1}(1 - F(x))dx = \int_0^\infty \frac{x^n}{n} f(x)dx = \frac{\mathbb{E}(X^n)}{n}.$$

Let us show another way to calculate this type of integral

$$\int_0^\infty x^{n-1} f_R(x)(x)dx = \frac{1}{E(S)} \int_0^\infty x^{n-1}(1 - F_S(x))dx = \frac{1}{E(S)} \int_{x=0}^\infty x^{n-1} \left[\int_{y=x}^\infty f_S(y)d(y)\right] dx$$

$$= \frac{1}{E(S)} \int_{y=0}^\infty \left[\int_{x=0}^y x^{n-1}dx\right] f_S(y)d(y) = \frac{1}{E(S)} \int_{y=0}^\infty \frac{y^n}{n} f_S(y)d(y) = \frac{E(S^n)}{nE(S)}.$$

In particular, for $n = 2$ we obtain

$$\mathbb{E}(R) = \frac{\mathbb{E}(S^2)}{2\mathbb{E}(S)}.$$

Pollaczek-Khintchine and Takács formulas

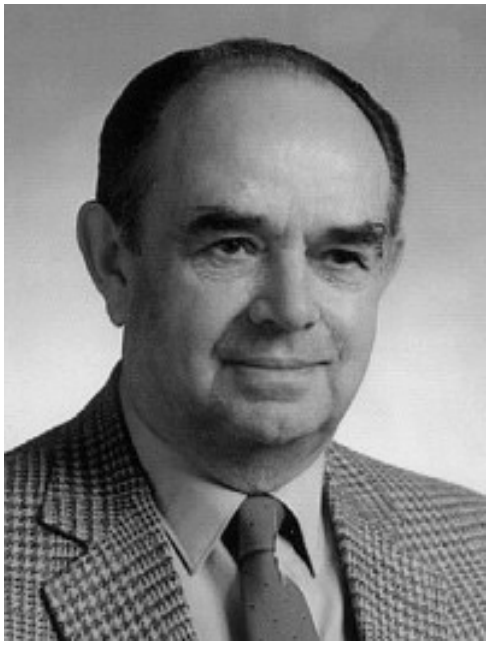
The following relations are commonly referred to as **Pollaczek-Khintchine transform equations**

$$G_N(z) = L_S(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{L_S(\lambda - \lambda z) - z}, \tag{2.34}$$

$$L_T(t) = L_S(t)\frac{t(1-\rho)}{t - \lambda + \lambda L_S(t)}, \tag{2.35}$$

$$L_W(t) = \frac{t(1-\rho)}{t - \lambda + \lambda L_S(t)}, \tag{2.36}$$

with the help of which, in principle, the distribution of the number of customers in the system, the density function of the response and waiting times can be obtained. Of course this time we must be able to invert the involved Laplace-transforms.



Lajos Takács, 1924-2015

Takács Recurrence Theorem

$$\mathbb{E}(W^k) = \frac{\lambda}{1-\rho} \sum_{i=1}^{k} \binom{k}{i} \frac{\mathbb{E}(S^{i+1})}{i+1} \mathbb{E}(W^{k-i}) \tag{2.37}$$

that is moments of the waiting time can be obtained in terms of lower moments of the waiting time and moments of the service time. It should be noted to get the kth moment of W the $k+1$th moment of the service time should exist.

Since W,S are independent and $T = W + S$ the kth moment of the response time can also be computed by

$$\mathbb{E}(T^k) = \sum_{l=0}^{k} \binom{k}{l} \mathbb{E}(W^l) \cdot \mathbb{E}(S^{k-l}). \tag{2.38}$$

By using these formulas the following relations it can be proved

$$\begin{aligned}
\mathbb{E}(W) &= \frac{\lambda\mathbb{E}(S^2)}{2(1-\rho)} = \frac{\rho\mathbb{E}(S)}{1-\rho}\left(\frac{1+C_S^2}{2}\right),\\
\mathbb{E}(T) &= \mathbb{E}(W) + \mathbb{E}(S),\\
\mathbb{E}(W^2) &= 2(\overline{W})^2 + \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)},\\
\mathbb{E}(T^2) &= \mathbb{E}(W^2) + \frac{\mathbb{E}(S^2)}{1-\rho},\\
Var(W) &= \mathbb{E}(W^2) - (\mathbb{E}(W))^2,\\
Var(T) &= Var(W+S) = Var(W) + Var(S).
\end{aligned}$$

Because

$$\mathbb{E}(N(N-1)) = \lambda^2\mathbb{E}(T^2)$$

after elementary but lengthy calculation we have

$$Var(N) = \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)} + \left(\frac{\lambda\mathbb{E}(S^2)}{2(1-\rho)}\right)^2 + \frac{\lambda(3-2\rho)\mathbb{E}(S^2)}{2(1-\rho)} + \rho(1-\rho).$$

Since

$$\begin{aligned}
\mathbb{E}(Q^2) &= \sum_{k=1}^{\infty}(k-1)^2P_k = \sum_{k=1}^{\infty}k^2P_k - 2\sum_{k=1}^{\infty}kP_k + \sum_{k=1}^{\infty}P_k\\
&= \mathbb{E}(N^2) - 2\overline{N} + \rho
\end{aligned}$$

by elementary computations we can prove that

$$Var(Q) = \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)} + \left(\frac{\lambda\mathbb{E}(S^2)}{2(1-\rho)}\right)^2 + \frac{\lambda\mathbb{E}(S^2)}{2(1-\rho)}.$$

Now let us turn our attention to the Laplace-transform of the busy period of the server.

Lajos Takács proved that

$$L_\delta(t) = L_S(t + \lambda - \lambda L_\delta(t)), \tag{2.39}$$

that is for the Laplace-transform $L_\delta(t)$ a function equation should be solved (which is usually impossible to invert).
However, by applying this equation the moments the busy period can be calculated.

First determine $\mathbb{E}(\delta)$. Using the properties of the Laplace-transform we have

$$\begin{aligned}
L'_\delta(0) &= (1 - \lambda L'_\delta(0))L'_S(0)\\
\mathbb{E}(\delta) &= (1 + \lambda\mathbb{E}(\delta))\mathbb{E}(S)\\
\mathbb{E}(\delta) &= \frac{\mathbb{E}(S)}{1-\rho} = \frac{1}{\lambda}\frac{\rho}{1-\rho}
\end{aligned}$$

which was obtained earlier by the well-known relation

$$\frac{\mathbb{E}(\delta)}{\frac{1}{\lambda}+\mathbb{E}(\delta)}=\rho.$$

After elementary but lengthy calculations it can be proved that

$$Var(\delta)=\frac{\mathbb{E}(S^2)}{(1-\rho)^3}-\frac{(\mathbb{E}(S))^2}{(1-\rho)^2}=\frac{Var(S)+\rho(\mathbb{E}(S))^2}{(1-\rho)^3}.$$

Now let us consider the generating function of the customers served during a busy period. It can be proved that

$$G_{N_d(\delta)}(z)=zL_S(\lambda-\lambda G_{N_d(\delta)}(z)) \tag{2.40}$$

which is again a functional equation but using derivations the higher moments can be computed.
Thus for the mean numbers we have

$$\begin{aligned}\mathbb{E}(N_d(\delta))&=1+\lambda\mathbb{E}(S)\mathbb{E}(N_d(\delta))\\ \mathbb{E}(N_d(\delta))&=\frac{1}{1-\rho},\end{aligned}$$

which can also be obtained by relation

$$\mathbb{E}(\delta)=\mathbb{E}(S)\mathbb{E}(N_d(\delta))$$

since

$$\begin{aligned}\frac{1}{\lambda}\frac{\rho}{1-\rho}&=\mathbb{E}(S)\cdot\mathbb{E}(N_d(\delta))\\ \mathbb{E}(N_d(\delta))&=\frac{\rho}{\rho(1-\rho)}=\frac{1}{1-\rho}.\end{aligned} \tag{2.41}$$

It can be proved that

$$Var(N_d(\delta))=\frac{\rho(1-\rho)+\lambda^2\mathbb{E}(S^2)}{(1-\rho)^3}. \tag{2.42}$$

It is interesting to note that the computation of $Var(\delta)$, $Var(N_d(\delta))$ does not require the existence of $\mathbb{E}(S^3)$, as it in the case of $Var(N)$, $Var(Q)$, $Var(T)$, $Var(W)$.

$M/G/1$ system with non-preemptive LCFS service discipline

In the following we show how the results concerning to the busy period analysis of a FCFS system can be used for the investigation of the waiting and response time of a system with non-preemptive LCFS (Last-Come- First-Served) service order. This means that the last customer does not interrupt the service of the current customer.

It should be noted that the mean waiting and response time of an $M/G/1$ under any well-known service discipline will be the same due to the Little-formula and the fact that the generating function of the steady-state distribution of the number of customers in the system is the same. As a consequence the mean number of customers in the system and the mean busy period length is the same. Moreover, the Laplace-transform of the busy period is the same, too. However, the higher moment will be different depending on the service order.

It can be proved that for $M/G/1$ systems we have

$$L_{W_{LCFS}}(t) = (1-\rho) + \rho \frac{1 - L_\delta(t)}{(t + \lambda - \lambda L_\delta(t))E(S)}$$

$$L_{T_{LCFS}}(t) = L_{W_{LCFS}}(t) L_S(t)$$

$$Var(W_{LCFS}) = \frac{\lambda E(S^3)}{3(1-\rho)^2} + \frac{\lambda^2(1+\rho)(E(S^2))^2}{4(1-\rho)^3}$$

$$Var(W_{FCFS}) = \frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2(E(S^2))^2}{4(1-\rho)^2}$$

$$Var(W_{SIRO}) = \frac{2\lambda E(S^3)}{3(1-\rho)(2-\rho)} + \frac{\lambda^2(2+\rho)(E(S^2))^2}{4(1-\rho)^2(2-\rho)}$$

Comparing the formulas term-by-term it is not difficult to prove that

$$Var(W_{FCFS}) < Var(W_{SIRO}) < Var(W_{LCFS})$$
$$Var(T_{FCFS}) < Var(W_{SIRO}) < Var(W_{LCFS})$$

As it is one of the most widely used queueing system the calculation of the main performance measure is of great importance. It can be done by the help of our Java applets

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.12 The $M/G/1$ Priority Queue

$M/G/1$ Queueing Systems (classes, no priority)

There are n customer classes. Customers from class i arrive in a Poisson pattern with mean arrival rate $\lambda_i, i = 1, 2, \ldots, n$. Each class has its own general service time with $\mathbb{E}[S_i] = 1/\mu_i, \quad \mathbb{E}[S_i^2], \quad \mathbb{E}[S_i^3]$. All customers served on a FCFS basis with no consideration for class. The total arrival stream to the system has a Poisson arrival pattern with

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

The first three moments of service time arc given by

$$\overline{S} = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n],$$

$$\mathbb{E}[S^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^2],$$

and

$$\mathbb{E}[S^3] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^3] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^3] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^3],$$

By Pollaczek's formula,

$$\overline{W} = \frac{\lambda\mathbb{E}[S^2]}{2(1-\rho)}.$$

The mean time in the system for each class is given by

$$\overline{T}_i = \overline{W} + \mathbb{E}[S_i], \quad i = 1, 2, \ldots, n.$$

The overall mean customer time in the system,

$$\overline{T} = \frac{\lambda_1}{\lambda}\overline{T}_1 + \frac{\lambda_2}{\lambda}\overline{T}_2 + \ldots + \frac{\lambda_n}{\lambda}\overline{T}_n.$$

The variance of the waiting time

$$Var(W) = \frac{\lambda\mathbb{E}[S^3]}{3(1-\rho)} + \frac{\lambda^2(\mathbb{E}[S^2])^2}{4(1-\rho)^2}.$$

The variance of T is given by

$$Var(T_i) = Var(W) + Var(S_i), \quad i = 1, 2, \ldots, n.$$

The second moment of T by class is

$$\mathbb{E}[T_i^2] = Var(T_i) + \overline{T}_i^2, \quad i = 1, 2, \ldots, n.$$

Thus, the overall second moment of T is given by

$$\mathbb{E}[T^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[T_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[T_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[T_n^2],$$
$$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

$M/G/1$ Non-preemptive (HOL) Priority Queueing Systems

There are n priority classes with each class having a Poisson arrival pattern with mean arrival rate λ_i. Each customer has the same exponential service time requirement. Then the overall arrival pattern is Poiisson with mean:

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

The server utilization

$$\overline{S} = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n],$$
$$\mathbb{E}[S^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^2],$$

and

$$\mathbb{E}[S^3] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^3] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^3] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^3],$$

Let

$$\rho_j = \lambda_1\mathbb{E}[S_1] + \lambda_2\mathbb{E}[S_2] + \ldots + \lambda_j\mathbb{E}[S_j], \quad j = 1, 2, \ldots, n,$$

and notice that

$$\rho_n = \rho = \lambda\overline{S}.$$

The mean times in the queues:

$$\overline{W}_j = \mathbb{E}[W_j] = \frac{\lambda\mathbb{E}[S^2]}{2(1-\rho_{j-1})(1-\rho_j)},$$
$$j = 1, 2, \ldots, n, \quad \rho_0 = 0.$$

The mean queue lengths are

$$\overline{Q}_j = \overline{\lambda}_j \cdot \overline{W}_j, \quad j = 1, 2, \ldots, n.$$

The unified time in the queue

$$\overline{W} = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n].$$

The mean times of staying in the system

$$\overline{T}_j = \mathbb{E}[T_j] = \mathbb{E}[W_j] + \mathbb{E}[S_j], \quad j = 1, 2, \ldots, n,$$

and the average of the customers staying at the system is

$$\overline{N}_j = \overline{\lambda}_j \cdot \overline{T}_j, \quad j = 1, 2, \ldots, n.$$

The total time in the system

$$\overline{T} = \overline{W} + \overline{S}.$$

The total queue length

$$\overline{Q} = \overline{\lambda} \cdot \overline{W},$$

and the average of the customers staying at the system

$$\overline{N} = \overline{\lambda} \cdot \overline{T}.$$

The variance of the total time stayed in the system by class

$$Var(T_j) = Var(S_j) + \frac{\lambda\mathbb{E}[S^3]}{3(1-\rho_{j-1})^2(1-\rho_j)}$$
$$+\frac{\lambda\mathbb{E}[S^2]\left(2\sum\limits_{i=1}^{j}\lambda_i\mathbb{E}[S_i^2] - \lambda\mathbb{E}[S^2]\right)}{4(1-\rho_{j-1})^2(1-\rho_j)^2}$$
$$+\frac{\lambda\mathbb{E}[S^2]\sum\limits_{i=1}^{j-1}\lambda_i\mathbb{E}[S_i^2]}{2(1-\rho_{j-1})^3(1-\rho_j)}, \quad j = 1, 2, \ldots, n.$$

The variance of the total time stayed in the system

$$Var(T) = \frac{\lambda_1}{\lambda}[Var(T_1) + \overline{T}_1^2] + \frac{\lambda_2}{\lambda}[Var(T_2) + \overline{T}_2^2]$$
$$+ \ldots + \frac{\lambda_n}{\lambda}[Var(T_n) + \overline{T}_n^2] - \overline{T}^2.$$

The variance of the waiting time by class

$$Var(W_j) = Var(T_j) - Var(S_j), \quad j = 1, 2, \ldots, n.$$

We know that $\mathbb{E}[W_j^2] = Var(W_j) + \overline{W}_j^2, \quad j = 1, 2, \ldots, n,$

so

$$\mathbb{E}[W^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n^2].$$

Finally

$$Var(W) = \mathbb{E}[W^2] - \overline{W}^2.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

$M/G/1$ Preemptive Resume Priority Queueing Systems

There are n customer classes. Class 1 customers receive the most favorable treatment; class n customers receive the least favorable treatment. Customers from class i arrive in a Poisson pattern with mean arrival rate λ_i,$t = 1, 2, \ldots, n$. Each class has its own general service time with $\mathbb{E}[S_i] = 1/\mu_i$, and finite second and third moments $\mathbb{E}[S_i^2]$, $\mathbb{E}[S_i^3]$. The priority system is preemptive resume, which means that if a customer of class j is receiving service when a customer of class $i < j$ arrives, the arriving customer preempts the server and the customer who was preempted returns to the head of the line for class j customers. The preempted customer resumes service at the point of interruption upon reentering the service facility. The total arrival stream to the system has a Poisson arrival pattern with

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

The first three moment of service time are given by:

$$\overline{S} = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n],$$
$$\mathbb{E}[S^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^2],$$
$$\mathbb{E}[S^3] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^3] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^3] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^3].$$

Let

$$\rho_j = \lambda_1\mathbb{E}[S_1] + \lambda_2\mathbb{E}[S_2] + \ldots + \lambda_j\mathbb{E}[S_j], \quad j = 1, 2, \ldots, n,$$

and notice that $\rho_n = \rho = \lambda\overline{S}$.

The mean time in the system for each class is

$$\overline{T}_j = \mathbb{E}[T_j] = \frac{1}{1-\rho_{j-1}} \left[\mathbb{E}[S_j] + \frac{\sum_{i=1}^{j} \lambda_i \mathbb{E}[S_i^2]}{2(1-\rho_j)} \right],$$

$\rho_0 = 0, \quad j = 1, 2, \ldots, n.$

Waiting times
$\overline{W}_j = \mathbb{E}[T_j] - \mathbb{E}[S_j], \quad j = 1, 2, \ldots, n.$

The mean length of the queue number j:

$$\overline{Q}_j = \lambda_j \overline{W}_j, \quad j = 1, 2, \ldots, n.$$

The total waiting time, $\overline{W}$, is given by:

$$\overline{W} = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n].$$

The mean number of customers staying in the system for each class is

$$\overline{N}_j = \lambda_j \overline{W}_j, \quad j = 1, 2, \ldots, n.$$

The mean total time is

$$\overline{T} = \frac{\lambda_1}{\lambda}\overline{T}_1 + \frac{\lambda_2}{\lambda}\overline{T}_2 + \ldots + \frac{\lambda_n}{\lambda}\overline{T}_n = \overline{W} + \overline{S}.$$

The mean number of customers waiting in the queue is

$$\overline{Q} = \overline{\lambda} \cdot \overline{W},$$

and the average number of customers staying in the system

$$\overline{N} = \overline{\lambda} \cdot \overline{T}.$$

The variance of the total time of staying in the system for each class is

$$Var(T_j) = \frac{Var(S_j)}{(1-\rho_{j-1})^2} + \frac{\mathbb{E}[S_j] \sum_{i=1}^{j-1} \lambda_i \mathbb{E}[S_i^2]}{(1-\rho_{j-1})^3}$$
$$+ \frac{\sum_{i=1}^{j} \lambda_i \mathbb{E}[S_i^3]}{3(1-\rho_{j-1})^2(1-\rho_j)} + \frac{\left(\sum_{i=1}^{j} \lambda_i \mathbb{E}[S_i^2]\right)^2}{4(1-\rho_{j-1})^2(1-\rho_j)^2}$$

$$+\frac{\left(\sum_{i=1}^{j}\lambda_i\mathbb{E}[S_i^2]\right)\left(\sum_{i=1}^{j-1}\lambda_i\mathbb{E}[S_i^2]\right)}{2(1-\rho_{j-1})^3(1-\rho_j)},\quad \rho_0=0,\quad j=1,2,\ldots,n.$$

The overall variance

$$Var(T)=\frac{\lambda_1}{\lambda}[Var(T_1)+\overline{T}_1^2]+\frac{\lambda_2}{\lambda}[Var(T_2)+\overline{T}_2^2]$$
$$+\ldots+\frac{\lambda_n}{\lambda}[Var(T_n)+\overline{T}_n^2]-\overline{T}^2.$$

The variance of waiting times for each class is

$$Var(W_j)=Var(T_j)-Var(S_j),\quad j=1,2,\ldots,n.$$

Because,

$$\mathbb{E}[W_j^2]=Var(W_j)+\overline{W}_j^2,\quad j=1,2,\ldots,n,$$

so

$$\mathbb{E}[W^2]=\frac{\lambda_1}{\lambda}\mathbb{E}[W_1^2]+\frac{\lambda_2}{\lambda}\mathbb{E}[W_2^2]+\ldots+\frac{\lambda_n}{\lambda}\mathbb{E}[W_n^2].$$

Finally

$$Var(W)=\mathbb{E}[W^2]-\overline{W}^2.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.13 The $M/G/c$ Processor Sharing Queue

$M/G/1$ Processor Sharing Queueing Systems

The Poisson arrival stream has an average arrival rate of λ and the average service rate is μ. The service time distribution is general with the restriction that its Laplace transform is rational, with the denominator having degree at least one higher than the numerator. Equivalently. the service time, s, is Coxian. The priority system is processor-sharing, which means that if a customer arrives when there are already $n-1$ customers in the system, the arriving customer (and all the others) receive service at the average rate μ/n. Then $P_n = \rho^n(1-\rho),\ n = 0, 1, \ldots,$ where $\rho = \lambda/\mu$. We also have

$$\overline{N} = \frac{\rho}{1-\rho}, \quad \mathbb{E}[T|S=t] = \frac{t}{1-\rho}, \text{ and } \overline{T} = \frac{\overline{S}}{1-\rho}.$$

Finally

$$\mathbb{E}[W|S=t] = \frac{\rho t}{1-\rho}, \text{ and } \overline{W} = \frac{\rho\overline{S}}{1-\rho}.$$

$M/G/c$ Processor Sharing Queueing Systems

The Poisson arrival stream has an average arrival rate of λ. The service time distribution is general with the restriction that its Laplace transform is rational, with the denominator having degree at least one higher than the numerator. Equivalently, the service time, s, is Coxian. The priority system is processor-sharing, which works as follows. When the number of customers in the service center, is less than c, then each customers is served simultaneously by one server; that is, each receives service at the rate μ. When $N > c$. each customer simultaneously receives service at the rate $c\mu/N$. We find that just as for the M/G/l processor-sharing system.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

2.14 The GI/M/1 Queue

We state the results for $GI/M/1$ queues, using the parameter σ given by

$$\sigma = L_A(\mu - \mu\sigma),$$

where $L_A(s)$ is the Laplace-transform of the interarrival time.
From now, $\rho = E(S)/E(A) = \lambda/\mu$.
For example, the distribution of customers in the system

$$P_k = \rho(1-\sigma)\sigma^{k-1}, \quad k > 0,$$
$$\pi_0 = 1-\rho,$$

the mean number of jobs in the system

$$\bar{N} = \frac{\rho}{1-\sigma},$$

the variance of the number of jobs in the system

$$Var(N) = \frac{\rho(1+\sigma-\rho)}{(1-\sigma)^2},$$

the mean response time

$$\bar{T} = \frac{1}{\mu} \cdot \frac{1}{1-\sigma},$$

the mean queue length

$$\bar{Q} = \frac{\rho \cdot \sigma}{1-\sigma},$$

the variance of the queue length

$$Var(Q) = \frac{\rho\sigma(1+\sigma(1-\rho))}{(1-\sigma)^2},$$

the mean waiting time

$$\bar{W} = \frac{1}{\mu} \cdot \frac{\sigma}{1-\sigma}.$$

Finally, for the waiting time distribution we have

$$F_W(x) = \begin{cases} 1-\sigma, & x = 0, \\ 1-\sigma \cdot e^{-\mu(1-\sigma)x}, & x > 0. \end{cases}$$

Let us see some examples.

In the case of an $M/M/1$ **queue**, we have $L_A(s) = \lambda/(s+\lambda)$, and it is easy to see that we obtain

$$\sigma = \frac{\lambda}{\mu} = \rho$$

and we have the well-known $M/M/1$ formulas.

A more interesting example is the $E_2/M/1$ **queue**, that is the Erlangian arrivals with 2 phases. In this case we get

$$L_A(s) = \left(\frac{\lambda}{s+\lambda}\right)^2.$$

Since $\sigma = 1$ is a solution of the resulting cubic equation it can be written in a product-form. Finding the roots reduces to the solution of the following quadratic equation

$$\mu^2\sigma - (2\lambda+\mu)\mu\sigma + \lambda^2 = 0$$

After elementary calculations we have

$$\sigma_{1,2} = \rho + \frac{1}{2} \pm \sqrt{\rho + \frac{1}{4}}.$$

Since $\sigma < 1$ we get

$$\sigma = \rho + \frac{1}{2} - \sqrt{\rho + \frac{1}{4}}.$$

It is not difficult to see that $0 < \sigma < 1$.

Finally, let us see a hipo-exponential distribution with 2 phases and parameters $(\mu, 2\mu)$.
This time we have to solve the following equation

$$\sigma = \frac{2\mu^2}{(\mu - \mu\sigma + \mu)((\mu - \mu\sigma + 2\mu)}.$$

After elementary calculations we get

$$\sigma^3 - 5\sigma^2 + 6\sigma - 2 = 0.$$

Since $\sigma = 1$ is a root of it, it remains to find the solution to

$$\sigma^2 - 4\sigma + 2 = 0.$$

For the roots we get

$$\sigma_{1,2} = 2 \pm \sqrt{2}$$

and the solution is $\sigma = 2 - \sqrt{2} < 1$.
Using the above formalas we obtain explicit expressions for the performance measures mentioned above.

The behaviour of an $M/G/1$ and of a $GI/M/1$ queue is very similar, especially if $C_S^2 \leq 1$. We compare the mean number of jobs $\bar{N}$ for $M/G/1$ and $GI/M/1$ queues having the same coefficient of variation C_X^2. Note that C_X^2 in the case of $GI/M/1$ denotes the coefficient of variation of the interarrival times, while in the $M/G/1$ case it denotes the coefficient of variation of the service times. Note also that in the $M/G/1$ case, $\bar{N}$ depends only on the first two moments of the service time distribution, while in the $GI/M/1$ case, the dependence is vis the Laplace-transform of the interarrival times.

2.15 Approximations

The level of this section is not basic, but without proof we give the corresponding approximations becuase we would like to show that in even more complicated cases we can give estimations the mean waiting time. But knowing the relation between the response, waiting and service times and applying the Little formula we can give approximations to the other measures, too.
The material of this section is based on Allen [3], and Bolch *et. al.* [12].

The GI/G/1 Queue

In the $GI/G/1$ case only approximation formulae and bounds exist. We can use $M/G/1$ and $GI/M/1$ results as upper or lower bounds, depending on the value of the coefficient of variation (see Table 2.3). Another upper bound is given by

Table 2.3: Upper bounds (UB) and lower bounds (LB) for the $GI/G/1$ mean waiting time

C_A^2	C_B^2	$M/G/1$	$GI/M/1$
1	1	LB	LB
1	1	LB	UB
1	1	UB	LB
1	1	UB	UB

$$\bar{W} < \frac{\sigma_A^2 + \sigma_B^2}{2(1-\rho)} \cdot \lambda.$$

A modification of this upper bound is in Bolch *et. al.* [12]

$$\bar{W} < \frac{1 + C_S^2}{(1/\rho^2) + C_S^2} \cdot \frac{\sigma_A^2 + \sigma_S^2}{2(1-\rho)} \cdot \lambda.$$

This formula is exact for $M/G/1$ and is a good approximation for $GI/M/1$ and $GI/G/1$ queues if ρ is not too small and C_A^2 or C_S^2 are not too big.
A lower bound is also known

$$\frac{\rho^2 \cdot C_S^2 + \rho(\rho - 2)}{2\lambda(1-\rho)} < \bar{W}$$

but more complex and better lower bounds are given in Kleinrock [62,63]. Many approximation formulae for the mean waiting time are mentioned in the literature. Four of them that are either very simple and straightforward or good approximations are introduced here. First, the well-known **Allen-Cunneen (AC)** approximation formula for $GI/G/m$ queue is in Allen [3]

$$\bar{W} \approx \frac{\rho/\mu}{1-\rho} \cdot \frac{C_A^2 + C_S^2}{2}. \tag{2.43}$$

This formula is exact for $M/G/1$ (Pollaczek-Khintchine formula) and a fair approximation elsewhere and is the basis for many other better approximations. A very good approximation is the **Kramer-Langenbach-Belz (KLB)** formula, a direct extension of Eq. 2.43 via a correction factor

$$\bar{W} \approx \frac{\rho/\mu}{1-\rho} \cdot \frac{C_A^2 + C_S^2}{2} \cdot G_{KLB}, \tag{2.44}$$

where the correction factor

$$G_{KLB} = \begin{cases} exp\left(-\frac{2}{3} \cdot \frac{1-\rho}{\rho} \cdot \frac{(1-C_A^2)^2}{C_A^2 + C_S^2}\right), & 0 \le C_A \le 1, \\ exp\left(-(1-\rho)\frac{C_A^2 - 1}{C_A^2 + 4C_S^2}\right), & C_A > 1. \end{cases} \tag{2.45}$$

Another extension of the Allen-Cunneen formula is the approximation of **Kulbatzki**

$$\bar{W} \approx \frac{\rho/\mu}{1-\rho} \cdot \frac{C_A^{f(C_A, C_S, \sigma)} + C_S^2}{2}, \tag{2.46}$$

with

$$f(C_A, C_S, \sigma) = \begin{cases} 1, & C_A \in [0,1], \\ [\rho(14.1C_A - 5.9) + (-13.7C_A + 4.1)]C_S^2 & \\ +[\rho(-59.7C_A + 21.1) + (54.9C_A - 16.3)]C_S & \\ +[\rho(C_A - 4.5) + (-1.5C_A + 6.55)], & 0 \leq C_A \leq 1, \\ -0.75\rho + 2.775, & C_A > 1, \end{cases} \tag{2.47}$$

It is interesting to note that Eq. 2.47 was obtained using simulation experiments.
A good approximation for the case $C_A^2 < 1$ is the Kimura approximation

$$\bar{W} \approx \frac{C_A^2 + C_S^2}{2} \bar{W}_{M/M/m} \left((1 - C_A^2) exp\left(\frac{2(1-\rho)}{3\rho} \right) + C_A^2 \right)^{-1} \tag{2.48}$$

The M/G/m Queue

We obtain the **Martin's** approximation formula for $M/G/m$ queues

$$\bar{W} = \bar{W}_0 + \frac{\bar{Q}}{m} \cdot \bar{T}.$$

Because of the m servers, an arriving customer has to wait, on the average, only for the service of $\bar{Q}/m$ customers. The remaining service time in this case is

$$\bar{W}_0 = P_m \cdot \bar{R}. \tag{2.49}$$

With

$$\bar{R} \approx \bar{T} \frac{(1 + C_S^2)}{2m} \tag{2.50}$$

we get

$$\bar{W} \approx \frac{P_m/\mu}{1 - \rho} \cdot \frac{(1 + C_S^2)}{2m}. \tag{2.51}$$

This is a special case of the Allen-Curineen formula for $GI/G/m$ queues and is exact for $M/M/m$ and $M/G/1$ queues. For the waiting probability P_m, we can use

$$P_m \approx \begin{cases} \frac{\rho^m + \rho}{2}, & \rho > 0.7, \\ \rho^{\frac{m+1}{2}}, & \rho < 0.7. \end{cases} \tag{2.52}$$

As an example, we compare the exact waiting probability with this approximation for $m = 5$ in Table 2.4. A good approximation for the mean waiting time in $M/G/m$ queues is due to **Cosmetatos**

$$\bar{W}_{M/G/m} \approx C_S^2 \bar{W}_{M/M/m} + (1 - C_S^2) \bar{W}_{M/D/m}$$

Use

$$\bar{W}_{M/D/m} = \frac{1}{2} \cdot \frac{1}{nc_{Dm}} \cdot \bar{W}_{M/M/m}$$

Table 2.4: approximate values of the probability of waiting

ρ	0.2	0.4	0.6	0.7	0.8	0.9	0.95	0.99
$\rho_{m_{ex}}$	0	0.06	0.23	0.38	0.55	0.76	0.88	0.97
$\rho_{m_{app}}$	0	0.06	0.21	0.34	0.56	0.75	0.86	0.97

where

$$nc_{Dm} = \left(1 + (1-\rho)(m-1)\frac{\sqrt{4+5m}-2}{16\rho m}\right)^{-1}$$

For $\bar{W}_{M/D/m}$ we can also use the **Crommelin** approximation formula

$$\bar{W}_{M/D/m} \approx \frac{1}{\mu}\sum_{k=1}^{\infty}\left(e^{-k\rho m}\left(\frac{(k\rho m)^{km}}{(km)!} - (1-1/\rho)\sum_{i=1}^{km}\frac{(k\rho m)^i}{i!}\right) + \left(1-\frac{1}{\rho}\right)\right)$$

Boxma, Cohen, and Huffels (BCH-formula) also use the preceding formulae for $\bar{W}_{M/D/m}$ as a basis for their approximation:

$$\bar{W}_{M/G/m} \approx \frac{1}{2}(1+C_S^2)\frac{2\bar{W}_{M/D/m}\bar{W}_{M/M/m}}{2a\bar{W}_{M/D/m} + (1+a)\bar{W}_{M/M/m}}$$

Where

$$a = \begin{cases} 1, & m = 1, \\ \frac{1}{1-m}\left(\frac{(C_S^2+1)}{\gamma_1} - m + 1\right), & m > 1, \end{cases}$$

and

$$\gamma_1 \approx \frac{1-C_S^2}{m+1} + \frac{C_S^2}{m}.$$

Tijms uses γ_1 from the BCH-formula in his approximation:

$$\bar{W}_{M/G/m} \approx \left((1-\rho)\gamma_1 m + \frac{\rho}{2}(C_S^2+1)\right)\bar{W}_{M/M/m}$$

The GI/G/m Queue

For $GI/G/m$ queues only bounds and approximation formulae are available. These are extensions of $M/G/m$ or $GI/G/1$ formulae. We begin with the well-known upper bound due to **Kingman**

$$\bar{W} \leq \frac{\sigma_A^2 + \sigma_B^2/m + (m-1)/(m^2\cdot\mu^2)}{2(1-\rho)} \cdot \lambda$$

and the lower bound of **Brumelle and Marchal**

$$\bar{W} \geq \frac{\rho^2 C_S^2 - \rho(2-\rho)}{2\lambda(1-\rho)} - \frac{m-1}{m}\cdot\frac{C_S^2+1}{2\mu}$$

As a heavy traffic approximation we have for $GI/M/m$ queues:

$$\bar{W} \approx \frac{\sigma_A^2 + \sigma_B^2/m^2}{2(1-\rho)} \cdot \lambda,$$

and we have the **Kingman-Kollerstrom** approximation for the waiting time distribution:

$$F_W(x) \approx 1 - exp\left(-\frac{2(1-\rho)}{\sigma_A^2 + \sigma_B^2/m^2} \cdot \frac{1}{\lambda} x\right).$$

The most known approximation formula for $GI/G/m$ queues is the Allen-Cunneen formula. We already introduced it for the special case $GI/G/1$. Note that the A-C formula is an extension of Martin's formula where we replace the 1 in the term $(1 + C_S^2)$ by C_A^2 to consider approximately the influence of the distribution of interarrival times:

$$\bar{W} \approx \frac{P_m/\mu}{1-\rho} \cdot \frac{C_A^2 + C_S^2}{2m}$$

For the probability of waiting a good approximation is provided by Eq. 2.52. As in the $GI/G/1$ case, the Allen-Cunneen approximation was improved by Kramer Langenbach-Belz using a correction factor:

$$\bar{W} \approx \frac{P_m/\mu}{1-\rho} \cdot \frac{C_A^2 + C_S^2}{2m} \cdot G_{KLB}$$

$$G_{KLB} = \begin{cases} exp\left(-\frac{2}{3}\frac{1-\rho}{P_m}\frac{(1-C_A^2)^2}{C_A^2+C_S^2}\right), & 0 \le C_A \le 1, \\ exp\left(-(1-\rho)\frac{C_A^2-1}{C_A^2+4C_S^2}\right), & C_A > 1, \end{cases}$$

and by Kulbatzki using the exponent $f(C_A, C_S, \rho)$ in place of 2 for C_A in Eq. 2.15:

$$\bar{W} \approx \frac{P_m/\mu}{1-\rho} \cdot \frac{C_A^{f(C_A,C_S,\rho)} + C_S^2}{2m}$$

For the definition of $f(C_A, C_S, \rho)$, see Eq. 2.47.
We start with the Kulbatzki $GI/G/1$ formula and use a heuristic correction factor to consider the number of servers m:

$$\bar{W} \approx \frac{\rho/\mu}{1-\rho} \cdot \frac{C_A^{f(C_A,C_S,\rho)} + C_S^2}{2} \cdot \rho^{\sqrt{0.5(m-1)}}.$$

This formula is applicable even if the values of m and the coefficients of variation are large.
In order to extend the Cosmetatos approximation from $M/G/m$ to $GI/G/m$ queues, $\bar{W}_{M/M/m}$ and $\bar{W}_{M/D/m}$ need to be replaced by $\bar{W}_{GI/M/m}$ and $\bar{W}_{GI/D/m}$ respectively:

$$\bar{W}_{GI/G/m} \approx C_S^2 \bar{W}_{GI/M/m} + (1 - C_S^2)\bar{W}_{GI/D/m}$$

where $\bar{W}_{GI/M/m}$ is given by the approximation

$$\bar{W}_{GI/M/m} \approx \begin{cases} \frac{1}{2}(C_A^2+1)exp\left(\frac{2}{3} \cdot \frac{1-\rho}{P_m} \cdot \frac{(1-C_A^2)^2}{1+C_A^2}\right)\bar{W}_{M/M/m}, & \text{for } 0 \le C_A \le 1, \\ \frac{1}{2}\left(C_A^{f(C_A,C_S,\rho)} + 1\right)\bar{W}_{M/M/m} & \text{for } C_A > 1, \end{cases}$$

and $\bar{W}_{GI/D/m}$ is obtained by

$$\bar{W}_{GI/D/m} = \frac{1}{2} \cdot \frac{1}{nc_{Dm}} \cdot \bar{W}_{GI/M/m}$$

with nc_{Dm} from Eq. 2.15 or

$$\bar{W}_{GI/D/m} \approx C_A^{h(\rho,m)f(C_A,0,\rho)} \cdot \bar{W}_{M/D/m}.$$
$$h(\rho,m) = 4\sqrt{(m-1)/(m+4)} \cdot (1-\rho) + 1$$

with $\bar{W}_{M/D/m}$ from Eq. 2.15.
A good approximation for the case C_A^2 is given by Kimura

$$\bar{W}_{GI/G/m} = \frac{C_A^2 + C_S^2}{2}\bar{W}_{M/M/m}$$
$$\cdot \left(\frac{1-C_A^2}{1-4c(m,\rho)} exp\left(\frac{2(1-\rho)}{3\rho}\right) + \frac{1-C_S^2}{1+c(m,\rho)} + C_A^2 + C_S^2 - 1 \right)^{-1}$$
$$c(m,\rho) = (1-\rho)(m-1)\frac{\sqrt{4+5m}-2}{16\rho m}.$$

Finally, the Boxma, Cohen. and Huffels formula can be extended to $GI/G/m$ queues:

$$\bar{W}_{GI/G/m} = \frac{1}{2}(1+C_S^2)\frac{2\bar{W}_{GI/D/m}\bar{W}_{GI/M/m}}{2a\bar{W}_{GI/D/m} + (1-a)\bar{W}_{GI/M/m}}$$

as well as the **Tijms** formula

$$\bar{W}_{GI/G/m} = \left((1-\rho)\gamma_1 m + \frac{\rho}{2}(C_S^2+1)\right)\bar{W}_{GI/M/m}$$

with a and γ_1 from Eq. 2.15

Chapter 3

Finite-Source Systems

So far we have been dealing with such queueing systems where arrivals followed a Poisson process, that is the source of customers is infinite. In this chapter we are focusing on the *finite-source population* models. They are also very important from practical point of view since in many situation the source is finite. Let us investigate the example of the so-called *machine interference problem* treated by many experts.

Let us consider n machines that operates independently of each other. The operation times and service times are supposed to be independent random variables with given distribution function. After failure the broken machines are repaired by a single or multiple repairmen according to a certain discipline. Having been repaired the machine starts operating again and the whole process is repeated.

This simple model has many applications in various fields, for example in manufacturing, computer science, reliability theory, management science, just to mention some of them. For a detailed references on the finite-source models and their application the interested reader is recommended to visit the following link

`http://irh.inf.unideb.hu/user/jsztrik/research/fsqreview.pdf`

3.1 The $M/M/r/r/n$ Queue, Engset-Loss System

As we can see depending on the system capacity r in an $M/M/r/r/n$ a customer may find the system full. Despite of the infinite-source model where the customer is lost, in the finite-source model this request returns to the source and stay there for a exponentially distributed time. Since all the random variables are supposed to be exponentially distributed the number of customers in the system is a birth-death process with the following rates

$$\begin{aligned} \lambda_k &= (n-k)\lambda \quad , \quad 0 \le k < r, \\ \mu_k &= k\mu \quad , \quad 1 \le k \le r, \end{aligned}$$

T.O Engset, 1865-1943

hence the distribution can be obtained as

$$P_k = \binom{n}{k}\rho^k P_0, \qquad 0 \le k \le r,$$

$$P_k = \frac{\binom{n}{k}\rho^k}{\sum_{i=0}^{r}\binom{n}{i}\rho^i},$$

which is called a **truncated binomial or Engset distribution** .
This is the distribution of a **finite-source loss or Engset system** .
Specially, if $r = n$ that is no loss and each customer has its own server the distribution has a very nice form, namely

$$P_k = \frac{\binom{n}{k}\rho^k}{\sum_{i=0}^{n}\binom{n}{i}\rho^i} = \frac{\binom{n}{k}\rho^k}{(1+\rho)^n}$$
$$= \binom{n}{k}\left(\frac{\rho}{1+\rho}\right)^k\left(1-\frac{\rho}{1+\rho}\right)^{n-k},$$

that is we have a binomial distribution with success parameter $p = \dfrac{\rho}{1+\rho}$.
That is p is the probability that a given request is in the system. It is easy to see that this distribution remains valid even for a $G/G/n/n/n$ system since

$$p = \frac{\mathbb{E}(S)}{\mathbb{E}(S)+\mathbb{E}(\tau)} = \frac{\rho}{1+\rho},$$

where $\rho = \frac{\mathbb{E}(S)}{\mathbb{E}(\tau)}$, and $\mathbb{E}(\tau)$ denotes the mean time a customer spends in the source.

As before it is easy to see that the *performance measures* are as follows

- *Mean number of customers in the system* $\overline{N}$

$$\overline{N} = \sum_{k=0}^{r} kP_k, \quad \overline{r} = \overline{N}, \quad U_S = \frac{\overline{r}}{r} = \frac{\overline{N}}{r},$$

- *Mean number of customers in the source* $\overline{m}$

$$\overline{m} = n - \overline{N}$$

- *Utilization of a source* U_t

$$U_t = \frac{\overline{m}}{n} = \frac{\mathbb{E}(\tau)}{\mathbb{E}(\tau) + \frac{1}{\mu}},$$

thus

$$\mathbb{E}(\tau) = \frac{1}{\mu} \frac{U_t}{1 - U_t}.$$

This help us to calculate the mean number of retrials of a customer from the source to enter to the system. That it we have

$$\mathbb{E}(N_R) = \lambda \mathbb{E}(\tau),$$

hence the mean number of rejection is $E(N_R) - 1$.

The blocking probability, that is the probability that a customer find the system full at his arrival, by the help of the Bayes's theorem can be calculated as

$$P_B(n, r) = \frac{(n - r)P_r(n, r)}{\sum_{i=0}^{r} (n - i)P_k(n, r)} = P_r(n - 1, r).$$

This can easily be verified by

$$\begin{aligned} P_B(n, r) &= \lim_{h \to 0} \frac{((n - r)\lambda h + o(h))P_r(n, r)}{\sum_{i=0}^{r} ((n - i)\lambda h + o(h))P_i(n, r)} = \frac{(n - r)P_r(n, r)}{\sum_{i=0}^{r} (n - i)P_i(n, r)} \\ &= \frac{(n - r)\binom{n}{r}\rho^r}{\sum_{i=0}^{r}(n - i)\binom{n}{i}\rho^i} = \frac{(n - r)\frac{n!}{r!(n-r)!}\rho^r}{\sum_{i=0}^{r}(n - i)\frac{n!}{i!(n-i)!}\rho^i} \\ &= \frac{\frac{(n-1)!}{r!(n-1-r)!}\rho^r}{\sum_{i=0}^{r}\frac{(n-1)!}{i!(n-1-i)!}\rho^i} = \frac{\binom{n-1}{r}\rho^r}{\sum_{i=0}^{r}\binom{n-1}{i}\rho^i} = P_r(n - 1, r). \end{aligned}$$

Let $E(n,r,\rho)$ denote the blocking probability, that is $E(n,r,\rho) = P_r(n-1,r)$, which is called **Engset's loss formula**.
In the following we show a recursion for this formula, namely

$$E(n,r,\rho) = \frac{\binom{n-1}{r}\rho^r}{\sum_{i=0}^{r}\binom{n-1}{i}\rho^i} = \frac{\binom{n-1}{r-1}\frac{n-r}{r}\rho^r}{\sum_{i=0}^{r-1}\binom{n-1}{i}\rho^i + \binom{n-1}{r-1}\frac{n-r}{r}\rho^r}$$
$$= \frac{\frac{n-r}{r}\rho E(n,r-1,\rho)}{1+\frac{n-r}{r}\rho E(n,r-1,\rho)} = \frac{(n-r)\rho E(n,r-1,\rho)}{r+(n-r)\rho E(n,r-1,\rho)}.$$

The initial value is

$$E(n,1,\rho) = P_1(n-1,1) = \frac{(n-1)\rho}{1+(n-1)\rho}.$$

It is clear that

$$\lim_{n\to\infty,\quad \lambda\to 0,\quad n\lambda\to\lambda'} E(n,r,\rho) = B(r,\rho'),$$

where

$$\rho' = \frac{\lambda'}{\mu}$$

which van be seen formally, too. Moreover, as $(n-r)\rho \to \rho'$ the well-known recursion for $B(r,\rho')$ is obtained which also justifies the correctness of the recursion for $E(n,r,\rho)$.

In particular, if $r = n$ then it is easy to see that $\overline{N} = n\frac{\rho}{1+\rho}$ and thus

$$U_S = \frac{\rho}{1+\rho}, \quad \overline{m} = \frac{n}{1+\rho}, \quad U_t = \frac{1}{1+\rho}, \quad \mathbb{E}(\tau) = \frac{1}{\lambda}, \quad \mathbb{E}(N_R) = 1, \quad P_B = 0,$$

which was expected.

In general case

$$\overline{\mu} = \overline{r}\mu = \overline{\lambda} = \sum_{k=0}^{r-1}\lambda(n-k)P_k \neq \lambda(n-\overline{N}), \quad \overline{T} = \frac{1}{\mu}.$$

Let us consider the distribution of the system at the instant when an arriving customer enters into the system.
By using the Bayes's law we have

$$\Pi_k = \lim_{h\to 0} \frac{(\lambda_k h + o(h))P_k}{\sum_{i=0}^{r-1}(\lambda_i h + o(h))P_i} = \frac{\lambda_k P_k}{\sum_{i=0}^{r-1}\lambda_i P_i}, \qquad k = 0, \cdots, r-1$$

$$\overline{T} = \frac{1}{\mu}\sum_{k=0}^{r-1}\frac{\lambda_k P_k}{\sum_{i=0}^{r-1}\lambda_i P_i} = \frac{1}{\mu}$$

$$\overline{\lambda}\cdot\overline{T} = \mu\overline{r}\cdot\frac{1}{\mu} = \overline{r} = \overline{N}$$

which **Little's formula** for the finite-source loss system.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

3.2 The $M/M/1/n/n$ Queue

It is the traditional machine interference problem, where the broken machines has to wait and the single repairman fixes the failed machine in FIFO order. Assume the the operating times are exponentially distributed with parameter λ and the repair rate is μ. All random variables are supposed to be independent of each other.

Let $N(t)$ denote the number of customers in the system at time t, which is a birth-death process with birth rates

$$\lambda_k = \begin{cases} (n-k)\lambda & \text{, ha } 0 \le k \le n, \\ \\ 0 & \text{, ha } k > n, \end{cases}$$

and with death rate

$$\mu_k = \mu, \qquad k \ge 1.$$

Thus for the distribution we have

$$P_k = \frac{n!}{(n-k)!}\varrho^k P_0 = (n-k+1)\varrho P_{k-1},$$

where

$$\varrho = \frac{\lambda}{\mu},$$

and

$$P_0 = \frac{1}{1 + \sum_{k=1}^{n} \frac{n!}{(n-k)!} \varrho^k} = \frac{1}{\sum_{k=0}^{n} \frac{n!}{(n-k)!} \varrho^k}.$$

Since the state space is finite the steady-state distribution always exists but if $\varrho > 1$ then more repairmen is needed.

For numerical calculation other forms are preferred that is why we introduce some notations.
Let $P(k;\lambda)$ a λ denote the Poisson distribution with parameter λ) and let $Q(k;\lambda)$ denote its cumulative distribution function, that is

$$P(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad 0 \le k < \infty;$$

$$Q(k;\lambda) = \sum_{i=0}^{k} P(i;\lambda), \qquad 0 \le k < \infty.$$

First we show that

$$P_k = \frac{P(n-k;R)}{Q(n;R)}, \qquad 0 \le k \le n,$$

where

$$R = \frac{\mu}{\lambda} = \varrho^{-1}.$$

By elementary calculations we have

$$\frac{P(n-k;R)}{Q(n;R)} = \frac{\frac{n!}{(n-k)!} \left(\frac{\mu}{\lambda}\right)^{n-k} e^{-\frac{\mu}{\lambda}}}{\sum_{i=0}^{n} \frac{n!}{i!} \left(\frac{\mu}{\lambda}\right)^{i} e^{-\frac{\mu}{\lambda}}} = \frac{\frac{n!}{(n-k)!} \left(\frac{\lambda}{\mu}\right)^{k}}{\sum_{i=0}^{n} \frac{n!}{i!} \left(\frac{\lambda}{\mu}\right)^{n-i}} = \frac{\frac{n!}{(n-k)!} \left(\frac{\lambda}{\mu}\right)^{k}}{\sum_{k=0}^{n} \frac{n!}{(n-i)!} \left(\frac{\lambda}{\mu}\right)^{i}} = P_k.$$

Hence a very important consequence is

$$P_0 = B(n,R).$$

The *main performance measures* can be obtained as follows

- *Utilization of the server* and the *throughput of the system*
  For the utilization of the server we have

  $$U_s = 1 - P_0 = 1 - B(n,R).$$

  By using the cumulative distribution function this cab be written as

  $$U_s = \frac{Q(n-1;R)}{Q(n;R)}.$$

  For the throughput of the system we obtain

  $$\lambda_t = \mu U_s.$$

- *Mean number of customers in the system* $\overline{N}$ can be calculated as

$$\overline{N} = \sum_{k=0}^{n} kP_k = n - \sum_{k=0}^{n}(n-k)P_k =$$

$$= n - \frac{1}{\varrho}\sum_{k=0}^{n}(n-k)\varrho P_k = n - \frac{1}{\varrho}\sum_{k=0}^{n-1} P_{k+1} =$$

$$= n - \frac{1}{\varrho}(1-P_0) = n - \frac{U_s}{\varrho}.$$

  In other form

$$\overline{N} = n - \frac{RQ(n-1;R)}{Q(n;R)} = n - \frac{U_s}{\varrho}.$$

- *Mean queue length, mean number of customers waiting* can be derived as

$$\overline{Q} = \sum_{k=1}^{n}(k-1)P_k = \sum_{k=1}^{n} kP_k - \sum_{k=1}^{n} P_k = n - \frac{\mu}{\lambda}(1-P_0) - (1-P_0) =$$

$$= n - (1-P_0)(1+\frac{\mu}{\lambda}) = n - \left(1+\frac{1}{\varrho}\right)U_s.$$

- *Mean number of customers in the source* can be calculated as

$$\overline{m} = \sum_{k=0}^{n}(n-k)P_k = n - \overline{N} = \frac{\mu}{\lambda}(1-P_0) = \frac{U_s}{\varrho}.$$

- *Mean busy period of the server*
  Since

$$U_s = 1 - P_0 = \frac{E\delta}{\frac{1}{n\lambda} + E\delta},$$

  thus

$$E\delta = \frac{1-P_0}{n\lambda P_0} = \frac{U_s}{n\lambda(1-U_s)}.$$

  In computer science and reliability theory application we often need the following measure

- *Utilization of a given source (machine, terminal)*

  The utilization of the ith source is defined by

$$U^{(i)} = \lim_{T\to\infty}\frac{1}{T}\int_0^T \chi(\text{at time } t \text{ the } i\text{th source is active})dt$$

  Then

$$U^{(i)} = P(\text{ there is a request in the } i\text{th source}) .$$

Hence the *overall utilization of the sources* is

$$U_n = \sum_{k=0}^{n}(n-k)P_k = \overline{m} = \frac{\mu}{\lambda}(1-P_0).$$

Thus the utilization of any source is

$$U_t = \frac{\mu}{n\lambda}(1-P_0) = \frac{\overline{m}}{n}.$$

This can be obtained in the following way as well,

$$U^{(i)} = \sum_{k=1}^{n}\frac{n-k}{n}P_k = \frac{\overline{m}}{n},$$

since the source are homogeneous we have

$$U_t = U^{(i)} \ .$$

- *Mean waiting time*

  By using the result of Tomkó 1 we have

$$U_t = \frac{1/\lambda}{1/\lambda + \overline{W} + 1/\mu} = \frac{\overline{m}}{n}.$$

  Thus

$$\lambda\overline{m} = \frac{n}{1/\lambda + \overline{W} + 1/\mu},$$

  and

$$\lambda\overline{m}\overline{W} = n - \overline{m}\left(1 + \frac{\lambda}{\mu}\right) = n - \frac{U_s}{\varrho}(1+\varrho) = \overline{Q},$$

  which the **Litle's law** for the mean waiting time. Hence

$$\overline{W} = \frac{\overline{Q}}{\lambda\overline{m}} = \frac{1}{\mu}\left(\frac{n}{U_s} - \frac{1+\varrho}{\varrho}\right).$$

  The mean response can be obtained as

$$\overline{T} = \overline{W} + \frac{1}{\mu} = \frac{1}{\mu}\left(\frac{n}{1-P_0} - \frac{1}{\varrho}\right) = \frac{1}{\mu}\left(\frac{n}{U_s} - \frac{1}{\varrho}\right).$$

  It is easy to prove that

$$\overline{m}\lambda\overline{T} = \overline{N},$$

  which is the **Little's law** for the mean response time. Clearly we have

$$\overline{m}\lambda\left(\overline{W} + \frac{1}{\mu}\right) = \overline{Q} + \overline{m}\varrho =$$

$$= n - \frac{U_s}{\varrho}(1+\varrho) + U_s = n - \frac{U_s}{\varrho} = \overline{N} \ .$$

- *Further relations*

$$U_s = 1 - P_0 = n\varrho U_t = \overline{m}\varrho,$$

and thus

$$\overline{m}\lambda = \mu U_s = \lambda_t \ .$$

It should be noted that the utilization of the server plays a key role in the calculation of all the main performance measures.

Distribution at the arrival instants

In the following we find the steady-state distribution of the system at arrival instants and in contrary to the infinity-source model is not he same as the distribution at a random point. To show this use the Bayes's theorem, that is

$$\Pi_k(n) = \lim_{h\to 0} \frac{(\lambda_k h + o(h))P_k}{\sum_{j=0}^{n-1}(\lambda_j h + o(h))P_j} = \frac{\lambda_k P_k}{\sum_{j=0}^{n-1}\lambda_j P_j} = \frac{\frac{n(n-1)\cdots(n-k)\lambda^k}{\mu_1\cdots\mu_k}P_0}{\sum_{j=0}^{n-1}\frac{n(n-1)\cdots(n-j)\lambda^j}{\mu_1\cdots\mu_j}P_0}$$
$$= \frac{\frac{(n-1)\cdots(n-k)\lambda^k}{\mu_1\cdots\mu_k}}{1+\sum_{j=1}^{n-1}\frac{(n-1)\cdots(n-j)\lambda^j}{\mu_1\cdots\mu_j}} = \frac{\frac{(n-1)\cdots(n-1-k+1)\lambda^k}{\mu_1\cdots\mu_k}}{1+\sum_{i=1}^{n-1}\frac{(n-1)\cdots(n-1-i+1)\lambda^i}{\mu_1\cdots\mu_i}} = P_k(n-1)$$

irrespective to the number of servers. It should be noted that this relation shows a very important result, namely that at arrivals the distribution of the system containing n sources is not he same as its distribution at random points, but equals to the random point distribution of a system with $n-1$ sources.

Distribution at the departure instants

We are interested in the distribution of the number of customers a departing customer leaves behind in the system. This calculations are independent of the number of servers. By applying the Bayes's theorem we have

$$D_k(n) = \lim_{h\to 0} \frac{(\mu_{k+1} h + o(h))P_{k+1}}{\sum_{j=1}^{n}(\mu_j h + o(h))P_j} = \frac{\mu_{k+1}P_{k+1}}{\sum_{j=1}^{n}\mu_j P_j} = \frac{\frac{\mu_{k+1}n(n-1)\cdots(n-k)\lambda^{k+1}}{\mu_1\cdots\mu_{k+1}}P_0}{\sum_{j=1}^{n}\frac{\mu_j n(n-1)\cdots(n-j+1)\lambda^j}{\mu_1\cdots\mu_j}P_0}$$
$$= \frac{\frac{(n-1)\cdots(n-k)\lambda^k}{\mu_1\cdots\mu_k}}{1+\sum_{j=2}^{n}\frac{(n-1)\cdots(n-j+1)\lambda^{j-1}}{\mu_1\cdots\mu_{j-1}}} = \frac{\frac{(n-1)\cdots(n-1-k+1)\lambda^k}{\mu_1\cdots\mu_k}}{1+\sum_{i=1}^{n-1}\frac{(n-1)\cdots(n-1-i+1)\lambda^i}{\mu_1\cdots\mu_i}} = P_k(n-1)$$

in the case when there is customer left in the system

$$D_0(n) = \frac{1}{1+\sum_{i=1}^{n-1}\frac{(n-1)\cdots(n-1-i+1)\lambda^i}{\mu_1\cdots\mu_i}} = P_0(n-1)$$

if the system becomes empty.

Recursive Relations

Similarly to the previous arguments it is easy to see that the density function of the response time can be obtained as

$$f_T(x) = \sum_{k=0}^{n-1} f_T(x|k)\Pi_k(n) = \sum_{k=0}^{n-1} \frac{\mu(\mu x)^k}{k!} e^{-\mu x} P_k(n-1).$$

Hence the mean value is

$$\overline{T}(n) = \sum_{k=0}^{n-1} \frac{k+1}{\mu} P_k(n-1) = \frac{1}{\mu}(\overline{N}(n-1)+1).$$

Similarly, for the waiting time we have

$$f_W(x) = \sum_{k=0}^{n-1} f_W(x|k)\Pi_k(n) = \sum_{k=0}^{n-1} \frac{\mu(\mu x)^{k-1}}{(k-1)!} e^{-\mu x} P_k(n-1),$$

thus its mean is

$$\overline{W}(n) = \sum_{k=0}^{n-1} \frac{k}{\mu} P_k(n-1) = \frac{1}{\mu}(\overline{N}(n-1)),$$

which is clear.
We want to verify the correctness of the formula

$$\overline{T}(n) = \frac{1}{\mu}(\overline{N}(n-1)+1).$$

As we have shown earlier the utilization can be expressed by the Erlang's loss formula, hence

$$\overline{N}(n) = n - \frac{1 - B(n, \frac{1}{\varrho})}{\varrho}.$$

Using the well-known recursive relation we have

$$B(n, \frac{1}{\varrho}) = \frac{\frac{1}{\varrho} B(n-1, \frac{1}{\varrho})}{n + \frac{1}{\varrho} B(n-1, \frac{1}{\varrho})} = \frac{B(n-1, \frac{1}{\varrho})}{n\varrho + B(n-1, \frac{1}{\varrho.})}.$$

Since

$$\overline{N}(n-1) = n - 1 - \frac{1 - B(n-1, \frac{1}{\varrho})}{\varrho}$$

thus

$$\varrho\overline{N}(n-1) = (n-1)\varrho - 1 - B\left(n-1, \frac{1}{\varrho}\right)$$

$$B\left(n-1, \frac{1}{\varrho}\right) = 1 + \varrho\overline{N}(n-1) - (n-1)\varrho.$$

After substitution we have

$$B\left(n,\frac{1}{\varrho}\right)=\frac{\frac{1}{\varrho}(1+\varrho\overline{N}(n-1)-(n-1)\varrho)}{n+\frac{1}{\varrho}(1+\varrho\overline{N}(n-1)-(n-1)\varrho)}$$
$$=\frac{1+\varrho\overline{N}(n-1)-(n-1)\varrho}{n\varrho+1+\varrho\overline{N}(n-1)-(n-1)\varrho}=\frac{1+\varrho\overline{N}(n-1)-(n-1)\varrho}{1+\varrho\overline{N}(n-1)+\varrho}.$$

Therefore

$$\overline{N}(n)=\frac{n\varrho-1+B(n,\frac{1}{\varrho})}{\varrho}=\frac{n\varrho-1+\frac{1+\varrho\overline{N}(n-1)-(n-1)\varrho}{1+\varrho\overline{N}(n-1)+\varrho}}{\varrho}$$
$$=\frac{n\varrho-\frac{n\varrho}{1+\varrho\overline{N}(n-1)+\varrho}}{\varrho}=n-\frac{n}{1+\varrho\overline{N}(n-1)+\varrho.}$$

Fnally

$$n-\overline{N}(n)=\frac{n}{1+\varrho\overline{N}(n-1)+\varrho}$$

$$1+\varrho(\overline{N}(n-1)+1)=\frac{n}{n-\overline{N}(n)}$$
$$\varrho(\overline{N}(n-1)+1)=\frac{\overline{N}(n)}{n-\overline{N}(n)},$$

which is a recursion for the mean number of customers in the system.

Now we able to prove our relation regarding the mean response time. Keeping in mind the recursive relation for $\overline{N}(n-1)$ we get

$$\overline{T}(n)=\frac{1}{\mu}(\overline{N}(n-1)+1)$$
$$\lambda\overline{T}(n)=\varrho(\overline{N}(n-1)+1)=\frac{\overline{N}(n)}{n-\overline{N}(n)}$$
$$\lambda(n-\overline{N}(n))\overline{T}(n)=\overline{N}(n),$$

which was proved earlier.

Now let us show how we can verify $\overline{T}(n)$ directly. It can easily be seen that

$$U_S(n)=1-B(n,\frac{1}{\varrho})=1-\frac{\frac{1}{\varrho}B(n-1,)}{n+\frac{1}{\varrho}B(n-1,\frac{1}{\varrho})}$$

$$=\frac{n}{n+\frac{1}{\varrho}B(n-1,\frac{1}{\varrho})}=\frac{n\varrho}{n\varrho+B(n-1,\frac{1}{\varrho})}=\frac{n\varrho}{n\varrho+1-U_S(n-1)},$$

that is there is a recursion for the utilization as well. It is also very important because by using this recursion all the main performance measures can be obtained. Thus if λ,μ,n are

given we can use the recursion for $U_S(n)$ and finally substitute it into the corresponding formula. Thus

$$U_S(n-1) = n\varrho + 1 - \frac{n\varrho}{U_S(n)} = 1 + n\varrho\left(1 - \frac{1}{U_S(n)}\right).$$

Since

$$\overline{N}(n-1) = n - 1 - \frac{U_S(n-1)}{\varrho},$$

we proceed

$$\overline{T}(n) = \frac{1}{\mu}(\overline{N}(n-1)+1) = \frac{1}{\mu}\left(n - 1 - \frac{U_S(n-1)}{\varrho} + 1\right)$$
$$= \frac{1}{\mu}\left(n - \frac{U_S(n-1)}{\varrho}\right) = \frac{1}{\mu}\left(n - \frac{1 + n\varrho(1 - \frac{1}{U_S(n)})}{\varrho}\right) = \frac{1}{\mu}\left(\frac{n}{U_S(n)} - \frac{1}{\varrho}\right),$$

which shows the correctness of the formula.

In the following let us show to compute $\overline{T}(n), \overline{W}(n), \overline{N}(n)$ recursively. As we have seen

$$\overline{T}(n) = \frac{1}{\mu}(\overline{N}(n-1)+1)$$
$$\overline{W}(n) = \overline{T}(n) - \frac{1}{\mu} = \frac{1}{\mu}\overline{N}(n-1),$$

we have to know how $\overline{N}(n)$ can be expressed in term of $\overline{T}(n)$.
It can be shown very easily, namely

$$\overline{N}(n) = \lambda(n - \overline{N}(n))\overline{T}(n) = \lambda n\overline{T}(n) - \lambda\overline{N}(n)\overline{T}(n)$$
$$\overline{N}(n)(1 + \lambda\overline{T}(n)) = \lambda n\overline{T}(n)$$
$$\overline{N}(n) = \frac{\lambda n\overline{T}(n)}{1 + \lambda\overline{T}(n)}.$$

The initial values are

$$\overline{T}(1) = \frac{1}{\mu}$$
$$\overline{N}(1) = \frac{\varrho}{1+\varrho}.$$

Now the iteration proceeds as

$$\overline{W}(n) = \frac{1}{\mu}\overline{N}(n-1)$$
$$\overline{T}(n) = \frac{1}{\mu} + \overline{W}(n)$$
$$\overline{N}(n) = \frac{\lambda n\overline{T}(n)}{(1 + \lambda\overline{T})(n)}$$

that is we use a double iteration. The main advantage is that only the mean values are needed. This method is referred to as **mean value analysis.**

In the previous section we have derived a recursion for $U_S(n)$ and thus we may expect that there is direct recursive relation for the other mean values as well since they depends on the utilization. As a next step we find a recursion for the *mean number of customers in the source* $\overline{m}(n)$. It si quite easy since

$$\overline{m}(n) = \frac{U_s(n)}{\rho} = \frac{n}{n\rho + 1 - U_s(n-1)} =$$

$$= \frac{\frac{n}{\rho}}{n + \frac{1}{\rho} - \overline{m}(n-1)} = \frac{n}{n\rho + 1 - \rho\overline{m}(n-1).}$$

By using this relation for the *utilization of the source* can be expressed as

$$U_t(n) = \frac{\overline{m}(n)}{n} = \frac{n}{n\rho + 1 - \rho\overline{m}(n-1)}\frac{1}{n} =$$

$$= \frac{\frac{1}{n-1}}{\frac{n\rho+1}{n-1} - \rho U_t(n-1)} = \frac{1}{n\rho + 1 - (n-1)\rho U_t(n-1)}.$$

For the *mean number of customers in the system* we have

$$\overline{N}(n) = n - \frac{U_s(n)}{\rho} = \frac{n\rho - U_s(n)}{\rho} = \frac{n\rho - \frac{n\rho}{n\rho + 1 - U_s(n-1)}}{\rho} =$$

$$= \frac{n^2\rho + n - nU_s(n-1) - 1}{n\rho + 1 - U_s(n-1)} = \frac{n(n\rho - U_s(n-1))}{n\rho + 1 - U_s(n-1)}.$$

Since

$$\overline{N}(n-1) = n - 1 - \frac{U_s(n-1}{\rho} = \frac{n\rho - U_s(n-1)}{\rho} - 1$$

$$\rho(\overline{N}(n-1) + 1) = n\rho - U_s(n-1)$$

$$U_s(n-1) = n\rho - \rho(\overline{N}(n-1) + 1),$$

thus after substitution we get

$$\overline{N}(n) = \frac{n\rho(\overline{N}(n-1) + 1)}{1 + \rho(\overline{N}(n-1) + 1)}.$$

Finally find the recursion for the *mean response time* . Starting with

$$\overline{T}(n) = \frac{1}{\mu}\left(\frac{n}{U_s(n)} - \frac{1}{\rho}\right)$$

using that

$$\overline{T}(n-1) = \frac{1}{\mu}\left(\frac{n-1}{U_s(n-1)} - \frac{1}{\rho}\right)$$

$$\mu\overline{T}(n-1) = \frac{n-1}{U_s(n-1)} - \frac{1}{\rho}$$

$$\mu\overline{T}(n-1) + \frac{1}{\rho} = \frac{n-1}{U_s(n-1)}$$

$$U_s(n-1) = \frac{(n-1)\rho}{\lambda\overline{T}(n-1)+1},$$

substituting into the recursion for $U_s(n)$ we obtain

$$\overline{T}(n) = \frac{1}{\mu}\frac{n\rho - U_s(n-1)}{\rho} = \frac{1}{\mu}\frac{n\lambda\overline{T}(n-1)+1}{\lambda\overline{T}(n-1)+1}.$$

Obviously the missing initial values are

$$\overline{m}(1) = U_t(1) = \frac{1}{1+\varrho}.$$

Distribution Function of the Response Time and Waiting Time

This subsection is devoted to one of the major problems in finite-source queueing systems. To find the distribution function of the response and waiting time is not easy. As it is expected the theorem of total probability should be used.

Let us determine the density function and then the distribution function. As we did many times in earlier chapters the law of total probability should be applied for the conditional density functions and the distribution at the arrival instants. So we can write

$$f_T(n,x) = \sum_{k=0}^{n-1} \mu\frac{(\mu x)^k}{k!}e^{-\mu x}\underbrace{\frac{\frac{(\frac{\mu}{\lambda})^{n-1-k}}{(n-1-k)!}e^{-\frac{\mu}{\lambda}}}{\sum_{i=0}^{n-1}\frac{(\frac{\mu}{\lambda})^i}{i!}e^{-\frac{\mu}{\lambda}}}}_{Q(n-1,\frac{\mu}{\lambda})} = \mu\frac{(\mu x+\frac{\mu}{\lambda})^{n-1}}{(n-1)!}\frac{e^{-(\mu x+\frac{\mu}{\lambda})}}{Q(n-1,\frac{\mu}{\lambda})}$$

$$= \frac{\mu P(n-1,\mu x+\frac{\mu}{\lambda})}{Q(n-1,\frac{\mu}{\lambda})}.$$

Similarly for the waiting time

$$f_W(n,x) = \sum_{k=1}^{n-1}\mu\frac{(\mu x)^{k-1}}{(k-1)!}e^{-\mu x}P_k(n-1) = \frac{\sum_{i=0}^{n-2}\mu\frac{(\mu x)^i}{i!}e^{-\mu x}\frac{(\frac{\mu}{\lambda})^{n-2-i}}{(n-2-i)!}e^{-\frac{\mu}{\lambda}}}{Q(n-1,\frac{\mu}{\lambda})}$$

$$= \frac{\mu P(n-2,\mu x+\frac{\mu}{\lambda})}{Q(n-1,\frac{\mu}{\lambda})}.$$

To get the distribution function we have to calculate the integral

$$F_T(n,x) = \int_0^x f_T(n,t)\,dt.$$

Using the substitution $y = \mu t + \frac{\mu}{\lambda}, \quad t = (y - \frac{\mu}{\lambda})\frac{1}{\mu}, \quad \frac{dt}{dy} = \frac{1}{\mu}$.
Hence

$$F_T(n,x) = \frac{\int_{\frac{\mu}{\lambda}}^{\mu x+\frac{\mu}{\lambda}} \frac{y^{n-1}}{(n-1)!}e^{-y}\,dy}{Q(n-1,\frac{\mu}{\lambda})} = \frac{\left[1-\sum_{i=0}^{n-1}\frac{y^i}{y!}e^{-y}\right]_{\frac{\mu}{\lambda}}^{\mu x+\frac{\mu}{\lambda}}}{Q(n-1,\frac{\mu}{\lambda})} = 1 - \frac{Q(n-1,\mu x+\frac{\mu}{\lambda})}{Q(n-1,\frac{\mu}{\lambda})}.$$

Similarly for the waiting time we have

$$F_W(n,x) = 1 - \frac{Q(n-2,\mu x+\frac{\mu}{\lambda})}{Q(n-1,\frac{\mu}{\lambda})}.$$

Now let us determine the distribution function by the help of the conditional distribution functions. Clearly we have to know the distribution function of the Erlang distributions, thus we can proceed as

$$F_T(x) = \sum_{k=0}^{n-1}\left(1-\sum_{j=0}^{k}\frac{(\mu x)^j}{j!}e^{-\mu x}\right)P_k(n-1)$$

$$= 1 - \sum_{k=0}^{n-1}\left(\sum_{j=0}^{k}\frac{(\mu x)^j}{j!}e^{-\mu x}\right)P_k(n-1)$$

$$= 1 - \sum_{k=0}^{n-1}Q(k,\mu x)P_k(n-1) = 1 - \sum_{k=0}^{n-1}Q(k,\mu x)\frac{\frac{(\frac{\mu}{\lambda})^{n-1-k}}{(n-1-k)!}e^{-\frac{\mu}{\lambda}}}{Q(n-1,\frac{\mu}{\lambda})}$$

$$= 1 - \frac{Q(n-1,\mu x+\frac{\mu}{\lambda})}{Q(n-1,\frac{\mu}{\lambda})}$$

Meantime we have used that

$$\int_0^\lambda \frac{t^j}{j!}e^{-t}\,dt = 1 - \sum_{i=0}^{j}\frac{\lambda^i}{i!}e^{-\lambda}$$

and thus

$$\sum_{j=0}^{l}\frac{\mu^{l-j}}{(l-j)!}e^{-\mu}\sum_{i=0}^{j}\frac{\lambda^i}{i!}e^{-\lambda}$$

can be written as

$$\sum_{j=0}^{l}\left(1-\int_0^{\lambda}\frac{t^j}{j!}e^{-t}\,dt\right)\frac{\mu^{l-j}}{(l-j)!}e^{-\mu}$$
$$=\underbrace{\sum_{j=0}^{l}\frac{\mu^{l-j}}{(l-j)!}e^{-\mu}}_{Q(l,\mu)}-\int_0^{\lambda}\frac{(t+\mu)^l}{l!}e^{-(t+\mu)}=Q(l,\mu)-\int_{\mu}^{\lambda+\mu}\frac{y^l}{l!}e^{-y}\,dy$$
$$=Q(l,\mu)-\left[1-\sum_{i=0}^{l}\frac{y^i}{i!}e^{-y}\right]_{\mu}^{\lambda+\mu}=Q(l,\lambda+\mu).$$

During the calculations we could see that the derivative of $Q(k,t)$ is $-P(k,t)$, which can be used to find the density function, that is

$$f_T(x)=\frac{\mu P(n-1,\mu x+\frac{\mu}{\lambda})}{Q(n-1,\frac{\mu}{\lambda})}.$$

Generating Function of the Customers in the System

Using the definition the generating function $G_N(s)$ can be calculated as

$$\begin{aligned}G_N(s)&=\sum_{k=0}^{n}s^k\frac{\left(\frac{\mu}{\lambda}\right)^{n-k}}{(n-k)!}P_0\\&=s^n\sum_{k=0}^{n}\frac{\left(\frac{1}{s\rho}\right)^{n-k}}{(n-k)!}P_0\\&=s^n e^{-\frac{1}{\rho}\left(1-\frac{1}{s}\right)}\frac{Q\left(n,\frac{1}{\rho s}\right)}{Q\left(n,\frac{1}{\rho}\right)}.\end{aligned}$$

This could be derived in the following way. Let denote by F the number of customers in the source. As we have proved earlier its distribution can be obtained as the distribution of an Erlang loss system with traffic intensity $\frac{1}{\rho}$. Since the generating function of this system has been obtained we can use this fact. Thus

$$\begin{aligned}G_N(s)=E(s^N)=E(s^{n-F})=s^nE(s^{-F})=s^nG_F\left(\frac{1}{s}\right)\\=s^n e^{-\frac{1}{\rho}\left(1-\frac{1}{s}\right)}\frac{Q\left(n,\frac{1}{\rho s}\right)}{Q\left(n,\frac{1}{\rho}\right)}.\end{aligned}$$

To verify the formula let us compute the mean number of customers in the system. By

the property of the generating function we have

$$\overline{N}(n) = G'_{N(n)}(1) = \left(s^n G_{F(n)}\left(\frac{1}{s}\right)\right)'_{s=1}$$
$$G'_{N(n)}(s) = n \cdot s^{n-1} G_{F(n)}\left(\frac{1}{s}\right) + s^n G'_{F(n)}\left(\frac{1}{s}\right)\left(-\frac{1}{s^2}\right),$$

thus

$$\overline{N}(n) = nG_{F(n)}(1) - G'_{F(n)}(1) = n - \frac{1}{\rho}\left(1 - B\left(n, \frac{1}{\rho}\right)\right) = n - \frac{U_S(n)}{\rho}.$$

Laplace-transform of the Response Time and Waiting Time

Solution 1

By the law of the total Laplace-transforms we have

$$L_T(s) = \sum_{k=0}^{n-1} \left(\frac{\mu}{\mu+s}\right)^{k+1} P_k(n-1)$$

since the conditional response time is Erlang distributed with parameters $(k+1, \mu)$. Substituting $P_k(n-1)$ we get

$$\begin{aligned}
L_T(s) &= \sum_{k=0}^{n-1} \left(\frac{\mu}{\mu+s}\right)^{k+1} \frac{\left(\frac{\mu}{\lambda}\right)^{n-1-k}}{(n-1-k)!} \frac{e^{-\frac{\mu}{\lambda}}}{Q\left(n-1, \frac{\mu}{\lambda}\right)} \\
&= \frac{\sum_{k=0}^{n-1} \left(\frac{\mu+s}{\mu}\right)^{-k-1} \cdot \left(\frac{\mu+s}{\mu}\right)^{n} \cdot \frac{\left(\frac{\mu}{\lambda}\right)^{n-1-k}}{(n-1-k)!} e^{-\frac{\mu}{\lambda}}}{\left(\frac{\mu+s}{\mu}\right)^{n} Q\left(n-1, \frac{\mu}{\lambda}\right)} \\
&= \left(\frac{\mu}{\mu+s}\right)^{n} e^{-\frac{\mu}{\lambda}} \frac{\sum_{k=0}^{n-1} \left(\frac{\mu+s}{\mu} \cdot \frac{\mu}{\lambda}\right)^{n-1-k} \cdot \frac{1}{(n-1-k)!}}{Q\left(n-1, \frac{\mu}{\lambda}\right)} \\
&= \left(\frac{\mu}{\mu+s}\right)^{n} e^{-\frac{\mu}{\lambda}} \frac{Q\left(n-1, \frac{\mu+s}{\lambda}\right) \cdot e^{\frac{\mu+s}{\lambda}}}{Q\left(n-1, \frac{\mu}{\lambda}\right)} \\
&= \left(\frac{\mu}{\mu+s}\right)^{n} e^{\frac{s}{\lambda}} \frac{Q\left(n-1, \frac{\mu+s}{\lambda}\right)}{Q\left(n-1, \frac{\mu}{\lambda}\right)}.
\end{aligned}$$

Solution 2

Let us calculate $L_T(s)$ by the help of the density function. Since the denominator is a constant we have to determine the Laplace-transform of the numerator, that is

$$\begin{aligned} L_{Num}(s) &= \int_0^\infty \mu \frac{\left(\mu x + \frac{\mu}{\lambda}\right)^{n-1}}{(n-1)!} e^{-\left(\mu x + \frac{\mu}{\lambda}\right)} \cdot e^{-sx} \mathrm{d}x \\ &= e^{-\frac{\mu}{\lambda}} \int_0^\infty \mu \frac{\left(\mu x + \frac{\mu}{\lambda}\right)^{n-1}}{(n-1)!} e^{-(\mu+s)x} \mathrm{d}x. \end{aligned}$$

By using the binomial theorem we get

$$\begin{aligned} L_{sz}(s) &= \frac{e^{-\frac{\mu}{\lambda}}}{(n-1)!} \int_0^\infty \mu \sum_{k=0}^{n-1} \binom{n-1}{k} (\mu x)^k \left(\frac{\mu}{\lambda}\right)^{n-1-k} \cdot e^{-(\mu+s)x} \mathrm{d}x \\ &= e^{-\frac{\mu}{\lambda}} \sum_{k=0}^{n-1} \frac{\left(\frac{\mu}{\lambda}\right)^{n-1-k}}{(n-1-k)!} \int_0^\infty \mu \frac{(\mu x)^k}{k!} e^{-(\mu+s)x} \mathrm{d}x \\ &= e^{-\frac{\mu}{\lambda}} \sum_{k=0}^{n-1} \frac{\left(\frac{\mu}{\lambda}\right)^{n-1-k}}{(n-1-k)!} \left(\frac{\mu}{\mu+s}\right)^{k+1} \\ &= e^{-\frac{\mu}{\lambda}} \left(\frac{\mu}{\mu+s}\right)^n \sum_{k=0}^{n-1} \frac{\left(\frac{\mu}{\lambda} \cdot \frac{\mu+s}{\mu}\right)^{n-1-k}}{(n-1-k)!} \\ &= e^{-\frac{\mu}{\lambda}} \left(\frac{\mu}{\mu+s}\right)^n Q\left(n-1, \frac{\mu+s}{\lambda}\right) \cdot e^{\frac{\mu+s}{\lambda}} \\ &= \left(\frac{\mu}{\mu+s}\right)^n Q\left(n-1, \frac{\mu+s}{\lambda}\right) \cdot e^{\frac{s}{\lambda}}. \end{aligned}$$

Since

$$L_T(s) = \frac{L_{Num}(s)}{Q\left(n-1, \frac{\mu}{\lambda}\right)},$$

thus

$$L_T(s) = \left(\frac{\mu}{\mu+s}\right)^n e^{\frac{s}{\lambda}} \frac{Q\left(n-1, \frac{\mu+s}{\lambda}\right)}{Q\left(n-1, \frac{\mu}{\lambda}\right)}.$$

Solution 3
The Laplace-transform of the numerator be can obtained as

$$L_{Num}(s) = \int_0^\infty \mu \frac{\left(\mu x + \frac{\mu}{\lambda}\right)^{n-1}}{(n-1)!} e^{-\left(\mu x + \frac{\mu}{\lambda}\right)} \cdot e^{-sx} \mathrm{d}x$$

Substituting $t = \mu x + \frac{\mu}{\lambda}$ we get

$$x = \frac{1}{\mu}\left(t - \frac{\mu}{\lambda}\right), \qquad \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{1}{\mu},$$

and thus

$$\begin{aligned} L_{Num}(s) &= \int_{\frac{\mu}{\lambda}}^\infty \mu \frac{t^{n-1}}{(n-1)!} e^{-t} e^{-\frac{s}{\mu}\left(t - \frac{\mu}{\lambda}\right)} \frac{1}{\mu} \mathrm{d}t \\ &= e^{\frac{s}{\mu}} \int_{\frac{\mu}{\lambda}}^\infty \frac{t^{n-1}}{(n-1)!} e^{-\left(1+\frac{s}{\mu}\right)t} \mathrm{d}t \\ &= e^{\frac{s}{\mu}} \left(\frac{\mu}{\mu+s}\right)^{n-1} \int_{\frac{\mu}{\lambda}}^\infty \frac{\left(\left(1+\frac{s}{\mu}\right)t\right)^{n-1}}{(n-1)!} e^{-\left(1+\frac{s}{\mu}\right)t} \mathrm{d}t. \end{aligned}$$

Substituting again $y = \frac{\mu+s}{\mu} t$ $\frac{\mathrm{d}t}{\mathrm{d}y} = \frac{\mu}{\mu+s}$, thus

$$\begin{aligned} L_{Num}(s) &= e^{\frac{s}{\lambda}} \left(\frac{\mu}{\mu+s}\right)^n \int_{\frac{\mu+s}{\lambda}}^\infty \frac{y^{n-1}}{(n-1)!} e^{-y} \mathrm{d}y \\ &= e^{\frac{s}{\lambda}} \left(\frac{\mu}{\mu+s}\right)^n \cdot Q\left(n-1, \frac{\mu+s}{\lambda}\right), \end{aligned}$$

therefore

$$L_T(s) = \frac{L_{Num}(s)}{Q\left(n-1, \frac{\mu}{\lambda}\right)} = e^{\frac{s}{\lambda}} \left(\frac{\mu}{\mu+s}\right)^n \cdot \frac{Q\left(n-1, \frac{\mu+s}{\lambda}\right)}{Q\left(n-1, \frac{\mu}{\lambda}\right)}.$$

That is all 3 solutions gives the same result. Thus, in principle the higher moments of the response time can be evaluated.

Since $L_T(s) = L_W(s) \cdot \frac{\mu}{\mu+s}$, thus

$$L_W(s) = \left(\frac{\mu}{\mu+s}\right)^{n-1} e^{\frac{s}{\lambda}} \frac{Q\left(n-1, \frac{\mu+s}{\lambda}\right)}{Q\left(n-1, \frac{\mu}{\lambda}\right)}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Example 26 *Consider* 6 *machines with mean lifetime of* 40 *hours. Let their mean repair time be* 4 *hours. Find the performance measures.*

Solution: $\lambda = \frac{1}{40}$ per hour, $\mu = \frac{1}{4}$per hour, $\rho = \frac{\lambda}{\mu} = \frac{4}{40} = 0.1$, $n = 6$, $P_0 = 0.484$

Failed machines	0	1	2	3	4	5	6
Waiting machines	0	0	1	2	3	4	5
P_n	$0,484$	0.290	0.145	0.058	0.017	0.003	0.000

$$\overline{Q} = 0.324, \quad U_{sz} = 0.516, \quad \overline{W} = 2.51 \text{ hour}, \quad \overline{T} = 2.51 + 0.25 = 6.51 \text{ hour}$$

$$\overline{e} = 40 \text{ hour}, \quad U_g = 0.86$$

$$\overline{m} = n \times U_g = 5.16, \quad \overline{N} = 6 - 5.16 = 0.84$$

$$E\delta = \frac{0.516}{6 \times \frac{1}{40} \times 0.484} = \frac{4 \times 5.16}{6 \times 0.484} \approx 7.10 \text{ hour}$$

■

Example 27 *Change the mean lifetime to* 2 *hours in the previous Example. Find the performance measures.*

Solution: $\frac{1}{\lambda} = 2$, $\frac{1}{\mu} = 4$, $\frac{\lambda}{\mu} = 2$, $n = 6$, $P_0 = \frac{1}{75973}$, which shows that a single repairman is not enough. We should increase the number of repairmen.

Failed machines	0	1	2	3	4	5	6
Waiting machines	0	0	1	2	3	4	5
P_k	$\frac{1}{75973}$	$\frac{1}{75973}$	0.001	0.012	0.075	0.303	0.606

$$U_s \approx 0.999, \quad \overline{Q} \approx 4.5, \quad \overline{W} \approx 22.5\text{hours}, \quad \overline{T} = 26.5 \text{ hours}$$

$$\overline{e} = 2 \text{ hours}, \quad U_g \approx 0.08, \quad \overline{m} \approx 0.5, \quad \overline{N} \approx 5.5, \quad E\delta \approx \infty.$$

All these measures demonstrate what we have expected because 1 is greater than 1. To decide how many repairmen is needed there are different criterias as we shall see in Section 3.4. To avoid this congestion we must ensure the condition $\frac{\lambda}{r\mu} < 1$ where r is the number of repairmen.

■

3.3 The Heterogeneous $\vec{M}/\vec{M}/1/n/n$ Queue

The results of this section have been published in the paper of Csige and Tomkó [24]. The reason of its introduction is to show the importance of the service discipline.

Let us consider n heterogeneous machines with exponentially distributed operating and repair time with parameter $\lambda_k > 0$ and $\mu_k > 0$, respectively for the kth machine, $k = 1, \cdots, n$. The failures are repaired by a single repairman according to Processor Sharing, FIFO, and Preemptive Priory disciplines. All involved random variables are supposed to be independent of each other.
Let $N(t)$, denote the number of failed machines at time t. Due to the heterogeneity of the machines this information is not enough to describe the behavior of the system because we have to know which machine is under service. Thus let us introduce an $N(t)$-dimensional vector with components $\left(x_1(t), \ldots, x_{v(t)}(t)\right)$ indicating the indexes of the failed machines. Hence for $N(t) > 0$ using FIFO discipline machine with index $x_1(t)$ is under service. Under Processor Sharing discipline when all machines are serviced by a proportional service rate, that is if $N(t) = k$ then the proportion is $1/k$ the order of indexes $(x_1(t), \ldots, x_n(t))$ is not important, but a logical treatment we order them as $\left(x_1(t) < x_2(t) < \ldots < x_{v(t)}(t)\right)$. In the case of Preemptive Priority assuming that the smaller index means higher priority we use the same ordering as before mentioning that in this case the machine with the first index is under service since he has the highest priority among the failed machines.
Due to the exponential distributions the process

$$X(t)(t) = \left(v(t); x_1(t), \ldots, x_{v(t)}(t)\right), \qquad (t \geq 0),$$

is a continuous-time Markov where the ordering of $x_1(t), \ldots, x_{v(t)}(t)$ depends ot the service discipline.

Let us consider the **Processor Sharing** service discipline.
Since $X(t)(t)$ is a finite state Markov chain thus if the parameters λ_k, μ_k, $(1 \leq k \leq n)$ are all positive then it is ergodic and hence the steady-state distribution exists. Of course this heavily depends on the service discipline.
Let the distribution of the Markov chain be denoted by

$$P_0(t), \quad , P'_{i_1,\ldots,i_k}(t).$$

It is not difficult to see that for this distribition we have

$$P'_0(t) = -\left[\sum_{i=1}^{n} \lambda_i\right] P_0(t) + \sum_{i=1}^{n} \mu_i P_i(t),$$

$$P'_{i_1,\ldots,i_k}(t) = \sum_{r=1}^{k} \lambda_{i_r} P_{i_1,\ldots,i_{r-1},i_{r+1},\ldots,i_k}(t) -$$

$$- \left[\nu_{i_1\ldots i_k} + \frac{1}{k}\sum_{r=1}^{k} \mu_{i_r}\right] P_{i_1,\ldots,i_k}(t) + \sum_{r \neq i_1\ldots i_k} \frac{\mu_r}{k+1} P_{i'_1 i'_2 \ldots i'_{k+1}}(t)$$

where $i_1^{'}, \ldots, i_{k+1}^{'}$ is the ordering of the indexes $i_1, \ldots, i_k, r$ and

$$\nu_{i_1 \ldots i_k} = \sum_{r \neq i_1 \ldots i_k} \lambda_r, \quad k = 1, \ldots, n-1.$$

$$P^{'}_{1,\ldots,n}(t) = \sum_{r=1}^{n} \lambda_r P_{1,\ldots,r-1,r+1,]\ldots,n}(t) - \left[\frac{1}{n}\sum_{r=1}^{n} \mu_r\right] P_{1,\ldots,n}(t).$$

The steady-state distribution which is denoted by

$$P_0 = \lim_{t\to\infty} P_o\,(t)\,,$$

$$P_{i_1 \ldots i_k} = \lim_{t\to\infty} P_{i_1 \ldots i_k}\,(t)$$

$$(1 \leq i_1 < i_2 < \ldots < i_k \leq n, \quad 1 \leq k \leq n).$$

is the solution of the following set of equations

$$\left[\sum_{i=1}^{n} \lambda_i\right] P_0 = \sum_{i=1}^{n} \mu_i P_i,$$

$$\left[\nu_{i_1 \ldots i_k} + \frac{1}{k}\sum_{r=1}^{k} \mu_{i_r}\right] P_{i_1 \ldots i_k} = \sum_{r=1}^{k} \lambda_{i_r} P_{i_1 \ldots i_{r-1} i_{r+1} \ldots i_k} +$$

$$+ \sum_{r \neq i_1 \ldots i_k} \frac{\mu_r}{k+1} P_{i_1^{'} i_2^{'} \ldots i_{k+1}^{'}},$$

$$\left[\frac{1}{n}\sum_{r=1}^{n} \mu_r\right] P_{1,\ldots,n} = \sum_{r=1}^{n} \lambda_r P_{1\ldots,r-1,r+1,\ldots,n}$$

with normalizing condition

$$P_0 + \sum P_{i_1 \ldots i_k} = 1$$

where the summation is mean by all possible combinations of the indexes.

The surprising fact is it can be obtained as

$$P_{i_1 \ldots i_k} = Ck! \prod_{r=1}^{k} \frac{\lambda_{i_r}}{\mu_{i_r}},$$

where C can be calculated from the normalizing condition.

For the FIFO and Preemptive Priority disciplines the balance equations and the solution is rather complicated and they are omitted. The interested reader is referred to the cited paper. However for all cases the performance measures can be computed the same way.

Performance Measures

- *Utilization of the server*

$$U_s = \frac{\mathbb{E}(\delta)}{\mathbb{E}(\delta) + \left[\sum_{i=1}^{n} \lambda_i\right]^{-1}} = 1 - P_0.$$

- *Utilization of the machines*

  Let $U^{(i)}$ denote the utilization of machine i. Then

$$U^{(i)} = \frac{\frac{1}{\lambda_i}}{\frac{1}{\lambda_i} + \overline{T}_i} = 1 - P^{(i)},$$

  where $\overline{T}_i$ denotes the mean response time for machine i, that is the mean time while it is broken, and

$$P^{(i)} = \sum_{k=1}^{n} \sum_{i \in (i_1,\ldots,i_k)} P_{i_1,\ldots,i_k},$$

  is the probability that the ith machine is failed. Thus

$$\overline{T}_i = \frac{P^{(i)}}{\lambda_i \left(1 - P^{(i)}\right)},$$

  and in FIFO case for the main waiting time we have

$$\overline{W}_i = \overline{T}_i - \frac{1}{\mu_i}.$$

  Furthermore it is easy to see that the mean number of failed machines can be obtained as

$$\overline{N} = \sum_{i=1}^{n} P^{(i)}.$$

  In addition

$$\sum_{i=1}^{n} \lambda_i \left(1 - P^{(i)}\right) \overline{T}_i = \sum_{i=1}^{n} P^{(i)}$$

  which is the **Little's formula** for heterogeneous customers. In particular, for homogeneous case we

$$(n - \bar{N})\lambda \bar{T} = \bar{N}$$

  which was proved earlier.

Various generalized versions of the machine interference problem with heterogeneous machines can be found in Pósafalvi and Sztrik [83, 84].

Let us see some sample numerical results for the illustration of the influence of the service disciplines on the main performance measures

Inpur parameters	Machine utilizations		
	FIFO	PROC-SHARING	PRIORITY
$n = 3$			
$\lambda_1 = 0.3 \quad \mu_1 = 0.7$	0.57	0.57	0.70
$\lambda_2 = 0.3 \quad \mu_2 = 0.7$	0.75 0.57	0.74 0.57	0.74 0.58
$\lambda_3 = 0.3 \quad \mu_3 = 0.7$	0.57	0.57	0.44
Overall machine utilization	1.72	1.72	1.72
$n = 3$			
$\lambda_1 = 0.5 \quad \mu_1 = 0.9$	0.48	0.51	0.64
$\lambda_2 = 0.3 \quad \mu_2 = 0.7$	0.75 0.56	0.76 0.56	0.77 0.56
$\lambda_3 = 0.2 \quad \mu_3 = 0.5$	0.62	0.58	0.44
Overall machine utilization	1.669	1.666	1.656
$n = 4$			
$\lambda_1 = 0.5 \quad \mu_1 = 0.9$	0.38	0.429	0.64
$\lambda_2 = 0.4 \quad \mu_2 = 0.7$	0.41	0.423	0.49
	0.903	0.906	0.922
$\lambda_3 = 0.3 \quad \mu_3 = 0.6$	0.46	0.451	0.36
$\lambda_4 = 0.2 \quad \mu_4 = 0.5$	0.54	0.500	0.24
Overall machine utilization	1.814	1.804	1.751

Table 3.1: Numerical results

3.4 The $M/M/r/n/n$ Queue

Consider the homogeneous finite-source model with $r, \quad r \leq n$ independent servers. Denoting by $N(t)$ the number of customers in the system at time t similarly to the previous sections it can easily be seen that it is a birth-death process with rates

$$\lambda_k = (n-k)\lambda, \qquad 0 \leq k \leq n-1,$$

$$\mu_k = \begin{cases} k\mu & , 1 \leq k \leq r, \\ r\mu & , r < k \leq n, \end{cases}$$

The steady-state distribution can be obtained as

$$P_k = \binom{n}{k} \rho^k P_0, \qquad 0 \leq k \leq r,$$

$$P_k = \frac{k!}{r! r^{k-r}} \binom{n}{k} \rho^k P_0, \qquad r \leq k \leq n$$

with normalizing condition

$$\sum_{k=0}^{n} P_k = 1$$

To determine P_0 we can use the following simpler recursion.

Let $\frac{a_k=P_k}{P_0}$ and using the relation for the consecutive elements of the birth-death process our procedure operates as follows

$$a_0 = 1,$$
$$a_k = \frac{n-k+1}{k}\varrho a_{k-1}, \qquad 0 \le k \le r-1,$$
$$a_k = \frac{n-k+1}{r}\varrho a_{k-1}, \qquad r \le k \le n.$$

Since

$$\sum_{k=0}^{n} P_k = 1$$

must be satisfied thus we get

$$P_0 = 1 - \sum_{k=1}^{n} P_k.$$

Dividing both sides by P_0 we have

$$1 = \frac{1}{P_0} - \sum_{k=1}^{n} \frac{P_k}{P_0} = \frac{1}{P_0} - \sum_{k=1}^{n} a_k,$$

hence

$$P_0 = \frac{1}{1 + \sum\limits_{k=1}^{n} a_k}.$$

Finally

$$P_k = a_k P_0 = P_k(n).$$

Let us determine *the main performance measures*

- *Mean and variance of the number of customers in the systems* can be computed as

$$\overline{N} = \sum_{k=0}^{n} kP_k, \quad Var(N) = \sum_{k=0}^{n} k^2 P_k - (\overline{N})^2.$$

- *Mean and variance of queue length* can be obtained by

$$\overline{Q} = \sum_{k=r+1}^{n} (k-r)P_k, \quad Var(Q) = \sum_{k=r+1}^{n} (k-r)^2 P_k - (\overline{Q})^2.$$

- *Mean number of customers in the source* can be calculated by

$$\overline{m} = n - \overline{N}.$$

- *Utilization of the system* is computed by

$$U_r = 1 - P_0.$$

- *Mean busy period of the systems* can be obtained by

$$E\delta^{(n)} = \frac{1 - P_0}{n\lambda P_0} = \frac{U_r}{n\lambda P_0}.$$

- *Mean number of busy servers* can be calculated by

$$\overline{r} = \sum_{k=1}^{r} kP_k + \sum_{k=r+1}^{n} rP_k = \sum_{k=1}^{r-1} kP_k + r\sum_{k=r}^{n} P_k$$

Furthermore,

$$U_s = \frac{\sum_{k=1}^{r} kP_k + r\sum_{k=r+1}^{n} P_k}{r} = \frac{\overline{r}}{r}.$$

- *Mean number of idle servers*

$$\overline{S} = r - \overline{r}.$$

Additional relation is

$$\overline{N} = \sum_{k=1}^{r} kP_k + \sum_{k=r+1}^{n} (k-r)P_k + r\sum_{k=r+1}^{n} P_k = \overline{Q} + \overline{r} = \overline{Q} + r - \overline{S} = n - \overline{m}.$$

- *Utilization of the sources* can be calculated by

$$U_t = \sum_{k=1}^{n} \frac{n-k}{n} P_k = \frac{\overline{m}}{n}.$$

- *The mean waiting and response times* can be derived by

$$U_t = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \overline{W} + \frac{1}{\mu}} = \frac{\overline{m}}{n},$$

thus for the mean waiting time we have

$$\overline{W} = \frac{\overline{N}}{\overline{m}}\frac{1}{\lambda} - \frac{1}{\mu} = \frac{1}{\mu}\left(\frac{\overline{N}}{\overline{m}\varrho} - 1\right).$$

Hence the mean response time is

$$\overline{T} = \overline{W} + \frac{1}{\mu} = \frac{\overline{N}}{\overline{m}\lambda},$$

consequently we get

$$\overline{m}\lambda\overline{T} = \overline{N},$$

which is the well-known **Little's formula**. Thus we get

$$\overline{m}\lambda\left(\overline{W} + \frac{1}{\mu}\right) = \overline{Q} + \overline{r},$$

that is

$$\overline{m}\lambda\overline{W} + \overline{m}\varrho = \overline{Q} + \overline{r}.$$

Show that

$$\overline{r} = \overline{m}\varrho,$$

because from this follows

$$\overline{m}\lambda\overline{W} = \overline{Q}$$

which is the **Little's formula** for the waiting time.

Since

$$P_{k+1} = \frac{(n-k)\lambda}{\mu_{k+1}} P_k,$$

where

$$\mu_j = \begin{cases} j\mu & , j \leq r, \\ r\mu & , j > r. \end{cases}$$

Furthermore, it is well-known that

$$\overline{r} = \sum_{k=1}^{r-1} kP_k + r\sum_{k=r}^{n} P_k.$$

We can proceed as

$$\varrho\overline{m} = \sum_{k=0}^{n} \varrho(n-k)P_k = \sum_{k=0}^{r-1} \varrho(n-k)P_k + \sum_{k=r}^{n-1} \varrho(n-k)P_k =$$

$$= \sum_{k=0}^{r-1} \frac{\lambda(n-k)(k+1)}{(k+1)\mu} P_k + r\sum_{k=r}^{n-1} \frac{\lambda(n-k)}{r\mu} P_k =$$

$$= \sum_{k=0}^{r-1} (k+1)P_{k+1} + r\sum_{k=r}^{n-1} P_{k+1} = \sum_{j=1}^{r} jP_j + r\sum_{j=r+1}^{n} P_j = \sum_{j=1}^{r-1} jP_j + r\sum_{j=r}^{n} P_j = \overline{r}.$$

Finally, we get

$$\varrho\overline{m} = \overline{r},$$

or in another form

$$\lambda\overline{m} = \mu\overline{r},$$

that is

$$\text{mean arrival rate} = \text{mean service rate},$$

which was expected because the system is in steady state. Consequently

$$\overline{W} = \frac{\overline{Q}}{\overline{m}\lambda} = \frac{\overline{Q}}{\overline{r}\lambda\varrho} = \frac{\overline{Q}}{\mu\overline{r}}.$$

- *Mean idle period of a server* can be computed as follows.
  If the idle servers start their busy period in the order as they finished the previous busy period, then their activity can be written as follows. If a server becomes idle and finds other $j-1$ servers idle, then his busy period start at the instant of the arrival of the j th customer.

$$\bar{e}_j = \frac{r-j}{(n-j)\lambda}, \quad \text{j} = 0,1,...,\text{r-1}$$
$$a_j = \frac{P_j(n-1)}{\sum_{i=0}^{r-1} P_i(n-1)} = \frac{\Pi_j(n)}{\sum_{i=0}^{r-1} \Pi_i(n)}$$
$$\Pi_j(n) = \frac{(n-j)P_j(n)}{\sum_{i=0}^{n-1}(n-i)P_i(n)} = P_j(n-1)$$
$$\bar{e} = \sum_{j=0}^{r-1} \bar{e}_j a_j.$$

- *Mean busy period of the servers* can be calculated as follows.
  Since
$$U_s = \frac{\bar{r}}{r} = \frac{E\delta}{\overline{e} + E\delta},$$
  thus
$$E\delta = \frac{U_s}{1-U_s}\overline{e}.$$

Distribution Function of the Waiting and Response Time

This subsection is devoted to the most complicated problem of this system, namely to the determination of the distribution function of the waiting and response times. First the density function is calculated and then we obtain the distribution function. You may remember that the distribution has been given in the form

$$P_k = \begin{cases} \binom{n}{k}\rho^k P_0 \\ \\ \dfrac{\binom{n}{k}k!\rho^k}{r!r^{k-r}}P_0. \end{cases}$$

Introducing $z = \frac{1}{\rho}$, this can be written as

$$P_k = \begin{cases} \binom{n}{k}z^{-k} P_0 \\ \\ \dfrac{\binom{n}{k}k!z^{-k}}{r!r^{k-r}}P_0 \end{cases}$$

thus

$$\begin{aligned} P_k &= \frac{\binom{n}{k} k! r^r (rz)^{-k}}{r!} P_0 \\ &= \frac{n! r^r (rz)^{n-k} \cdot e^{-rz}}{(n-k)! r! (rz)^n \cdot e^{-rz}} P_0 \\ &= \frac{r^r}{r!} \frac{P(n-k, rz)}{P(n, rz)} P_0, \qquad k \geq r. \end{aligned}$$

Since

$$\begin{aligned} &\Pi_k(n) = P_k(n-1), \quad \text{thus} \\ &\Pi_k(n) = \frac{r^r}{r!} \frac{P(n-1-k, rz)}{P(n-1, rz)} P_0(n-1), \qquad \text{ha } k = r, \ldots, n-1. \end{aligned}$$

It is easy to see that the probability of waiting is

$$\sum_{k=r}^{n-1} \Pi_k(n) = \sum_{k=r}^{n-1} P_k(n-1) = P_W = P(W > 0).$$

Inserting z this can be rewritten as

$$\begin{aligned} P_W &= \sum_{k=r}^{n-1} \frac{r^r}{r!} \frac{P(n-1-k, rz)}{P(n-1, rz)} P_0(n-1) \\ &= \frac{r^r}{r!} P_0(n-1) \frac{\sum_{i=0}^{n-1-r} P(i, rz)}{P(n-1, rz)} \\ &= \frac{r^r}{r!} \frac{Q(n-1-r, rz)}{P(n-1, rz)} P_0(n-1). \end{aligned}$$

We show that the distribution function of the waiting time can be calculated as

$$F_W(x) = 1 - \frac{r^r Q(n-1-r, r(z+\mu x))}{r! P(n-1, rz)} P_0(n-1),$$

and thus

$$F_W(0) = 1 - \frac{r^r Q(n-1-r, rz)}{r! P(n-1, rz)} P_0(n-1)$$

which is probability that an arriving customer finds idle server. For the density function we have

$$\begin{aligned} &f_W(0) = 1 - P_W, \\ &f_W(x) = \mu r \frac{r^r P(n-1-r, r(z+\mu x))}{r! P(n-1, rz)} P_0(n-1), \qquad x > 0. \end{aligned}$$

If we calculate the integral $\int_{0+}^{\infty} f_W(x)\mathrm{d}x$-t that is 0 is not considered then

$$\int_{0+}^{\infty} f_W(x)\mathrm{d}x = \frac{r^r P_0(n-1)}{r!P(n-1,rz)} \cdot \int_{0+}^{\infty} \mu r \frac{(r(z+\mu t))^{n-1-r}}{(n-1-r)!} e^{-r(z+\mu t)}\mathrm{d}t.$$

By the substitution $y = r(z+\mu t)$ we have $\frac{\mathrm{d}t}{\mathrm{d}y} = \frac{1}{\mu}$ for the integral part we get

$$\int_{rz}^{\infty} \frac{y^{n-1-r}}{(n-1-r)!} e^{-y}\mathrm{d}y = Q(n-1-r,rz)$$

that is

$$\int_{0+}^{\infty} f_W(x)\mathrm{d}x = \frac{r^r Q(n-1-r,rz)}{r!P(n-1,rz)} P_0(n-1) = P_W,$$

as it was expected. Thus

$$\int_{0}^{\infty} f_W(x)\mathrm{d}x = f_W(0) + \int_{0+}^{\infty} f_W(x)\mathrm{d}x = 1.$$

Let us determine the density function for $x > 0$. That is

$$\begin{aligned}
f_W(x) &= \sum_{k=r}^{n-1} r\mu \frac{(r\mu x)^{k-r}}{(k-r)!} e^{-r\mu x} P_k(n-1) \\
&= \sum_{k=r}^{n-1} r\mu \frac{(r\mu x)^{k-r}}{(k-r)!} e^{-r\mu x} \frac{r^r}{r!} \frac{P(n-1-k,rz)}{P(n-1,rz)} P_0(n-1) \\
&= \frac{r\mu r^r P_0(n-1)}{r!P(n-1,rz)} \sum_{k=r}^{n-1} \frac{(r\mu x)^{k-r}}{(k-r)!} \frac{(rz)^{n-1-k}}{(n-1-k)!} e^{-r(z+\mu x)} \\
&= \frac{r\mu r^r P_0(n-1) e^{-r(z+\mu x)}}{r!P(n-1,rz)} \sum_{i=0}^{n-1-r} \frac{(r\mu x)^i}{i!} \frac{(rz)^{n-1-r-i}}{(n-1-r-i)!} \\
&= \frac{r\mu r^r P_0(n-1)}{r!P(n-1,rz)} \frac{(r(z+\mu x))^{n-1-r}}{(n-1-r)!} \cdot e^{-r(z+\mu x)} \\
&= \frac{r\mu r^r P_0(n-1) P(n-1-r, r(z+\mu x))}{r!P(n-1,rz)},
\end{aligned}$$

as we got earlier, but we have to remember that

$$f_W(0) = 1 - P_W.$$

Therefore

$$\begin{aligned} P(W > x) &= \int_x^\infty f_W(t)\mathrm{d}t \\ &= \frac{r^r P_0(n-1)}{r!P(n-1,rz)} \int_x^\infty r\mu \frac{(r(z+\mu t))^{n-1-r}}{(n-1-r)!} e^{-r(z+\mu t)}\mathrm{d}t \\ &= \frac{r^r P_0(n-1)}{r!P(n-1,rz)} \int_{r(z+\mu x)}^\infty \frac{y^{n-1-r}}{(n-1-r)!} e^{-y}\mathrm{d}y \\ &= \frac{r^r P_0(n-1)Q(n-1-r, r(z+\mu x))}{r!P(n-1,rz)}. \end{aligned}$$

Thus for the distribution function we have

$$F_W(x) = 1 - P(W > x)$$

which was obtained earlier.

To verify the correctness of the formula let $r = 1$. After substitution we get

$$P(W > x) = \frac{P_0(n-1)Q(n-2, z+\mu x)}{P(n-1,z)},$$

but

$$P_0(n-1) = \frac{P(n-1,z)}{Q(n-1,z)},$$

thus

$$P(W > x) = \frac{Q(n-z, z+\mu x)}{Q(n-1,z)}.$$

The derivation of the distribution function of the response time is analogous. Because the calculation is rather lengthy it is omitted, but can be found in the Solution Manual for Kobayashi [64].
As it can be seen in Allen [3], Kobayashi [64], the following formulas are valid for $r \geq 2$

$$F_T(x) = 1 - C_1 e^{-\mu x} + C_2 Q(n-r-1, r(z+\mu x)),$$

where

$$\begin{aligned} C_1 &= 1 + C_2 Q(n-r-1, rz), \\ C_2 &= \frac{r^r P_0(n-1)}{r!(r-1)(n-r-1)!P(n-1,rz)}. \end{aligned}$$

Hence the density function can be obtained as

$$f_T(x) = \mu C_1 e^{-\mu x} - C_2 r\mu P(n-r-1, r(z+\mu x)).$$

It should be noted that for the normalizing constant we have the following recursion

$$P_0^{-1}(n) = 1 + \frac{n}{rz} P_0^{-1}(n-1) + \frac{n}{z} \sum_{i=0}^{r-1} \frac{\binom{n-1}{i}}{z^i} \left(\frac{1}{i+1} - \frac{1}{r} \right), \qquad n > r,$$

with initial value

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

Since the conditional waiting time is Erlang distributed, it is easy to see that

$$E(W^2) = \sum_{k=r}^{K-1} \frac{(k-r+1) + (k-r+1)^2}{(r\mu)^2} \Pi_k, \quad Var(W) = E(W^2) - (E(W))^2,$$

$$Var(T) = Var(W) + 1/\mu^2.$$

Laplace-transform of the Waiting and Response Times

First determine the Laplace-transform of the waiting time.
It is easy to see that by using the theorem of total Laplace-transform we have

$$L_W(s) = 1 - P_W + \sum_{k=r}^{n-1} \left(\frac{r\mu}{r\mu + s} \right)^{k-r+1} P_k(n-1).$$

We calculate this formula step-by-step. Namely we can proceed as

$$\sum_{k=r}^{n-1} \left(\frac{r\mu}{r\mu + s} \right)^{k-r+1} \frac{r^r P_0(n-1) P(n-1-k, rz)}{r! P(n-1, rz)}$$
$$= \frac{r^r P_0(n-1) e^{-rz}}{r! P(n-1, rz)} \sum_{k=r}^{n-1} \left(\frac{r\mu}{r\mu + s} \right)^{k-r+1} \frac{(rz)^{n-1-k}}{(n-1-k)!}.$$

Then

$$\sum_{k=r}^{n-1} \left(\frac{r\mu}{r\mu + s} \right)^{k-r+1} \frac{(rz)^{n-1-k}}{(n-1-k)!}$$
$$= \sum_{i=0}^{n-1-r} \left(\frac{r\mu}{r\mu + s} \right)^{i+1} \frac{(rz)^{n-1-r-i}}{(n-1-r-i)!},$$

where $i = k - r$. Thus the last equation can be written as

$$\left(\frac{r\mu}{r\mu + s} \right)^{n-r} \cdot \sum_{i=0}^{n-1-r} \frac{\left(\frac{r\mu+z}{\lambda}\right)^{n-1-r-i}}{(n-1-r-i)!} = \left(\frac{r\mu}{r\mu + s} \right)^{n-r} e^{\frac{r\mu+z}{\lambda}} Q\left(n-1-r, \frac{r\mu+z}{\lambda} \right).$$

Finally collecting all terms we get

$$L_W(s) = 1 - P_W + \left(\frac{r\mu}{r\mu+s}\right)^{n-r} \frac{r^r P_0(n-1)e^{-rz}}{r!P(n-1,rz)} e^{\frac{r\mu+s}{\lambda}} Q\left(n-1-r, \frac{r\mu+s}{\lambda}\right)$$
$$= 1 - P_W + \frac{r^r e^{\frac{s}{\lambda}} P_0(n-1) Q\left(n-1-r, \frac{r\mu+s}{\lambda}\right)}{r!P(n-1,rz)} \left(\frac{r\mu}{r\mu+s}\right)^{n-r}.$$

To verify the correctness of the formula let $r = 1$.
Thus after inserting we have

$$L_W(s) = P_0(n-1) + \left(\frac{\mu}{\mu+s}\right)^{n-1} \frac{e^{\frac{s}{\lambda}} P_0(n-1) Q\left(n-2, \frac{\mu+s}{\lambda}\right)}{P(n-1,z)}$$
$$= \frac{P(n-1,z)}{Q(n-1,z)} + \left(\frac{\mu}{\mu+s}\right)^{n-1} \frac{e^{\frac{s}{\lambda}} Q\left(n-2, \frac{\mu+s}{\lambda}\right)}{Q(n-1,z)}$$
$$= \frac{\left(\frac{\mu}{\mu+s}\right)^{n-1}}{Q(n-1,z)} \left[\frac{\left(z\left(\frac{\mu+s}{\mu}\right)\right)^{n-1}}{(n-1)!} e^{-z} + e^{\frac{s}{\lambda}} Q\left(n-2, \frac{\mu+s}{\lambda}\right)\right]$$
$$= \frac{\left(\frac{\mu}{\mu+s}\right)^{n-1}}{Q(n-1,z)} \left[\frac{\left(\frac{\mu+s}{\lambda}\right)^{n-1} e^{-\frac{\mu+s}{\lambda}} e^{\frac{s}{\lambda}}}{(n-1)!} + e^{\frac{s}{\lambda}} Q\left(n-2, \frac{\mu+s}{\lambda}\right)\right]$$
$$= \frac{\left(\frac{\mu}{\mu+s}\right)^{n-1} e^{\frac{s}{\lambda}} Q\left(n-1, \frac{\mu+s}{\lambda}\right)}{Q\left(n-1,z\right)},$$

as we got earlier.

Keeping in mind the relation between the waiting time and the response time and the properties of the Laplace-transform we have

$$L_T(s) = \left(\frac{\mu}{\mu+s}\right) L_W(s),$$

which is in the case of $r = 1$ reduces to

$$L_T(s) = \left(\frac{\mu}{\mu+s}\right)^{n} \frac{e^{\frac{s}{\lambda}} Q\left(n-1, \frac{\mu+s}{\lambda}\right)}{Q(n-1,z)}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Example 28 *A factory possesses* 20 *machines having mean lifetime of* 50 *hours. The mean repair time is* 5 *hours and the repairs are carried out by* 3 *repairmen. Find the performance measures of the system.*

Solution:

$$\rho = \frac{\lambda}{\mu} = \frac{\frac{1}{\mu}}{\frac{1}{\lambda}} = \frac{5}{50} = \frac{1}{10} = 0.1$$

By using the recursive approach we get

$$a_0 = 1$$

$$a_1 = \frac{20-0}{0+1} \times 0.1 \times 1 = 2$$

$$a_2 = \frac{20-1}{1+1} \times 0.1 \times 2 = 1.9$$

$$a_3 = \frac{20-2}{2+1} \times 0.1 \times 1.9 = 1.14$$

$$a_4 = \frac{20-3}{3} \times 0.1 \times 1.14 = 0.646$$

$$\vdots$$

and so on.
Hence

$$P_0 = \frac{1}{1 + \sum_{k=1}^{n} a_k} = \frac{1}{1 + 6.3394} = 0.13625.$$

Innen

$$P_1 = a_1 \times P_0 = 2 \times 0.13625 = 0.2775$$

$$P_2 = a_2 \times P_0 = 1.9 \times 0.13625 = 0.2588 \text{ etc}$$

The distribution can be seen in the next Table for
$n = 20$, $r = 3$, $\rho = 0.1$

K	Number of busy under repair repairmen	Number of waiting machines (Q)	Number of idle repairmen (S)	Steady-state. distribution (P_k)
0	0	0	3	0.13625
1	1	0	2	0.27250
2	2	0	1	0.25888
3	3	0	0	0.15533
4	3	1	0	0.08802
5	3	2	0	0.04694
6	3	3	0	0.02347
7	3	4	0	0.01095
8	3	5	0	0.00475
9	3	6	0	0.00190
10	3	7	0	0.00070
11	3	8	0	0.00023
12	3	9	0	0.00007

Hence the performance measures are

$$\overline{Q} = 0.339, \qquad \overline{S} = 1.213, \qquad \overline{N} = \overline{Q} + r - \overline{S} = 2.126$$

$$P(W > 0) = 0.3323, \qquad P(e) = 0.6677, \qquad \overline{W} = \frac{\overline{Q}}{\lambda(n - \overline{N})} = 0.918 \text{ hours, } 58 \text{ minutes}$$

$$\overline{m} = 20 - 2.126 = 17.874, \qquad U^{(n)} = 0.844$$

$$E\delta^{(n)} = \frac{U^{(n)}}{n\lambda P_0} = \frac{5}{2} \times \frac{0.844}{0.136} \approx 15.5 \text{ hours}, \qquad \overline{r} = 1.787, \qquad \overline{s} = 1.213$$

$$U_S = \frac{\overline{r}}{r} = \frac{1.787}{3} = 0.595, \qquad \overline{e} = \frac{\overline{s}}{P(e)\lambda} = \frac{1.213}{0.667 \times \frac{1}{50}} = \frac{50 \times 1.213}{0.667} \approx 90.8 \text{ hours}$$

$$E\delta = \frac{\overline{r}}{P(e)\lambda} = \frac{1.787}{0.667 \times \frac{1}{50}} = \frac{50 \times 1.787}{0.667} \approx 132.1 \text{ hours}$$

$$U_g = \frac{\overline{m}}{n} = \frac{17.874}{20} \approx 0.893$$

$$\overline{T} = \overline{W} + \frac{1}{\mu} = 0.981 + 5 = 5.981 \text{ hours}$$

$$K_1 = \frac{\text{mean number of waiting machines}}{\text{total number of machines}} = \frac{\overline{Q}}{n} = \frac{0.339}{20} = 0.0169$$

$$K_2 = \frac{\text{mean number of idle repairmen}}{\text{total number of repairmen}} = \frac{\overline{S}}{r} = \frac{1.213}{3} = 0.404$$

Let us compare these measures to the system where we have 6 machines and a single repairman. The lifetime and repair time characteristics remain the same. The result can be seen in the next Table

Number of machines	6	20
Number of repairman	1	3
Number of machines per repairman	6	$6\frac{2}{3}$
Waiting coefficient for the servers K_2	0.4845	0.4042
Waiting coefficient for the machines K_1	0.0549	0.01694

■

Example 29 *Let us continue the previous Example with cost structure. Assume that the waiting cost is* 18 000 *Euro/hour and the cost for an idle repairman is* 600 *Euro/hour. Find the optimal number of repairmen. It should be noted that different cost functions can be constructed.*

Solution:

The mean cost per hour as a function of r can be seen in the next Table which are calculated by the help of the distribution listed below for $r = 3, 4, 5, 6, 7$.

r	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
3	0.136	0.272	0.258	0.155	0.088	0.047	0.023	0.011	0.005
4	0.146	0.292	0.278	0.166	0.071	0.028	0.010	0.003	0.001
5	0.148	0.296	0.281	0.168	0.071	0.022	0.006	0.001	0.000
6	0.148	0.297	0.282	0.169	0.072	0.023	0.006	0.001	$\cdots$
7	0.148	0.297	0.282	0.169	0.072	0.023	0.006	$\cdots$	$\cdots$

The mean cost per hour is

r	$\overline{Q}$	$\overline{S}$	$E(Cost)$ Euro
3	0.32	1.20	6480
4	0.06	2.18	2388
5	0.01	3.17	2082
6	0	4.17	2502
7	0	5.16	3096

Hence the optimal number is $r = 5$.
This simple Example shows us that there are different criteria for the optimal operation.
■

3.5 The $M/M/r/K/n$ Queue

This system is an combination of the finite-source systems considered in the previous sections. It is the most general system since for $K = r$ we have the Engset system treated in Section 3.1, for $r = 1, \quad K = n$ get the system analyzed in Section 3.2, for $K = n$ we obtain the system of Section 3.4. For the value $r < K < n$ we have delay-loss system, that is customers can arrive into the system until the number of customers in the system is $K - 1$ but then the must return to the source because the system is full.
As before it is easy to see that the number of customers in the systems is a birth-death process with rates

$$\lambda_k = (n-k) \quad , \qquad 0 \le k < K,$$

$$\mu_k = \begin{cases} k\mu & , \qquad 1 \le k \le r, \\ r\mu & , \qquad r < k \le K \end{cases}$$

where $1 \le r \le n$, $r \le K \le n$. It is rather complicated system and have not been investigated, yet. The main problem is that there are no closed form formulas as before, but using computers all the performance measures can be obtained. The normalizing constant $P_0(n, r, K)$ should satisfies the normalizing condition

$$\sum_{i=0}^{K} P_k(n, r, K) = 1.$$

As before it can easily be seen that

$$P_k(n, r, K) = \begin{cases} \binom{n}{k} \rho^k P_0(n, r, K) & , \qquad 0 \le k < r, \\ \\ \dfrac{\binom{n}{k} k! \rho^k}{r! r^{k-r}} P_0(n, r, K) & , \qquad r \le k \le K \end{cases}$$

The main *performance measures* can be computed as

$$\overline{N} = \sum_{k=0}^{K} k P_k, \quad Var(N) = \sum_{k=0}^{K} k^2 P_k - (\overline{N})^2,$$

$$\overline{Q} = \sum_{k=r}^{K} (k-r) P_k, \quad Var(Q) = \sum_{k=r}^{K} (k-r)^2 P_k - (\overline{Q})^2,$$

$$\overline{r} = \sum_{k=1}^{r-1} k P_k + r \sum_{k=r}^{K} P_k, \quad \overline{m} = n - \overline{N},$$

$$U_S = \frac{\overline{r}}{r}, \qquad U_t = \frac{n-\overline{N}}{n}, \qquad \overline{\lambda} = \overline{\mu} = \mu\overline{r},$$
$$\overline{T} = \frac{\overline{N}}{\overline{\lambda}}, \qquad \overline{W} = \frac{\overline{Q}}{\overline{\lambda}}, \qquad \overline{W} = \overline{T} - \frac{1}{\mu},$$
$$\frac{n-\overline{N}}{n} = \frac{\mathbb{E}(\tau)}{\mathbb{E}(\tau)+\overline{T}}, \qquad \mathbb{E}(\tau) = \frac{(n-\overline{N})\overline{T}}{\overline{N}}, \qquad \overline{N_R} = \mathbb{E}(\tau)\lambda.$$

By using the Bayes's rule it is easy to see that for the *probability of blocking* we have

$$P_B(n,r,K) = \frac{(n-K)P_K(n,r,K)}{\sum_{i=0}^{K}(n-i)P_i(n,r,K)} = P_K(n-1,r,K).$$

In particular, if $K = n$, then

$$\overline{\lambda} = \lambda(n-\overline{N}) = \mu\overline{r},$$

thus

$$\overline{T} = \frac{\overline{N}}{\lambda(n-\overline{N})}, \qquad \mathbb{E}(\tau) = \frac{1}{\lambda}, \qquad P_B = 0,$$

as it was expected.
Furthermore, by elementary calculations it can be seen that the normalizing constant $P_0(n,r,K)$ can be expressed recursively with respect to K under fixed r, n. Namely we have

$$(P_0(n,r,K))^{-1} = (P_0(n,r,K-1))^{-1} + \frac{\binom{n}{K}K!\rho^K}{r!r^{K-r}},$$

with initial value

$$(P_0(n,r,r))^{-1} = \sum_{i=0}^{r}\binom{n}{i}\rho^i.$$

By using the Bayes's rule it is easy to see that the probability that an arriving customer finds k customers in the system is

$$\Pi_k^*(n,r,K) = P_k(n-1,r,K), \qquad k = 0,\cdots,K$$

but the probability that a customer arriving into the systems finds k customers there is

$$\Pi_k(n,r,K) = \frac{(n-k)P_k(n,r,K)}{\sum_{i=0}^{K-1}(n-i)P_i(n,r,K)}, \qquad k = 0,\ldots,K-1.$$

Hence the probability of waiting and the density function of the waiting time can be expressed as

$$\begin{aligned} P_W(n,r,K) &= \sum_{k=r}^{K-1} \Pi_k(n,r,K) \\ f_W(0) &= 1 - P_W(n,r,K) \\ f_W(x) &= \sum_{k=r}^{K-1} \frac{(r\mu)(r\mu x)^{k-r+1}}{(k-r+1)!} e^{-r\mu x} \cdot \Pi_k(n,r,K) \end{aligned}$$

By using the Bayes's rule it can easily be verified that

$$\Pi_k(n,r,K) = \frac{P_k(n-1,r,K)}{1 - P_K(n-1,r,K)},$$

and analogously to the earlier arguments for the density function we obtain

$$f_W(x) = \frac{\mu r r^r P(K-1-r, rz)}{r! P(K-1, rz)} \frac{P_0(n-1)}{1 - P_K(n-1,r,K)}.$$

In particular, if $K = n$, that is all customer may enter into the system, then $P_K(n-1,r,K) = 0$ and thus we got the formulas derived before.

$$\begin{aligned} P(W > x) &= \sum_{k=r}^{K-1} \sum_{j=0}^{k-r} \frac{(r\mu x)^j}{j!} \bar{e}^{r\mu x} \Pi_k(n,r,K) \\ P(W \leq x) &= 1 - P(W > x) \\ P(W = 0) &= \sum_{k=0}^{r-1} \Pi_k(n,r,K). \end{aligned}$$

In $M/M/1/K/n$ systems we have

$$\begin{aligned} P(T > x) &= \sum_{k=0}^{K-1} \sum_{j=0}^{k} \frac{(\mu x)^j}{j!} \bar{e}^{\mu x} \Pi_k(n,1,K) \\ P(T < x) &= 1 - P(T > x) \\ P(T > 0) &= \sum_{k=0}^{K-1} \Pi_k(n,r,K) = 1 \\ P(T < 0) &= 0. \end{aligned}$$

By reasonable modifications for the distribution function we have

$$F_W(x) = 1 - \frac{r^r Q(K-1-r, r(z+\mu x))}{r! P(K-1, rz)} \frac{P_0(n-1,r,K)}{1 - P_K(n-1,r,K)}.$$

The corresponding Laplace-transform can be computed as

$$L_W(s) = 1 - P_W(n,r,K) + \left(\frac{r\mu}{r\mu+s}\right)^K \frac{r^r e^{\frac{s}{\lambda}}}{r!} \frac{Q\left(K-1-r, \frac{r\mu+s}{\lambda}\right) P_0(n-1,r,K)}{P(K-1,rz)(1-P_K(n-1,r,K))}.$$

Since the conditional waiting time is Erlang distributed, it is easy to see that

$$E(W^2) = \sum_{k=r}^{K-1} \frac{(k-r+1)+(k-r+1)^2}{(r\mu)^2} \Pi_k(n,r,K), \quad Var(W) = E(W^2) - (E(W))^2,$$

$$Var(T) = Var(W) + 1/\mu^2.$$

Utilization of the system is computed by

$$U_r = 1 - P_0.$$

Mean busy period of the systems can be obtained by

$$E\delta^{(n,r,K)} = \frac{1-P_0}{n\lambda P_0} = \frac{U_r}{n\lambda P_0}.$$

Mean idle period of a server can be evaluated as follows.
If the idle servers start their busy period in the order as they finished the previous busy period, then their activity can be written as follows. If a server becomes idle and finds other $j-1$ servers idle, then his busy period start at the instant of the arrival of the j th customer.

$$\bar{e}_j = \frac{r-j}{(n-j)\lambda}, \quad \text{j} = 0,1,...,\text{r-1}$$

$$a_j = \frac{P_j(n-1,r,K)}{\sum_{i=0}^{r-1} P_i(n-1,r,K)} = \frac{\Pi_j(n,r,K)}{\sum_{i=0}^{r-1} \Pi_i(n,r,K)} \qquad K = r,...,n$$

$$\bar{e} = \sum_{j=0}^{r-1} \bar{e}_j a_j, \qquad a = \frac{\bar{r}}{r}, \qquad E(\delta) = \frac{a}{1-a}\bar{e}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

3.6 The $M/M/c/K/n$ Queue with Balking and Reneging

Exactly the same was as we dealt with an $M/M/c/K$ system we can introduce the balking probabilities and reneging intensities. The balking can be represented as series of monotonically decreasing functions of the system size multiplying the corresponding arrival rate. Let b_k be this function, so that $\lambda_k = b_n\lambda$ and $b_{k+1} \leq b_k \leq 1, k > 0, b_0 = 1$, that is the probability of joining the system provided it is in state k.

Possible examples that may be useful for the $b_k = 1/(k+1), k = 1, ..., K$ Now if k customers are in the system, an estimate for the average waiting time might be $k/c\mu$, if the customer had an idea of μ. In this case $b_k = e^{-\frac{k\alpha}{c\mu}}$. The $M/M/c/K/n$ system can be obtained as $b_k = 1, k = 0, ..., K$.
Let $r_k h + o(h) =$ probability of reneging during h given k customers in the system, that is the reneging intensity is r_k. A good possibility for the reneging function r_k is
$r_k = 0,\ k = 0, ..., K$ classical system, $r_k = (k-c)\theta$, $r_k = e^{\frac{k\alpha}{c\mu}}, k = c, ...K$, and zero otherwise, where θ is the parameter of the exponentially distributed impatience time of a customer.

It is not so difficult to see, that the number of customers in the systems is a birth-death process with

$$\lambda_k = (n-k)\lambda b_k,\ k = 0, \cdots, K-1$$

$$\mu_k = \begin{cases} k\mu, & k = 1, \cdots, c \\ c\mu + r_k, & k = c, \cdots, K. \end{cases}$$

As usual, the steady-state distribution can be obtained as

$$P_k = \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} P_0, \qquad P_0 = \left(1 + \sum_{j=1}^{K} \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j}\right)^{-1}$$

The **main performance measures** can be calculated as follows

$$U_r = 1 - P_0,\ E(\delta_r) = \frac{1}{\lambda} \cdot \frac{U_s}{1 - U_s}$$
$$\bar{N} = \sum_{k=1}^{K} kP_k,\ \bar{Q} = \sum_{k=c}^{K} (k-c)P_k$$
$$\bar{N^2} = \sum_{k=1}^{K} k^2 P_k,\ \bar{Q^2} = \sum_{k=c}^{K} (k-c)^2 P_k$$
$$Var(N) = \bar{N^2} - (\bar{N})^2,\ Var(Q) = \bar{Q^2} - (\bar{Q})^2$$

$$\bar{c} = \sum_{k=1}^{c-1} kP_k + \sum_{k=c}^{K} cP_k,\ \bar{N} = \bar{Q} + \bar{c},\ U_c = \bar{c}/c$$
$$\bar{m} = n - \bar{N},\ U_t = \bar{m}/n$$
$$\bar{\lambda} = \sum_{k=0}^{K-1} \lambda_k P_k,\ \bar{\mu} = \sum_{k=1}^{K} \mu_k P_k,\ \bar{\lambda} = \bar{\mu}$$
$$\bar{T} = \bar{N}/\bar{\lambda},\ \bar{W} = \bar{Q}/\bar{\lambda}$$
$$\bar{r} = \sum_{k=c}^{K} r_k P_k,\ \text{mean reneging rate}$$

The probability that an entering customer finds k customers in the system is

$$\Pi_k = \frac{\lambda_k P_k}{\bar{\lambda}}, \qquad k = 0, \ldots K-1.$$

$$P(\text{an arriving customer enters the system}) = \frac{\bar{\lambda}}{\sum_{k=0}^{K-1}(n-k)\lambda P_k},$$

$$P(\text{a departing customer leaves the system without service}) = \frac{\bar{r}}{\bar{\mu}}$$

$$P(\text{waiting}) = \sum_{k=c}^{K-1} \Pi_k, \qquad P(\text{blocking}) = \frac{\lambda_K P_K}{\sum_{k=0}^{K} \lambda_k P_k}.$$

In the case of a balking system we can calculate the variance of waiting and response time and the distribution function of the waiting time, too.
Namely, we have

$$\overline{W} = \frac{\overline{Q}}{\overline{\lambda}} = \sum_{k=c}^{K-1} \frac{(k-c+1)}{(c\mu)} \Pi_k, \qquad \overline{T} = \frac{\overline{N}}{\overline{\lambda}} = \overline{W} + 1/\mu$$

Since the conditional waiting time is Erlang distributed, it is easy to see that

$$E(W^2) = \sum_{k=c}^{K-1} \frac{(k-c+1) + (k-c+1)^2}{(c\mu)^2} \Pi_k, \quad Var(W) = E(W^2) - (E(W))^2,$$

$$Var(T) = Var(W) + 1/\mu^2.$$

Distribution function of the waiting time
As in the previous parts for $F_W(t)$ the theorem of total probability is applied resulting

$$F_W(t) = F_W(0) + \sum_{n=c}^{K-1} \Pi_n \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx$$
$$= F_W(0) + \sum_{n=c}^{K-1} \Pi_n \left(1 - \int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx\right).$$

Similarly to the previous section we have

$$F_W(t) = F_W(0) + \sum_{n=c}^{K-1} \Pi_n - \sum_{n=c}^{K-1} \Pi_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}$$
$$= 1 - \sum_{n=c}^{K-1} \Pi_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

3.7 The $M/G/1/n/n/PS$ Queue

This system is a generalization of system $M/M/1/n/n/FIFO$ treated in Section 3.2. The essential differences are the distribution of the service time and the service disciple. Since the service times are not exponentially distributed the number of customers as a stochastic process is not a Markov chain. In this Section we introduce the model which has been published in Yashkov [128].

The requests arrive from a finite-source where they spend an exponentially distributed time with parameter λ. The required service time S is generally distributed random variable with $\mathbb{E}S < \infty$. Let us denote by $G(x)$ and $g(x)$ its distribution function, density function, respectively, assuming that $(G(0^+) = 0)$. The service discipline is Processor Sharing, that is all customers in the service facility are being served but the rate is proportional to the number of customers in service.

The method of *supplementary variables* is used for the description of the behavior of the system.
Let us introduce the following random variables.
Let $\nu(t)$ denote the number of customers in the system at time t, and for $\nu(t) > 0$ let $\xi_1(t), \ldots, \xi_{\nu(t)}(t)$ denote the elapsed service time of the requests.

The stochastic process

$$X(t) = \big(\nu(t); \xi_1(t), \ldots, \xi_{\nu(t)}(t)\big)$$

is a continuous-time Markov process with discrete and continuous components which are called *piecewise linear Markov process*.

It should be noted the many practical problems can be modeled by the help of these processes and the interested reader is referred to the book of Gnedenko–Kovalenko [39].

Let

$$P_k\left(t, x_1, \ldots, x_k\right) dx_1 \ldots dx_k = P\left(\nu(t) = k; x_i \leq \xi_i < x_i + dx_i, i = 1, \ldots, k\right),$$

that is $P_k(t, x_1, \ldots, x_k)$, $k = 1, \ldots, n$ denotes the density function that at time t there are k customers in the system and their elapsed service times are $x_1, \ldots, x_k$.
Let δ be a small positive real number. Then for the density functions $P_k(t, x_1, \ldots, x_k)$ we have the following set of equations

$$P_k\left(t; x_1, \ldots, x_k\right) =$$

$$= P_k\left(t - \delta; x_1 - \frac{\delta}{k}, \ldots, x_k - \frac{\delta}{k}\right) \prod_{i=1}^{k} \frac{1 - G\left(x_i\right)}{1 - G\left(x_i - \frac{\delta}{k}\right)} \left[1 - \lambda\left(n - k\right)\delta\right] +$$

$$+ \left(k + 1\right) \int_0^{\infty} P_{k+1}\left(t - \delta; x_1 - \frac{\delta}{k}, \ldots, x_{k+1} - \frac{\delta}{k+1}\right) \times$$

$$\times \prod_{i=1}^{k} \frac{1 - G\left(x_i\right)}{1 - G\left(x_i - \frac{\delta}{k+1}\right)} \cdot \frac{G\left(x_{k+1}\right) - G\left[x_{k+1} - \frac{\delta}{k+1}\right]}{1 - G\left[x_{k+1} - \frac{\delta}{k+1}\right]} dx_{k+1}.$$

Dividing both sides by $\prod_{i=1}^{k} \left[1 - G\left(x_i\right)\right]$ and taking the limits as $\delta \to 0$, $t \to \infty$ we have the stationary equations, namely

$$\left[\frac{1}{k} \sum_{i=1}^{k} \frac{\partial}{\partial x_i} + \lambda\left(n - k\right)\right] q_k\left(x_1, \ldots, x_k\right) =$$

$$\int_0^{\infty} q_{k+1}\left(x_1, \ldots, x_{k+1}\right) g\left(x_{k+1}\right) dx_{k+1}, \;\; k = 1, \ldots, n-1,$$

where

$$q_k\left(x_1, \ldots, x_k\right) = \lim_{t \to \infty} P_k\left(t; x_1, \ldots, x_k\right) / \prod_{i=1}^{k} \left[1 - G\left(x_i\right)\right]$$

are called *normalized density functions.*

Similarly, for P_0 and $q_n(x_1, \ldots, x_n)$ we obtain

$$\lambda n P_0 = \int_0^{\infty} q_1\left(x_1\right) g\left(x_1\right) dx_1,$$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} q_n\left(x_1, \ldots, x_n\right) = 0.$$

Beside these equation we need the boundary conditions which are

$$q_1(0) = \lambda n P_0,$$

$$q_k(0, x_1, \ldots, x_{k-1}) = \lambda (n-k+1)\, q_{k-1}(x_1, \ldots, x_{k-1}),$$
$$k = 1, \ldots, n.$$

The solution to these set of integro-differential equations is surprisingly simple, namely

$$q_k(x_1, \ldots, x_k) = P_0 \lambda^k n! / [(n-k)!],$$

which can be proved by direct substitution.
Consequently

$$P_k(x_1, \ldots, x_k) = P_0 \lambda^k \frac{n!}{(n-k)!} \prod_{i=1}^{k} [1 - G(x_i)],$$
$$i = 1, \ldots, n.$$

Let us denote by P_k the steady-state probability of the number of customers in the system. Clearly we have

$$P_k = \int_0^\infty \ldots \int_0^\infty P_k(x_1, \ldots, x_k)\, dx_1 \ldots dx_k = P_0 \frac{n!}{(n-k)!} (\lambda \mathbb{E}S)^k.$$

Probability P_0 can be obtained by using the normalizing condition $\sum_{i=0}^{n} P_i = 1$.

Recall that it is the same as the distribution in the $M/M/1/n/n$ system if $\varrho = \lambda \mathbb{E}S$.

It is not difficult to see that for this $M/G/1/n/n/PS$ system the *performance measures* can be calculated as

$$(i) \qquad \overline{N} = \sum_{k=1}^{n} k P_k$$

$$(ii) \quad U^{(i)} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \overline{T}} = \frac{n - \overline{N}}{n},$$

thus

$$\overline{T} = \frac{1}{\lambda} \frac{\overline{N}}{n - \overline{N}},$$

hence

$$\lambda (n - \overline{N}) \overline{T} = \overline{N}$$

which is the **Little's formula**.
Clearly, due to the Processor Sharing discipline the response time is longer then the required service time, and there is no waiting time since each customer are being served.

The difference is $\overline{T} - \mathbb{E}(S)$.

It can be proved, see Cohen [22], that for an $\vec{G}/\vec{G}/1/n/n/PS$ system the steady-state probability that customers with indexes $i_1, \ldots, i_k$ are in the system can be written as

$$P(i_1, \ldots, i_k) = C \cdot k! \prod_{j=1}^{k} \rho i_j, \qquad \rho_i = \frac{\mathbb{E}(S_i)}{\mathbb{E}(\tau_i)}, \quad i = 1, \ldots, n.$$

For homogeneous case we get

$$P_k = C \cdot k! \binom{n}{k} \rho^k.$$

3.8 The $\vec{G}/M/r/n/n/FIFO$ Queue

This section is devoted to a generalized version the finite-source model with multiple servers where the customers are supposed to have heterogeneous generally distributed source times and homogeneous exponentially distributed service times. They are served according to the order of their arrivals. The detailed description of this model can be found in Sztrik [100].

Customers arrive 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 (FFIFO) discipline. If there is no idle server, then a waiting line is formed and the units are delayed. The service times of the units are supposed to be identically and exponentially distributed random variables with parameter μ. After completing service, customer with index i returns to the source and stays there for a random time τ_i having general distribution function $F_i(x)$ with density $f_i(x)$. All random variables are assumed to be independent of each other.

Determination of the steady-state distribution

As in the previous section the modeling is more difficult since the involved random times are not all exponentially distributed and thus we have to use the *method of supplementary variables*.

Let the random variable $\nu(t)$ denote the number of customers staying in the source at time t and let $\left(\alpha_1(t), \ldots, \alpha_{\nu(t)}\right)$ indicate their indexes ordered lexicographically, that is in increasing order of their indexes.

Let us denote by $\left(\beta_1(t), \ldots, \beta_{n-\nu(t)}\right)$ the indexes of the requests waiting or served at the service facility in the order of their arrival. It is not difficult to see that the process

$$Y(t) = \left(\nu(t); \alpha_1(t), \ldots, \alpha_{\nu(t)}; \beta_1(t), \ldots, \beta_{n-\nu(t)}\right), \qquad (t \geq 0)$$

is not Markovian unless the distribution functions $F_i(x)$, $i = 1, \ldots, n$ are exponential.

To use the supplementary variable technique let us introduce the supplementary variable $\xi\alpha_i(t)$ to denote the elapsed source time of request with index $\alpha_i, \quad i = 1, \cdots, n$. Define

$$X(t) = \left(\nu(t); \alpha_1(t), \ldots, \alpha_{\nu(t)}; \xi\alpha_1(t), \ldots, \xi\alpha_{\nu(t)}; \beta_1(t), \ldots . \beta_{n-\nu(t)}\right)$$

This is a multicomponent piecewise linear Markov process.

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 process $(X(t), t \geq 0)$ consists of the set of points

$$(i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{n-k})$$

where

$$(i_1, \ldots, i_k) \in C_k^n, (j_1, \ldots, j_{n-k}) \in V_k^n, x_i \in R_+, i = 0, 1, \ldots, k, k = 0, 1, \ldots, n.$$

Process $X(t)$ is in state $(i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{n-k})$ if k customers with indexes $(i_1, \ldots, i_k)$ have been staying in the source for times $(x_1, \ldots, x_k)$, respectively while the rest need service and their indexes in the order of arrival are $(j_1, \ldots, j_{n-k})$.

To derive the Kolmogorov-equations we should consider the transitions that can occur in an arbitrary time interval $(t, t+h)$. For $0 \leq n-k < r$ the transition probabilities are then the following

$$P[X(t+h) = (i_1, \ldots, i_k; x_1 + h, \ldots, x_k + h; j_1, \ldots, j_{n-k}) \mid$$

$$X(t) = (i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{n-k})]$$

$$= (1 - (n-k)\mu h) \prod_{l=1}^{k} \frac{1 - F_{i_l}(x_l + h)}{1 - F_{i_l}(x_l)} + o(h),$$

$$P[X(t+h) = (i_1, \ldots, i_k; x_1 + h, \ldots, x_k + h; j_1, \ldots, j_{n-k}) \mid$$

$$X(t) = (i_1', \ldots, j_{n-k}', \ldots, i_k'; x_1', \ldots, y', \ldots, x_k'; j_1, \ldots, j_{n-k-1})]$$

$$= \frac{f_{j_{n-k}}(y)\, h}{1 - F_{j_{n-k}}(y)} \prod_{l=1}^{k} \frac{1 - F_{i_l}(x_l + h)}{1 - F_{i_l}(x_l)} + o(h),$$

where $(i_1', \ldots, j_{n-k}', \ldots, i_k')$ denotes the lexicographical order of indexes $(i_1, \ldots, i_k, j_{n-k})$ while $(x_1', \ldots, y', \ldots, x_k')$ indicates the corresponding times.
For $r \leq n-k \leq n$ the transition probabilities can be obtained as

$$P[X(t+h) = (i_1, \ldots, i_k; x_1 + h, \ldots, x_k + h; j_1, \ldots, j_{n-k}) \mid$$

$$X(t) = (i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{n-k})]$$

$$= (1 - r\mu h) \prod_{l=1}^{k} \frac{1 - F_{i_l}(x_l + h)}{1 - F_{i_l}(x_l)} + o(h),$$

$$P\Big[X\left(t+h\right)=\left(i_1,\ldots,i_k;x_1+h,\ldots,x_k+h;j_1,\ldots,j_{n-k}\right)\ \mid$$

$$X\left(t\right)=\left(i_1^{'},\ldots,j_{n-k}^{'},\ldots,i_k;x_1,\ldots,y^{'},\ldots,x_k^{'};j_1,\ldots,j_{n-k-1}\right)\Big]=$$

$$=\frac{f_{j_{n-k}}\left(y\right)h}{1-F_{j_{n-k}}\left(y\right)}\prod_{l=1}^{k}\frac{1-F_{i_l}\left(x_l+h\right)}{1-F_{i_l}\left(x_l\right)}+o\left(h\right).$$

For the distribution of $X(t)$ introduce the following functions

$$Q_{0;j_1,\ldots,j_n}\left(t\right)=P\left(\nu\left(t\right)=0;\beta_1\left(t\right)=j_1,\ldots,\beta_n\left(t\right)=j_n\right),$$

$$Q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k;t\right)=$$

$$P\left(\nu\left(t\right)=k;\alpha_1\left(t\right)=i_1,\ldots,\alpha_k\left(t\right)=i_k;\xi_{i_1}\leq x_1,\ldots,\xi_{i_k}\leq x_k;\right.$$

$$\left.\beta_1\left(t\right)=j_1,\ldots,\beta_{n-k}\left(t\right)=j_{n-k}\right).$$

Let λ_i is defined by $1/\lambda_i=\mathbb{E}(\tau_i)$. Then we have

Theorem 2 *If $1/\lambda_i<\infty$, $i=1,\ldots,n$, then the process $\left(X\left(t\right),t\geq 0\right)$ possesses a unique limiting (stationary, steady-state) distribution independent of the initial conditions, namely*

$$Q_{0;j_1,\ldots,j_n}=\lim_{t\to\infty}Q_{0;j_1,\ldots,j_n}\left(t\right),$$

$$Q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k\right)=\lim_{t\to\infty}Q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k;t\right).$$

Notice that $X(t)$ belongs to the class of piecewise-linear Markov processes, subject to discontinuous changes treated by Gnedenko and Kovalenko [39]. Our statement follows from a theorem on page 211 of this monograph.

Since by assumption $F_i(x)$ has density function, for fixed k Theorem 2 provides the existence and uniqueness of the following limits

$$q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k\right)dx_1\ldots dx_k=$$

$$=P\left(\nu\left(t\right)=k;\alpha_1\left(t\right)=i_1,\ldots,\alpha_k\left(t\right)=i_l;x_l\leq\xi_{i_l}<x_l+dx_l,l=1,\ldots,k;\right.$$

$$\left.\beta_1\left(t\right)=j_1,\ldots,\beta_{n-k}\left(t\right)=j_{n-k}\right),\qquad k=1,\ldots,n$$

where $q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k\right)$ denotes the density function of state $(i_1,\ldots,i_k;x_1,\ldots,x_k;j_1,\ldots,j_{n-k})$ when $t\to\infty$.

Let us introduce the so-called *normed density function* defined by

$$\tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k\right)=\frac{q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k\right)}{\left(1-F_{i_1}\left(x_1\right)\right)\ldots\left(1-F_{i_k}\left(x_k\right)\right)}.$$

Then we have

Theorem 3 *The normed density functions satisfy the following system of integro-differential equations (3.1), (3.3) with boundary conditions (3.2), (3.4)*

$$\left[\frac{\partial}{\partial x_1}+\ldots+\frac{\partial}{\partial x_k}\right]^* \tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_k) \tag{3.1}$$

$$=-(n-k)\,\mu\tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_k)+$$

$$+\int_0^\infty \tilde{q}_{i_1',\ldots,j_{n-k}',\ldots,i_k';j_1,\ldots,j_{n-k-1}}\left(x_1',\ldots,y',\ldots,x_k'\right) f_{j_n}(y)\,dy,$$

$$\tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_{l-1},0,x_{l+1},\ldots,x_k) = \tag{3.2}$$

$$=\mu\sum_{V^{i_l}_{j_1,\ldots,j_{n-k}}} \tilde{q}_{i_1,\ldots,i_{l-1};i_{l+1},\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_k)$$

$$for\ l=1,\ldots,k,\quad 0\le n-k<r$$

$$\left[\frac{\partial}{\partial x_1}+\ldots+\frac{\partial}{\partial x_k}\right]^* \tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_k) = \tag{3.3}$$

$$=-r\mu\tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_k)+$$

$$+\int_0^\infty \tilde{q}_{i_1',\ldots,j_{n-k}',\ldots,i_k';j_1,\ldots,j_{n-k-1}}(x_1',\ldots,y',\ldots,x_k')f_{j_n}(y)dy$$

$$\tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_{l-1},0,x_{l+1},\ldots,x_k) = \tag{3.4}$$

$$=\mu\sum_{V^{i_l}_{j_1,\ldots,j_{r-1}}} \tilde{q}_{i_1,\ldots,i_{l-1};i_{l+1},\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_k),$$

$$for\ l=1,\ldots,k,\quad r\le n-k<n-1$$

furthermore

$$r\mu Q_{0;j_1,\ldots,j_n}=\int_0^\infty \tilde{q}_{j_n;j_1,\ldots,j_{n-1}}(y)f_{j_n}(y)\,dy.$$

The symbol $[\ \ ]^*$ *will be explained later while*

$$V^{i_l}_{j_1,\ldots,j_s}=[(i_l,j_1,\ldots,j_s),(j_1,i_l,j_2,\ldots,j_s),\ldots,(j_1,\ldots,j_s,i_l)]\in V^n_{s+1}.$$

Proof: Since the process $(X(t),t\ge 0)$ is Markovian its densities must satisfy the Kolmogorov-equations. A derivation is based on the examination of the sample paths of the process during an infinitesimal interval of width h. The following relations hold

$$q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1+h,\ldots,x_k+h) =$$

$$= q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_k)\left(1-(n-k)\mu h\right)\prod_{l=1}^{k}\frac{1-F_{i_l}(x_l+h)}{1-F_{i_l}(x_l)}+$$

$$+\prod_{l=1}^{k}\frac{1-F_{i_l}(x_l+h)}{1-F_{i_l}(x_l)}+\int_0^{\infty}\tilde{q}_{i_1^{'},\ldots,j_{n-k}^{'},\ldots,i_k^{'};j_1,\ldots,j_{n-k-1}}\left(x_1^{'},\ldots,y^{'}\ldots,x_k^{'}\right)\times$$

$$\times\frac{f_{j_{n-k}}^{'}(y)\,h}{1-F_{j_{n-k}}(x_l)}dy+o(h),$$

$$q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1+h,\ldots,x_{l-1}+h,0,x_{l+1}+h,\ldots,x_k+h)\,h=$$

$$=o(h)+\prod_{\substack{s=1\\s\neq l}}^{k}\frac{1-F_{i_s}(x_s+h)}{1-F_{i_s}(x_s)}\times$$

$$\times\mu h\sum_{V_{j_1,\ldots,j_{n-k}}^{i_l}}\tilde{q}_{i_1,\ldots,i_{l-1};i_{l+1},\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_k)$$

$$\text{for } 0\leq n-k<r,\quad l=1,\ldots,k.$$

Similarly

$$q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1+h,\ldots,x_k+h)=$$

$$= q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_k)\left(1-r\mu h\right)\prod_{l=1}^{k}\frac{1-F_{i_l}(x_l+h)}{1-F_{i_l}(x_l)}+$$

$$+\prod_{l=1}^{k}\frac{1-F_{i_l}(x_l+h)}{1-F_{i_l}(x_l)}\int_0^{\infty}\tilde{q}_{i_1^{'},\ldots,j_{n-k}^{'},\ldots,ie_k^{'};j_1,\ldots,j_{n-k-1}}\left(x_1^{'},\ldots,y^{'},\ldots,x_k^{'}\right)\times$$

$$\times\frac{f_{j_{n-k}}^{'}(y)\,h}{1-F_{j_{n-k}}(x_l)}dy+o(h),$$

$$q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1+h,\ldots,x_{l-1}+h,0,x_{l+1}+h,\ldots,x_k+h)\,h=$$

$$=o(h)+\prod_{\substack{s=1\\s\neq l}}^{k}\frac{1-F_{i_s}(x_s+h)}{1-F_{i_s}(x_s)}\times$$

$$\times\mu h\sum_{V_{j_1,\ldots,j_{n-k}}^{i_l}}\tilde{q}_{i_1,\ldots,i_{l-1};i_{l+1},\ldots,i_k;j_1,\ldots,j_{n-k}}(x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_k)$$

$$\text{for } 0\leq n-k<r,\quad l=1,\ldots,k.$$

Finally

$$Q_{0;j_1,\ldots,j_n}=Q_{0;j_1,\ldots,j_n}\left(1-r\mu h\right)+$$

$$+\int_0^{\infty}\tilde{q}_{j_n;j_1,\ldots,j_{n-1}}(y)\,\frac{f_{j_n}(y)\,h}{1-F_{j_n}(y)}dy+o(h).$$

Thus the statement of this theorem can easily be obtained. Dividing the left-hand side of equations by $\prod_{l=1}^{k}\left(1-F_{i_l}\left(x_l+h\right)\right)$ and taking into account the definition of the normed densities taking the limit as $h \to 0$ we get the desired result.

In the left-hand side of (3.1)(3.3) used for the notation of the limit in the right-hand side, the usual notation for partial differential quotients has been applied. Strictly considering this is not allowed, since the existence of the individual partial differential quotient is not assured. This is why the operator is notated by $[\ ]^*$. Actually this is a $(1,1,\ldots,1) \in R^k$ directional derivative, see Cohen [22].

To determine the steady-state probabilities

$$\left[Q_{0;j_1,\ldots,j_n}, Q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\right],$$

$$(i_1,\ldots,i_k) \in C_k^n, \quad (j_1,\ldots,j_{n-k}) \in V_{n-k}^N, \qquad k=1,\ldots,n.$$

we have to solve equations (3.1)(3.3) subject to the boundary conditions (3.2)(3.4).
If we set

$$Q_{0;j_1,\ldots,j_n} = c_0,$$

$$\tilde{q}_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}\left(x_1,\ldots,x_k\right) = c_k, \qquad k=1,\ldots,n,$$

then by direct substitution it can easily be verified that they satisfy these equations with boundary conditions. Moreover these c_k can be obtained by the help of c_n, namely

$$c_k = \left(r!r^{n-r-k}\mu^{n-k}\right)^{-1}c_n, \qquad 0 \le k \le n-r,$$

$$c_k = \left((n-k)!\mu^{n-k}\right)^{-1}c_n, \qquad n-r \le k \le n.$$

Since these equations completely describe the system, this is the required solution.

Let $Q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}}$ denote the steady-state probability that customers with indexes $(i_1,\ldots,i_k)$ are in the source and the order of arrivals of the rest to the service facility is $(j_1,\ldots,j_{n-k})$. Furthermore, denote by $Q_{i_1,\ldots,i_k}$ the stationary probability that requests with indexes $(i_1,\ldots,i_k)$ i are staying in the source.
It can easily be seen

$$Q_{i_1,\ldots,i_k;j_1,\ldots,j_{n-k}} = \left(\lambda_{i_1},\ldots,\lambda_{i_k}\right)^{-1}c_k, \qquad k=1,\ldots,n.$$

By using the relation we obtained for c_k we have

$$Q_{i_1,\ldots,i_k} = (n-k)!\left(r!r^{n-r-k}\mu^{n-k}\lambda_{i_1},\ldots,\lambda_{i_k}\right)^{-1}c_n,$$

$$(i_1,\ldots,i_k) \in C_k^n, \qquad k=0,1,\ldots,n-r.$$

Similarly

$$Q_{i_1,\ldots,i_k} = \left(\mu^{n-k}\lambda_{i_1},\ldots,\lambda_{i_k}\right)^{-1}c_n,$$

$$(i_1,\ldots,i_k) \in C_k^n, \qquad k=n-r,\ldots,n.$$

Let us denote by $\hat{Q}_k$ and $\hat{P}_l$ the steady-state probability of the number of customers in the source, in the service facility, respectively. Hence it is easy to see that

$$Q_{i_1,\ldots,i_n} = Q_{1,\ldots,n} = \hat{Q}_n,$$
$$\hat{Q}_k = \hat{P}_{n-k}, \qquad k = 0,\ldots,n.$$

Furthermore

$$c_n = \hat{Q}_n\left(\lambda_1,\ldots,\lambda_n\right),$$
$$\hat{Q}_k = \sum_{(i_1,\ldots,i_k)\in C_k^n} Q_{i_1,\ldots,i_k},$$

where $\hat{Q}_n$ can be obtained by the help of the normalizing condition $\sum_{k=0}^{n} \hat{Q}_k = 1$.

In the homogeneous case these formulas reduce to

$$\hat{Q}_k = \frac{n!}{r!k!r^{n-k-r}}\left(\frac{\lambda}{\mu}\right)^{n-k}\hat{Q}_n, \qquad for \;\; 0 \le k \le n-r,$$

$$\hat{Q}_k = \binom{n}{k}\left(\frac{\lambda}{\mu}\right)^{n-k}\hat{Q}_n, \qquad for\; n-r \le k \le n,$$

which is the result of the paper Bunday and Scraton [17], and for $r = 1$ is the formulas obtained by Schatte [92]. Thus the distribution of the number of customers in the service facility is

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

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

This is exactly the same result that we got for the $< M/M/r/n/n >$ model.

It should be underlined that the distribution of the number of customers in the system does not depend on the form of $F_i(x)$ but the mean $1/\lambda_i$, that is it is robust.

■

Performance Measures

- *Utilization of the sources*

  Let $Q^{(i)}$ denote the steady-state distribution that source i is busy with generating a new customer, that is

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

  Hence the utilization of source i can be obtained as

$$U^{(i)} = Q^{(i)}.$$

- *Utilization of the servers*

  As we have calculated earlier the utilization of a server can be derived as

$$U_{CPU} = \frac{1}{r}\left(\sum_{k=1}^{r} k\hat{P}_k + r\sum_{k=r+1}^{n}\hat{P}_k\right) = \frac{\bar{r}}{r},$$

  where $\bar{r}$ denotes the mean number of busy servers. Thus the overall utilization of the servers is $\bar{r}$.

- *Mean waiting and response times*

  By the results of Tomkó [118] we have

$$Q^{(i)} = (1/\lambda_i)\left(1/\lambda_i + \bar{W}_i + 1/\mu\right)^{-1}.$$

  Thus for the mean waiting time of the customer with index i we obtain

$$\bar{W}_i = \frac{1}{\lambda_i}\cdot\frac{1-Q^{(i)}}{Q^{(i)}} - \frac{1}{\mu}, \qquad i = 1,\ldots,n.$$

  Consequently the mean response time $\bar{T}_i$ of the ith request can be calculated as

$$\bar{T}_i = \bar{W}_i + 1/\mu = \left(1-Q^{(i)}\right)\left(\lambda_i Q^{(i)}\right)^{-1}, \qquad i = 1,\ldots,n.$$

  Since

$$\sum_{i=1}^{n}\left(1-Q^{(i)}\right) = \bar{N},$$

  where $\bar{N}$ denotes the mean number customers at the service facility. This can be rewritten as

$$\sum_{i=1}^{n}\lambda_i\bar{T}_i Q^{(i)} = \bar{N},$$

  which the **Little's formula** for the $\vec{G}/M/r/n/n/FIFO >$ queueing system.

It should be noted that using the terminology of the machine interference problem $U^{(i)}$, $\bar{W}_i, \bar{T}_i$ denote the utilization, mean waiting time and the mean time spent in a broken state of the ith machine.

This model can be generalized such a way that the service intensities depend on the number of customers in the source, see Sztrik [102, 103].

Chapter 4

Exercises

4.1 Infinite-Source Systems

Exercise 1 *Solve the following system of equations by the help of difference equations*

$$\lambda P_0 = \mu P_1$$
$$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}, n \geq 1.$$

Solution:
It is easy to see that it can be rewritten as

$$\lambda P_{n-1} - (\lambda + \mu)P_n + \mu P_{n+1} = 0, n = 1, 2, \ldots$$

which is a 2-nd order difference equation with constant coefficient. Its general solution can be obtained in the form

$$P_n = c_1 x_1^n + c_2 x_2^n, n = 1, 2, \ldots$$

where x_1, x_2 are the solutions to

$$\mu x^2 - (\lambda + \mu)x + \lambda = 0.$$

It can easily be verified that $x_1 = 1, \qquad x_2 = \varrho$ and thus

$$P_n = c_1 + c_2 \varrho^n, n = 1, 2, \ldots.$$

However $P_1 = \varrho P_0$, and because $\sum_{n=0}^{\infty} P_n = 1$, thus $c_1 = 0$ and $c_2 = P_0 = 1 - \varrho$.
■

Exercise 2 *Find the generating function of the number of customers in the system for an $M/M/1$ queueing system by using the steady-state balance equations. Then derive the corresponding distribution.*

Solution:
Staring with the set of equations

$$\lambda P_0 = \mu P_1$$
$$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}, n \geq 1$$

by multiplying both sides by s^i and then adding the terms we obtain

$$\lambda G_N(s) + \mu G_N(s) - \mu P_0 = \lambda s G_N(s) + \frac{\mu}{s} G_N(s) - \frac{\mu}{s} P_0.$$

Thus we can calculate as

$$\begin{aligned} G_N(s)\left(\lambda(1-s) + \mu\left(1 - \frac{1}{s}\right)\right) &= \mu\left(1 - \frac{1}{s}\right) P_0, \\ G_N(s)\left(\mu\left(1 - \frac{1}{s}\right) - \lambda s\left(1 - \frac{1}{s}\right)\right) &= \mu\left(1 - \frac{1}{s}\right) P_0, \\ G_N(s) &= \frac{\mu}{\mu - \lambda s} P_0. \end{aligned}$$

Since $G_N(1) = 1$, therefore

$$P_0 = \frac{\mu - \lambda}{\mu} = 1 - \varrho.$$

That is

$$G_N(s) = \frac{1 - \varrho}{1 - s\varrho},$$

which is exactly the generating function of a modified geometric distribution with parameter $(1 - \varrho)$. It can be proved as follows, if

$$P(N = k) = (1 - \varrho)\varrho^k, k = 0, \ldots$$

then its generating function is

$$G_N(s) = \sum_{k=0}^{\infty} s^k (1 - \varrho)\varrho^k = \frac{1 - \varrho}{1 - s\varrho}.$$

■

Exercise 3 *Find the generating function of the number of customers waiting in a queue for an $M/M/1$ queueing system.*

Solution:
Clearly

$$\begin{aligned} G_Q(s) &= (P_0 + P_1)s^0 + \sum_{k=2}^{\infty} s^{k-1} P_k = P_0 + \sum_{k=1}^{\infty} s^{k-1} P_k = 1 - \varrho + \sum_{k=1}^{\infty} s^{k-1} \varrho^k (1 - \varrho) \\ &= 1 - \varrho + \varrho \sum_{i=0}^{\infty} s^i (1 - \varrho)\varrho^i = 1 - \varrho + \varrho G_N(s) = 1 - \varrho(1 - G_N(s)). \end{aligned}$$

For verification let us calculate the mean queue length, thus

$$G'_Q(1) = \varrho G'_N(1) = \frac{\varrho^2}{1 - \varrho}.$$

■

Exercise 4 *Find the Laplace-transform of T and W for an $M/M/1$ queueing system.*

Solution:
It is easy to see that

$$L_T(s) = \sum_{k=0}^{\infty} \left(\frac{\mu}{\mu+s}\right)^{k+1} \varrho^k(1-\varrho) = (1-\varrho)\frac{\mu}{\mu+s}\sum_{k=0}^{\infty}\left(\frac{\mu\varrho}{\mu+s}\right)^k$$
$$= (1-\varrho)\frac{\mu}{\mu+s}\frac{1}{1-\frac{\mu\varrho}{\mu+s}} = (1-\varrho)\frac{\mu}{\mu+s}\frac{\mu+s}{\mu(1-\varrho)+s} = \frac{\mu(1-\varrho)}{\mu(1-\varrho)+s},$$

which was expected, since T follows an exponential distribution with parameter $\mu(1-\varrho)$.
To get the Laplace-transform of W we have

$$L_W(s) = \sum_{k=0}^{\infty}\left(\frac{\mu}{\mu+s}\right)^k \varrho^k(1-\varrho) = 1-\varrho+\sum_{k=1}^{\infty}\left(\frac{\mu}{\mu+s}\right)^k \varrho^k(1-\varrho)$$

$$= 1-\varrho+\frac{\varrho\mu(1-\varrho)}{\mu(1-\varrho)+s},$$

which should be

$$\frac{L_T(s)}{\frac{\mu}{\mu+s}}$$

since

$$L_T(s) = L_W(s)\frac{\mu}{\mu+s}.$$

To show this it can be calculated that

$$L_W(s) = L_T(s)\frac{\mu+s}{\mu} = \frac{\mu(1-\varrho)}{\mu(1-\varrho)+s}\frac{\mu+s}{\mu} = 1-\varrho+\varrho\frac{\mu(1-\varrho)}{\mu(1-\varrho)+s}.$$

Let us verify the result by deriving the mean values $\overline{T}$ and $\overline{W}$.

$$L_T'(0) = -\frac{1}{\mu(1-\varrho)},$$
$$L_W'(0) = \varrho L_T'(0) = -\frac{\varrho}{\mu(1-\varrho)},$$

thus

$$\overline{T} = \frac{1}{\mu(1-\varrho)}, \overline{W} = \frac{\varrho}{\mu(1-\varrho)},$$

which was obtained earlier.
■

Exercise 5 *Show that for an $M/M/1/K$ queueing system*

$$\lim_{K\to\infty} \overline{N}(K) = \frac{\rho}{1-\rho}, \qquad \rho < 1$$

.

Solution:
It is well-known if $\rho < 1$ then

$$\lim_{K\to\infty} \rho^K = 0.$$

Since $\overline{N} = \frac{\rho\left(1-(K+1)\rho^K+K\rho^{K+1}\right)}{(1-\rho)(1-\rho^{K+1})}$, it is enough to show that

$$\lim_{K\to\infty} K\rho^K = 0.$$

This can be proved by the L'Hospital's rule, namely

$$\lim_{K\to\infty} \frac{K}{\rho^{-K}} = \frac{\infty}{\infty}$$
$$\lim_{K\to\infty} \frac{K}{\rho^{-K}} = \lim_{K\to\infty} \frac{1}{-ln\rho\rho^{-K}} = \frac{\rho^K}{-ln\rho} = 0.$$

■

Exercise 6 *Show that for an $M/M/1/K$ queueing system the Laplace-transform*

$$L_T(s) = \frac{\mu P_0}{1-P_K} \frac{1-\left(\frac{\lambda}{\mu+s}\right)^K}{\mu-\lambda+s}$$

satisfies $L_T(0) = 1$.

Solution:

$$\begin{aligned} L_T(0) &= \frac{\mu P_0}{1-P_K}\frac{1-\rho^K}{\mu-\lambda} = \frac{P_0}{1-P_K}\frac{1-\rho^K}{1-\rho} \\ &= \frac{\frac{1-\rho}{1-\rho^{K+1}}}{1-\frac{\rho^K(1-\rho)}{1-\rho^{K+1}}} \cdot \frac{1-\rho^K}{1-\rho} \\ &= \frac{1-\rho}{1-\rho^{K+1}} \cdot \frac{1-\rho^{K+1}}{1-\rho^{K+1}-\rho^K+\rho^{K+1}} \cdot \frac{1-\rho^K}{1-\rho} \\ &= 1. \end{aligned}$$

■

Exercise 7 *Find $\overline{T}$ by the help of the Laplace-transform for an $M/M/1/K$ queueing system.*

Solution:

Since

$$L_T(s) = \frac{\mu P_0}{1-P_K} \frac{1-\left(\frac{\lambda}{\mu+s}\right)^K}{\mu-\lambda+s}$$

then

$$L'_T(s) = \frac{\mu P_0}{1-P_K} \left(\frac{K\left(\frac{\mu+s}{\lambda}\right)^{-(K+1)} \cdot \frac{1}{\lambda}(\mu-\lambda+s) - \left(1-\left(\frac{\lambda}{\mu+s}\right)^K\right)}{(\mu-\lambda+s)^2} \right)$$

$$\begin{aligned} L'_T(0) &= \frac{\mu P_0}{1-P_K} \left(\frac{K\rho^{K+1}\left(\frac{1}{\rho}-1\right)+\rho^K-1}{(\mu-\lambda)^2} \right) \\ &= \frac{P_0\rho}{\lambda(1-P_K)(1-\rho)^2} \left(K\rho^{K+1}\left(\frac{1}{\rho}-1\right)+\rho^K-1 \right) \\ &= \frac{1}{\lambda(1-P_K)} \cdot \frac{\rho P_0(K\rho^K - K\rho^{K+1}+\rho^K-1)}{(1-\rho)^2} \\ &= \frac{1}{\lambda(1-P_K)} \frac{\rho P_0((K+1)\rho^K - K\rho^{K+1}-1)}{(1-\rho)^2} \\ &= -\frac{\overline{N}}{\lambda(1-P_K)}, \end{aligned}$$

that is

$$\overline{T} = \frac{\overline{N}}{\lambda(1-P_K)},$$

which was obtained earlier.
The higher moments can be calculated, too.
■

Exercise 8 *Consider a closed queueing network with 2 nodes containing K customers. Assume that at each node the service times are exponentially distributed with parameter μ_1 and μ_2, respectively. Find the mean performance measures at each node.*

Solution:
It is easy to see that the nodes operate the same way and they can be considered as an $M/M/1/K$ queueing system. Hence the performance measures can be computed by using the formulas with $\rho_2 = \frac{\mu_1}{\mu_2}$ and $\rho_1 = \frac{\mu_2}{\mu_1}$, respectively.
Furthermore, one can easily verify that

$$U_S(1)\mu_1 = U_S(2)\mu_2,$$

where $U_S(i)$, $i = 1, 2$ is the utilization of the server.

■

Exercise 9 *Find the generating function for an $M/M/n/n$ queueing system.*

Solution:

$$G_N(s) = \sum_{k=0}^{n} s^k \frac{\varrho^k}{k!} P_0 = \sum_{k=0}^{n} \frac{(s\varrho)^k}{k!} e^{-s\varrho} P_0 e^{s\varrho} = e^{-\varrho(1-s)} \frac{Q(n, s\varrho)}{Q(n, \varrho)}.$$

To verify the formula let us calculate $\overline{T}$.
Since $\overline{N} = G'_N(1)$, therefore take the derivative, that is we get

$$G'_N(s) = \varrho e^{-\varrho(1-s)} \frac{Q(n, \varrho s)}{Q(n, \varrho)} - e^{-\varrho(1-s)} \frac{\varrho P(n, \varrho s)}{Q(n, \varrho)}.$$

hence

$$G'_N(1) = \varrho - \varrho B(n, \varrho) = \varrho(1 - B(n, \varrho)),$$

which was obtained earlier.

■

Exercise 10 *Find $Var(N)$ for an $M/M/n/n$ queueing system.*

Solution:
Since $Var(N) = \mathbb{E}(N^2) - (\mathbb{E}(N))^2$, let us calculate first $\mathbb{E}(N^2)$. That is

$$\begin{aligned}
\mathbb{E}(N^2) &= \sum_{k=1}^{n} k^2 P_k = \sum_{k=1}^{n} (k(k-1) + k) P_k = \sum_{k=1}^{n} k(k-1) P_k + \sum_{k=1}^{n} k P_k \\
&= \sum_{k=2}^{n} k(k-1) \frac{\varrho^k}{k!} P_0 + \mathbb{E}(N) = \sum_{i=0}^{n-2} \varrho^2 \frac{\varrho^i}{i!} P_0 + \mathbb{E}(N) \\
&= \varrho^2 (1 - P_n - P_{n-1}) + \mathbb{E}(N) = \varrho^2 \left(1 - P_n \left(1 + \frac{n}{\varrho}\right)\right) + \mathbb{E}(N).
\end{aligned}$$

Since $\mathbb{E}(N) = \varrho(1 - B(n, \varrho))$, therefore

$$\begin{aligned}
Var(N) &= \varrho^2 (1 - B(n, \varrho)(1 + \frac{n}{\varrho})) - (\varrho(1 - B(n, \varrho)))^2 + \mathbb{E}(N) \\
&= \varrho^2 (1 - B(n, \varrho)(\frac{\varrho + n}{\varrho})) - (\varrho(1 - B(n, \varrho)))^2 + \mathbb{E}(N) \\
&= \varrho^2 - \varrho^2 B(n, \varrho) - n\varrho B(n, \varrho) - \varrho^2 - \varrho^2 B^2(n, \varrho) + 2\varrho^2 B(n, \varrho) + \mathbb{E}(N) \\
&= \mathbb{E}(N) + \varrho^2 B(n, \varrho) - n\varrho B(n, \varrho) - \varrho^2 B^2(n, \varrho) \\
&= \mathbb{E}(N) - \varrho B(n, \varrho)(n - \varrho(1 - B(n, \varrho))) \\
&= \mathbb{E}(N) - \varrho B(n, \varrho)(n - \mathbb{E}(N)).
\end{aligned}$$

■

Exercise 11 *Show that $B(m,a)$ is a monotone decreasing sequence and its limit is 0.*

Solution:

$$B(m,a) = \frac{aB(m-1,a)}{m + aB(m-1,a)} < \frac{a}{m},$$

and thus it tends to 0 as m increasing. The sequence is monotone decreasing iff

$$B(m,a) - B(m-1,a) < 0, \forall m$$

that is

$$\begin{aligned} \frac{aB(m-1,a)}{m + aB(m-1,a)} - B(m-1,a) &< 0 \\ \frac{B(m-1,a)(a - m - aB(m-1,a))}{m + aB(m-1,a)} &< 0 \\ a - m - aB(m-1,a) &< 0 \\ B(m-1,a) &> \frac{a-m}{a}, \end{aligned}$$

which is satisfied if $a \leq m$. Since $1 \geq B(m-1,a) \geq 0$ therefore $1 \geq \frac{a-m}{a} \geq 0$, and thus $a \geq m$,$m \geq 0$, that is $0 \leq m \leq a$. It means that $B(m,a)$ is monotone decreasing for m which was expected since as the number of servers increases the probability of loss should decrease.

■

Exercise 12 *Find a recursion for $C(m,a)$.*

Solution:
Let $a = \frac{\lambda}{\mu}$, then by the help of

$$C(m,a) = \frac{B(m,a)}{1 - \frac{a}{m}(1 - B(m,a))}$$

we should write a recursion for $C(m,a)$ since $B(m,a)$ can be obtained recursively. First we show how $B(m-1,a)$ can be expressed by the help of $C(m-1,a)$ and then substituting into the recursion

$$B(m,a) = \frac{aB(m-1,a)}{m + aB(m-1,a)}$$

we get the desired formula. So let us express $B(m,a)$ via $C(m,a)$ that is

$$C(m,a) = \frac{mB(m,a)}{m - a(1 - B(m,a))}$$

$$C(m,a)(m-a) + C(m,a)aB(m,a) = mB(m,a)$$

thus

$$B(m,a) = \frac{(m-a)C(m,a)}{m - aC(m,a)},$$

which is positive since $m > a$ is the stability condition for an $M/M/m$ queueing system. This shows that

$$B(m,a) < C(m,a),$$

which was expected because of the nature of the problem.
Consequently

$$B(m-1,a) = \frac{(m-1-a)C(m-1,a)}{m-1-aC(m-1,a)},$$

and $m-1 > a$ is also valid due to the stability condition. Let us first express $C(m,a)$ by the help of $B(m-1,a)$ then substitute. To do so

$$\begin{aligned} C(m,a) &= \frac{\frac{aB(m-1,a)}{m+aB(m-1,a)}m}{m-a\big(1-\frac{aB(m-1,a)}{m+aB(m-1,a)}\big)} = \frac{\frac{amB(m-1,a)}{m+aB(m-1,a)}}{m-a\big(\frac{m+aB(m-1,a)-aB(m-1,a)}{m+aB(m-1,a)}\big)} \\ &= \frac{aB(m-1,a)}{m+aB(m-1,a)-a} = \frac{aB(m-1,a)}{m-a(1-B(m-1,a))}. \end{aligned}$$

Now let us substitute $C(m-1,a)$ into here. Let us express the numerator and denominator in a simpler form, namely

$$NUM = a\frac{(m-1-a)C(m-1,a)}{m-1-aC(m-1,a)}$$

$$\begin{aligned} DENOM &= m-a\left(1-\frac{(m-1-a)C(m-1,a)}{m-1-aC(m-1,a)}\right) \\ &= m-a\frac{m-1-aC(m-1,a)-(m-1)C(m-1,a)+aC(m-1,a)}{m-1-aC(m-1,a)} \\ &= m-a\frac{(m-1)(1-C(m-1,a))}{m-1-aC(m-1,a)} \\ &= \frac{m(m-1)-maC(m-1,a)-a(m-1)(1-C(m-1,a))}{m-1-aC(m-1,a)} \\ &= \frac{m(m-1)-a(m-1)-aC(m-1,a)}{m-1-aC(m-1,a)} \\ &= \frac{(m-1)(m-a)-aC(m-1,a)}{m-1-aC(m-1,a)}. \end{aligned}$$

Thus

$$C(m,a) = \frac{a(m-1-a)C(m-1,a)}{(m-1)(m-a)-aC(m-1,a)},$$

and the initial value is $C(1) = a$. Thus the probability of waiting can be computed recursively. It is important because the main performance measures depends on this value.

Now, let us show that $C(m,a)$ is a monotone decreasing sequence and tends to 0 as m, increases which is expected. It is not difficult to see that

$$C(m,a) < \frac{a(m-1-a)C(m-1,a)}{(m-1)(m-a)-a}$$

and if we show that

$$\frac{a(m-1-a)}{(m-1)(m-a)a} < 1,$$

then we have

$$C(m,a) < C(m-1,a).$$

To do so it is easy to see that

$$\begin{aligned} a(m-1-a) &< (m-1)(m-a)-a \\ m^2-m-ma+a-a-am+a+a^2 &> 0 \\ m^2-(1+2a)m+a+a^2 &> 0 \\ m_{1,2} &= \frac{1+2a \pm \sqrt{(1+2a)^2-4(a^2+a)}}{2} \\ m_{1,2} &= \frac{1+2a \pm \sqrt{1+4a^2+4a-4a^2-4a}}{2} \\ m_{1,2} &= \frac{1+2a \pm 1}{2} \end{aligned}$$

that is if $m > a+1$ then the values of the parabola are positive. However, this condition is satisfied since the stability condition is $m-1 > a$.
Furthermore, since

$$C(m,a) = \frac{B(m,a)}{1-\frac{a}{m}(1-B(m,a))}$$

therefore $\lim_{m\to\infty} C(m,a) = 0$, which was expected.
This can be proved by direct calculations, since

$$C(m,\varrho) = \frac{\varrho^m}{m!}\frac{m}{m-\varrho}P_0(m)$$

and from

$$\lim_{m\to\infty} P_0(m) = e^{-\varrho}, \qquad \lim_{m\to\infty} \frac{\varrho^m}{m!}\frac{m}{m-\varrho} = 0,$$

the limit is 0. It is clear because there is no waiting in an infinite-server system.
■

Exercise 13 *Verify that the distribution function of the response time for a $M/M/r$ queueing system in the case of $r = 1$ reduces to the formula obtained for an $M/M/1$ system.*

Solution:

$$\begin{aligned} P(T>x) &= e^{-\mu x}\left(1 + C(n,\varrho)\frac{1-e^{-\mu(r-1-\varrho)x}}{r-1-\varrho}\right) = e^{-\mu x}\left(1+\varrho\frac{1-e^{\mu\varrho x}}{-\varrho}\right) \\ &= e^{-\mu x}(1-1+e^{\mu\varrho x}) = e^{-\mu(1-\varrho)x}. \end{aligned}$$

Thus

$$F_T(x) = 1 - e^{-\mu(1-\varrho)x}.$$

■

Exercise 14 *Show that* $\lim_{z\to 1} G_N(z) = 1$ *for an* $M/G/1$ *queueing system.*

Solution:

$$\lim_{z\to 1} G_N(z) = \lim_{z\to 1} (1-\rho)L_S(\lambda-\lambda z)\cdot\frac{z-1}{z-L_S(\lambda-\lambda z)} = \frac{0}{0},$$

therefor the L'Hospital's rule is applied. It is easy to see that

$$\lim_{z\to 1}\frac{z-1}{z-L_S(\lambda-\lambda z)} = \lim_{z\to 1}\frac{1}{1+\lambda L_S'(\lambda-\lambda z)} = \frac{1}{1-\rho},$$

and thus

$$\lim_{z\to 1} G_N(z) = \lim_{z\to 1} (1-\rho)L_S(\lambda-\lambda z)\cdot\lim_{z\to\infty}\frac{z-1}{z-L_S(\lambda-\lambda z)} = \frac{1-\rho}{1-\rho} = 1.$$

■

Exercise 15 *Show that if the residual service time in an* $M/G/1$ *queueing system is denoted by* R *then its Laplace-transform can be obtained as* $L_R(t) = \frac{1-L_S(t)}{t\cdot\mathbb{E}(S)}$.

Solution:

$$L_R(t) = \int_0^\infty e^{-tx}\frac{1-F_S(x)}{\mathbb{E}(S)}\mathrm{d}x.$$

Using integration by parts we have

$$L_R(t) = \left[-\frac{e^{-tx}}{t}\frac{1-F_S(x)}{\mathbb{E}(S)}\right]_0^\infty + \int_0^\infty \frac{e^{-tx}}{t}\frac{(-f(x))}{\mathbb{E}(S)}\mathrm{d}x = \frac{1-L_S(t)}{t\mathbb{E}(S)}.$$

Verify the limit $\lim_{t\to 0} L_R(t) = 1$.
It is easy to see that

$$\lim_{t\to 0} L_R(t) = \lim_{t\to\infty}\frac{1-L_S(t)}{t\mathbb{E}(S)} = \frac{0}{0},$$

therefore apply the L'Hospital's rule. Thus

$$\lim_{t\to 0} L_R(t) = \lim_{t\to 0}\frac{-L_S'(t)}{\mathbb{E}(S)} = \frac{\mathbb{E}(S)}{\mathbb{E}(S)} = 1.$$

■

Exercise 16 *By the help of $L_R(t)$ prove that if $S \in Exp(\mu)$, then $R \in Exp(\mu)$.*

Solution:

$$L_R(t) = \frac{1 - L_S(t)}{t\mathbb{E}(S)} = \frac{1 - \frac{\mu}{\mu+t}}{\frac{t}{\mu}} = \frac{\mu}{\mu + t},$$

thus $R \in Exp(\mu)$.

■

Exercise 17 *By the help of the formulas for an $M/G/1$ system derive the corresponding formulas for an $M/M/1$ system.*

Solution:

In this case

$$L_S(t) = \frac{\mu}{\mu + t},$$

therefore the Laplace-transform of the response time is

$$\begin{aligned} L_T(t) &= L_S(t)\frac{t(1-\rho)}{t - \lambda + \lambda L_S(t)} \\ &= \frac{\mu}{\mu + t}\frac{t(1-\rho)}{t - \lambda + \frac{\lambda\mu}{\mu+t}} \\ &= \frac{\mu}{\mu + t}\frac{t(1-\rho)(\mu+t)}{\mu t + t^2 - \lambda\mu - \lambda t + \lambda\mu} \\ &= \frac{t(\mu-\lambda)}{t(t+\mu-\lambda)} = \frac{\mu-\lambda}{t+\mu-\lambda} = \frac{\mu(1-\rho)}{\mu(1-\rho)+t}, \end{aligned}$$

that is $T \in Exp(\mu(1-\rho))$, as we have seen earlier.

For the generating function of the number of customers in the system we have

$$\begin{aligned} G_N(z) &= L_S(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{L_S(\lambda - \lambda z) - z} \\ &= \frac{\mu}{\lambda - \lambda z + \mu} \cdot \frac{(1-\rho)(1-z)}{\frac{\mu}{\lambda - \lambda z + \mu} - z} \\ &= \frac{\mu}{\lambda - \lambda z + \mu} \cdot \frac{(1-\rho)(1-z)(\lambda - \lambda z + \mu)}{\mu - \lambda z + \lambda z^2 - \mu z} \\ &= \frac{\mu(1-\rho)(1-z)}{\mu(1-z) - \lambda z(1-z)} = \frac{\mu(1-\rho)}{\mu - \lambda z} = \frac{1-\rho}{1-\rho z}, \end{aligned}$$

as we proved in the case of an $M/M/1$ system.

For the mean waiting and response times we get

$$\overline{W} = \frac{\rho\mathbb{E}(S)}{1-\rho} \cdot \frac{1+C_S^2}{2} = \frac{\rho}{\mu(1-\rho)} \cdot \frac{1+1}{2} = \frac{\rho}{\mu(1-\rho)},$$
$$\overline{T} = \overline{W} + \frac{1}{\mu} = \frac{1}{\mu}\left(\frac{\rho}{1-\rho}+1\right) = \frac{1}{\mu(1-\rho)}.$$

To calculate the variance we need

$$\begin{aligned}\mathbb{E}(W^2) &= 2(\overline{W})^2 + \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)} = 2\left(\frac{\rho}{\mu(1-\rho)}\right)^2 + \frac{\lambda\frac{3!}{\mu^3}}{3(1-\rho)} \\ &= 2\frac{\rho^2}{\mu^2(1-\rho)^2} + \frac{2\lambda}{\mu^3(1-\rho)} = \frac{2\mu\rho^2 + 2\lambda(1-\rho)}{\mu^3(1-\rho)^2} \\ &= \frac{2\lambda\rho + 2\lambda - 2\lambda\rho}{\mu^3(1-\rho)^2} = \frac{2\lambda}{\mu^3(1-\rho)^2},\end{aligned}$$

thus

$$\begin{aligned}Var(W) &= \frac{2\lambda}{\mu^3(1-\rho)^2} - \left(\frac{\rho}{\mu(1-\rho)}\right)^2 \\ &= \frac{2\lambda - \mu\rho^2}{\mu^3(1-\rho)^2} = \frac{2\lambda - \lambda\rho}{\mu^3(1-\rho)^2} = \frac{(2-\rho)\rho}{\mu^2(1-\rho)^2},\end{aligned}$$

as we have seen earlier.

Furthermore

$$\begin{aligned}Var(T) &= Var(W) + Var(S) = \frac{(2-\rho)\rho}{\mu^2(1-\rho)^2} + \left(\frac{1}{\mu}\right)^2 \\ &= \frac{2\rho - \rho^2 + 1 - 2\rho + \rho^2}{\mu^2(1-\rho)^2} = \frac{1}{\mu^2(1-\rho)^2} = \left(\frac{1}{\mu(1-\rho)}\right)^2.\end{aligned}$$

The variance of the number of customers in the system is

$$\begin{aligned}
Var(N) &= \frac{\lambda^3\mathbb{E}(S^3)}{3(1-\rho)} + \left(\frac{\lambda^2\mathbb{E}(S^2)}{2(1-\rho)}\right)^2 + \frac{\lambda^2(3-2\rho)\mathbb{E}(S^2)}{2(1-\rho)} + \rho(1-\rho) \\
&= \frac{\lambda^3\frac{3!}{\mu^3}}{3(1-\rho)} + \left(\frac{\lambda^2\frac{2}{\mu^2}}{2(1-\rho)}\right)^2 + \frac{\lambda^2(3-2\rho)\frac{2}{\mu^2}}{2(1-\rho)} + \rho(1-\rho) \\
&= \frac{2\lambda^3}{\mu^3(1-\rho)} + \left(\frac{\rho^2}{1-\rho}\right)^2 + \frac{\rho^2(3-2\rho)}{1-\rho} + \rho(1-\rho) \\
&= \frac{2\rho^3}{1-\rho} + \frac{\rho^4}{(1-\rho)^2} + \frac{\rho^2(3-2\rho)}{1-\rho} + \rho(1-\rho) \\
&= \frac{2\rho^3(1-\rho) + \rho^4 + \rho^2(1-\rho)(3-2\rho)}{(1-\rho)^2} + \frac{\rho(1-\rho)^3}{(1-\rho)^2} \\
&= \frac{2\rho^3 - 2\rho^4 + \rho^4 + 3\rho^2 - 2\rho^3 - 3\rho^3 + 2\rho^4 + \rho + 3\rho^3 - 3\rho^2 - \rho^4}{(1-\rho)^2} \\
&= \frac{\rho}{(1-\rho)^2},
\end{aligned}$$

as we have seen earlier.

Finally

$$\begin{aligned}
Var(Q) &= \frac{\lambda^3\mathbb{E}(S^3)}{3(1-\rho)} + \left(\frac{\lambda^2\mathbb{E}(S^2)}{2(1-\rho)}\right)^2 + \frac{\lambda^2\mathbb{E}(S^2)}{2(1-\rho)} \\
&= \frac{\lambda^3\frac{3!}{\mu^3}}{3(1-\rho)} + \left(\frac{\lambda^2\frac{2}{\mu^2}}{2(1-\rho)}\right)^2 + \frac{\lambda^2\frac{2}{\mu^2}}{2(1-\rho)} \\
&= \frac{2\rho^3}{1-\rho} + \frac{\rho^4}{(1-\rho)^2} + \frac{\rho^2}{1-\rho} = \frac{2(1-\rho)\rho^3 + \rho^4 + \rho^2(1-\rho)}{(1-\rho)^2} \\
&= \frac{2\rho^3 - 2\rho^4 + \rho^4 + \rho^2 - \rho^3}{(1-\rho)^2} = \frac{\rho^3 - \rho^4 + \rho^2}{(1-\rho)^2} = \frac{\rho^2(1+\rho-\rho^2)}{(1-\rho)^2}.
\end{aligned}$$

These verifications help us to see if these complicated formulas reduces to the simple ones.

■

Exercise 18 *Based on the transform equation*

$$G_N(z) = L_S(\lambda - \lambda z)(1-\rho)\frac{1-z}{L_S(\lambda-\lambda z) - z}$$

find $\overline{N}$*-t.*

Solution:
It is well-known that $\overline{N} = G'_N(1)$, that is why we have to calculate the derivative at the

right hand side. However, the term $\frac{1-z}{L_S(\lambda-\lambda z)-z}$ takes an indetermined value at $z=1$ hence the L'Hospital's rule is used. Let us first define a function

$$f(z) = \frac{L_S(\lambda - \lambda z) - z}{1 - z}$$

.

Hence one can see that

$$L_N(z) = (1-\rho)\frac{L_S(\lambda - \lambda z)}{f(z)}.$$

Applying the expansion procedure

$$L_S(\lambda - \lambda z) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k \mathbb{E}(S^k)}{k!}(\lambda - \lambda z)^k$$

we have

$$\begin{aligned} f(z) &= \frac{1 + \sum_{k=1}^{\infty} \frac{(-1)^k \mathbb{E}(S^k)}{k!}(\lambda - \lambda z)^k - z}{1 - z} \\ &= 1 - \lambda\mathbb{E}(S) + \lambda^2 \frac{\mathbb{E}(S^2)(1-z)}{2} + \dots. \end{aligned}$$

Thus $f(1) = 1 - \rho$ and $f'(1) = -\frac{\lambda^2\mathbb{E}(S^2)}{2}$.

After these calculations we get

$$L'_N(z) = \frac{(1-\rho)f(z)L'_S(\lambda - \lambda z)(-\lambda) - L_S(\lambda - \lambda z)\cdot f'(z)}{(f(z))^2}$$

and hence

$$\begin{aligned} \overline{N} = G'_N(1) &= \frac{(1-\rho)f(1)\lambda\mathbb{E}(S) + \frac{\lambda^2\mathbb{E}(S^2)}{2}}{(1-\rho)^2} \\ &= \frac{(1-\rho)\left((1-\rho)\rho + \frac{\lambda^2\mathbb{E}(S^2)}{2}\right)}{(1-\rho)^2} \\ &= \rho + \frac{\lambda^2\mathbb{E}(S^2)}{2(1-\rho)} = \rho + \frac{\rho^2}{1-\rho}\cdot\frac{1+C_S^2}{2}, \end{aligned}$$

which was obtained in a different way.

■

Exercise 19 *Find $Var(W)$ by the help of $L_W(s) = \frac{s(1-\rho)}{s-\lambda+\lambda L_S(s)}$.*

Solution:
Let us define a function

$$f(s) = \frac{s - \lambda + \lambda L_S(s)}{s},$$

which is after expansion can be written as

$$\begin{aligned} 1 - \frac{\lambda}{s} + \frac{\lambda}{s} L_S(s) &= 1 - \frac{\lambda}{s} + \frac{\lambda}{s} \cdot \sum_{i=0}^{\infty} (-1)^i \frac{\mathbb{E}(S^i)}{i!} s^i \\ &= 1 - \lambda\mathbb{E}(S) + \frac{\lambda\mathbb{E}(S^2)}{2} s - \frac{\lambda\mathbb{E}(S^3)}{3!} s^2 + \dots . \end{aligned}$$

Therefore

$$\begin{aligned} f'(s) &= \frac{\lambda\mathbb{E}(S^2)}{2} - \frac{2\lambda\mathbb{E}(S^3)}{3!} s + \frac{3\lambda\mathbb{E}(S^4)}{4!} s^2 + \dots, \\ f^{(2)}(s) &= -\frac{2\lambda\mathbb{E}(S^3)}{3!} + \frac{3 \cdot 2\lambda\mathbb{E}(S^4)}{4!} s + \dots . \end{aligned}$$

Hence

$$\begin{aligned} f(0) &= 1 - \rho, \\ f'(0) &= \frac{\lambda\mathbb{E}(S^2)}{2}, \\ f''(0) &= -\frac{\lambda\mathbb{E}(S^3)}{3}. \end{aligned}$$

Consequently, because

$$L_W(s) = \frac{1 - \rho}{f(s)}$$

we have

$$\begin{aligned} L'_W(s) &= -(1-\rho)\frac{f'(s)}{(f(s))^2}, \\ L''_W(s) &= -(1-\rho)\frac{f''(s)(f(s))^2 - 2f(s)(f'(s))^2}{(f(s))^4}. \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}(W) = -L'_W(0) &= (1-\rho)\frac{f'(0)}{(f(0))^2} = \frac{(1-\rho)\frac{\lambda\mathbb{E}(S^2)}{2}}{(1-\rho)^2} \\ &= \frac{\lambda\mathbb{E}(S^2)}{2(1-\rho)} = \frac{\rho\mathbb{E}(S)}{1-\rho} \cdot \frac{1 + C_S^2}{2}. \end{aligned}$$

Similarly

$$\begin{aligned}\mathbb{E}(W^2) = L''_W(0) &= -(1-\rho)\frac{f''(0)(f(0))^2 - 2f(0)(f'(0))^2}{(f(0))^4} \\ &= -\frac{(1-\rho)(1-\rho)^2\left(-\frac{\lambda\mathbb{E}(S^3)}{3}\right) - 2(1-\rho)\left(\frac{\lambda\mathbb{E}(S^2)}{2}\right)^2}{(1-\rho)^4} \\ &= 2(\mathbb{E}(W))^2 + \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)}.\end{aligned}$$

Thus

$$\begin{aligned}Var(W) &= \mathbb{E}(W^2) - (\mathbb{E}(W))^2 \\ &= 2(\mathbb{E}(W))^2 + \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)} - (\mathbb{E}(W))^2 \\ &= (\mathbb{E}(W))^2 + \frac{\lambda\mathbb{E}(S^3)}{3(1-\rho)}.\end{aligned}$$

Finally

$$Var(T) = Var(W+S) = Var(W) + Var(S).$$

■

Exercise 20 *By using the Laplace-transform show that*

$$\mathbb{E}(R^k) = \frac{\mathbb{E}(S^{k+1})}{(k+1)\mathbb{E}(S)}.$$

Solution:
As we have seen earlier

$$L_R(s) = \frac{1 - L_S(s)}{s\mathbb{E}(S)},$$

and it is well-known that

$$\begin{aligned}L_S(s) &= \sum_{i=0}^{\infty} \frac{L_S^{(i)}(0)}{i!} s^i \\ &= \sum_{i=0}^{\infty} \frac{(-1)^i \mathbb{E}(S^i)}{i!} s^i.\end{aligned}$$

Thus for $L_R(s)$ we get

$$L_R(s) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}\mathbb{E}(R^k)s^k.$$

Therefore

$$L_R(s) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}\mathbb{E}(R^k)s^k = \frac{1 - L_S(s)}{s\mathbb{E}(S)}$$
$$= \frac{1 - \left(1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}\mathbb{E}(S^k)s^k\right)}{s\mathbb{E}(S)}$$
$$= \sum_{k=1}^{\infty} \frac{(-1)^k\mathbb{E}(S^{k+1})}{(k+1)!\mathbb{E}(S)} \cdot s^k = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}\frac{\mathbb{E}(Sk+1)}{(k+1)\mathbb{E}(S)}s^k.$$

Consequently

$$\mathbb{E}(R^k) = \frac{\mathbb{E}(S^{k+1})}{(k+1)\mathbb{E}(S)}, \qquad k = 1, 2, \ldots$$

■

Exercise 21 *Find the generating function of the number of customers arrived during a service time for an $M/G/1$ system.*

Solution:
By applying the theorem of total probability we have

$$P(\nu_A(S) = k) = \int_0^{\infty} \frac{(\lambda x)^k}{k!}e^{-\lambda x} f_S(x)\, dx.$$

Hence its generating function can be written as

$$G_{\nu_A(S)}(z) = z^k \int_0^{\infty} \frac{(\lambda x)^k}{k!}e^{-\lambda x} f_S(x)\, dx = \int_0^{\infty} \sum_{k=0}^{\infty} \frac{(z\lambda x)^k}{k!}e^{-\lambda x} f_S(x)\, dx$$
$$= \int_0^{\infty} e^{z\lambda x}e^{-\lambda x} f_S(x)\, dx = \int_0^{\infty} e^{-\lambda x(1-z)} f_S(x)\, dx = L_S(\lambda(1-z)).$$

■

4.2 Finite-Source Systems

Exercise 22 *If* $P(k,\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$ *and* $Q(k,\lambda) = \sum_{i=0}^{k}\frac{\lambda^i}{i!}e^{-\lambda}$, *then show the following important formula*

$$\sum_{j=0}^{k} P(k-j,a_1)Q(j,a_2) = Q(k,a_1+a_2).$$

Solution:
It is well-known that

$$Q(n,a) = \int_a^\infty P(n,y)\,dy,$$

therefore

$$\sum_{j=0}^{k} P(k-j,a_1)Q(j,a_2) = \sum_{j=0}^{k}\frac{a_1^{k-j}}{(k-j)!}e^{-a_1}\sum_{i=0}^{j}\frac{a_2^i}{i!}e^{-a_2}$$

$$\sum_{j=0}^{k}\frac{a_1^{k-j}}{(k-j)!}e^{-a_1}\int_{a_2}^\infty\frac{y^j}{j!}e^{-y}\,dy = \int_{a_2}^\infty\frac{(y+a_1)^k}{k!}e^{-(a_1+y)}\,dy = \int_{a_1+a_2}^\infty\frac{t^k}{k!}e^{-t}\,dt = Q(k,a_1+a_2),$$

where we introduced the substitution $t = y + a_1$.
■

Exercise 23 *Find the mean response time for an* $M/M/1/n/n$ *queueing system by using the Laplace-transform.*

Solution:

It is well-known that $\overline{T} = -L_T'(0)$, that is why let us calculate $L_T'(0)$.

$$\begin{aligned} L_T'(s) &= \left[\left(\frac{\mu}{\mu+s}\right)^n e^{\frac{s}{\lambda}}\frac{Q\left(n-1,\frac{\mu+s}{\lambda}\right)}{Q\left(n-1,\frac{\mu}{\lambda}\right)}\right]' \\ &= \left(\left(\frac{\mu}{\mu+s}\right)^n e^{\frac{s}{\lambda}}\right)'\cdot\frac{Q\left(n-1,\frac{\mu+s}{\lambda}\right)}{Q\left(n-1,\frac{\mu}{\lambda}\right)} + \left(\frac{\mu}{\mu+s}\right)^n e^{\frac{s}{\lambda}}\cdot\frac{\left(Q\left(n-1,\frac{\mu+s}{\lambda}\right)\right)'}{Q\left(n-1,\frac{\mu}{\lambda}\right)}. \end{aligned}$$

Thus

$$L_T'(o) = -\frac{n}{\mu} + \frac{1}{\lambda} - \frac{1}{\lambda}B\left(n-1,\frac{\mu}{\lambda}\right) = -\frac{n}{\mu} + \frac{1}{\lambda}U_S(n-1),$$

and hence

$$\overline{T}(n) = \frac{n}{\mu} - \frac{U_S(n-1)}{\lambda}.$$

Since

$$\begin{aligned}\overline{T}(n) &= \frac{1}{\mu}(\overline{N}(n-1)+1)\\ &= \frac{1}{\mu}\left(n-1-\frac{U_S(n-1)}{\rho}+1\right)\\ &= \frac{n}{\mu}-\frac{U_S(n-1)}{\lambda},\end{aligned}$$

which was obtained earlier. The higher moments of $T(n)$ can be obtained and hence further measured can be calculated.
Similarly, the moments of $W(n)$ can be derived.
■

Exercise 24 *Find the mean response time, denoted by $\overline{T}A$, for an $M/M/1/n/n$ system by using the density function approach.*

Solution:

$$\begin{aligned}z &\doteq \frac{\mu}{\lambda}\\ \overline{T} &= \frac{1}{Q(n-1,z)}\int_0^\infty x\mu\frac{(\mu x+z)^{n-1}}{(n-1)!}e^{-(\mu x+z)}\mathrm{d}x\\ &= \frac{e^{-z}}{Q(n-1,z)}\int_0^\infty x\mu\sum_{k=0}^{n-1}\frac{(\mu x)^k}{k!}\frac{z^{n-1-k}}{(n-1-k)!}e^{-\mu x}\mathrm{d}x\\ &= \frac{e^{-z}}{Q(n-1,z)}\sum_{k=0}^{n-1}\frac{z^{n-1-k}}{(n-1-k)!}\int_0^\infty x\mu\frac{(\mu x)^k}{k!}e^{-\mu x}\mathrm{d}x\\ &= \frac{e^{-z}}{Q(n-1,z)}\sum_{k=0}^{n-1}\frac{z^{n-1-k}}{(n-1-k)!}\cdot\frac{k+1}{\mu}\\ &= \frac{1}{\mu}\sum_{k=0}^{n-1}(k+1)\frac{\frac{z^{n-1-k}}{(n-1-k)!}\cdot e^{-z}}{Q(n-1,z)}\\ &= \frac{1}{\mu}\sum_{k=0}^{n-1}\frac{k+1}{\mu}P_k(n-1) = \frac{1}{\mu}(\overline{N}(n-1)+1).\end{aligned}$$

The mean waiting time can be obtained similarly, starting the summation from 1 and taking a Erlang distribution with one phase less.
■

Exercise 25 *Find $Var(N)$ for an $M/M/1/n/n$ system.*

Solution:

Let us denote by F the number of customers in the source. Hence $F + N = n$, and thus $Var(N) = Var(F)$.
As we have proved the distribution of F equals to the distribution of an $M/M/n/n$ system with traffic intensity $z = \frac{1}{\rho}$ we have

$$\begin{aligned} Var(N) &= \frac{1}{\rho}\left(1 - B\left(n, \frac{1}{\rho}\right)\right) - \frac{1}{\rho}B\left(n, \frac{1}{\rho}\right)\left(n - \frac{1}{\rho}\left(1 - B\left(n, \frac{1}{\rho}\right)\right)\right) \\ &= \frac{U_S}{\rho} - \frac{1 - U_S}{\rho}\left(n - \frac{U_S}{\rho}\right). \end{aligned}$$

If the number of sources is denoted then this formula can be written as

$$Var(N(n)) = \frac{U_s(n)}{\rho} - \frac{1 - U_s(n)}{\rho}\left(n - \frac{U_s(n)}{\rho}\right).$$

This result helps us to determine $Var(T(n))$-t and $Var(W(n))$-t. Since $W(n)$ can be consider as a random sum, where the summands are the exponentially distributed service times with parameter μ, and the counting process is the number of customers in the system at the arrival instant of a customer, denoted by $N_A^{(n)}$, we can use the formula obtained for the variance of a random sum, namely

$$\begin{aligned} Var(W(n)) &= \mathbb{E}(N_A^{(n)}) \cdot \frac{1}{\mu^2} + \frac{1}{\mu^2} Var(N_A^{(n)}) \\ &= \overline{N}(n-1) \cdot \frac{1}{\mu^2} + \frac{1}{\mu^2} Var(N(n-1)) \\ &= \frac{1}{\mu^2}(\overline{N}(n-1) + Var(N(n-1))), \end{aligned}$$

where

$$\begin{aligned} \overline{N}(n-1) &= n - 1 - \frac{U_s(n-1)}{\rho} \\ Var(N(n-1)) &= \frac{U_s(n-1)}{\rho} - \frac{1 - U_s(n-1)}{\rho}\left(n - 1 - \frac{U_s(n-1)}{\rho}\right). \end{aligned}$$

Similarly, since $T(n) = W(n) + S$, then

$$Var(T(n)) = Var(W(n)) + \frac{1}{\mu^2}.$$

■

Chapter 5

Queueing Theory Formulas

5.1 Notations and Definitions

Table 1. Basic Queueing Theory Notations and Definitions

a	Server utilization.
a_i	Utilization of component i in a queueing network.
$A[t]$	Distribution function of interarrival time. $A[t] = P[\tau \leq t]$
b	Random variable describing the busy period for a server
$B[c, \rho]$	Erlang's B formula. Probability all servers busy in M/M/c/c system. Also called Erlang's loss formula.
c	Number of servers in in a service facility.
$C[c, \rho]$	Erlang's C formula. Probability all servers busy in M/M/c/c system. Also called Erlang's delay formula
C_X^2	Squared coefficient of a variation of a positive random variable, $C_X^2 = \frac{Var[X]}{\mathbb{E}[X]^2}$.
D	Symbol for constant (deterministic) interarrival or service time.
E_k	Symbol for Erlang-k distribution of interarrival or service time.
$\mathbb{E}[Q\|Q > 0]$	Expected (mean or average) queue length of nonempty queues.
$\mathbb{E}[W\|W > 0]$	Expected queueing time.
FCFS	First Come First Served queue discipline.
FIFO	First In First Out queue discipline. Identical with FCFS.
$F_T(t)$	The distribution function of T, $F_T(t) = P[T < t]$.
$F_W(t)$	The distribution function of W, $F_W(t) = P[W < t]$.
G	Symbol for general probability distribution of service time. Independence usually assumed.
GI	Symbol for general independent interarrival time distribution.
H_2	Symbol for two-stage hyperexponential distribution. Can be generalized to k stages.
K	Maximum number of customers allowed in queueing system. Also size of population in finite population models.

Table 1. Basic Queueing Theory Notations and Definitions (continued)

λ	Mean arrival rate of customers into the system.
$\overline{\lambda}$	Actual mean arrival rate into the system, for which some arrivals arc turned away, e.q., the M/M/c/c system.
λ_T	Mean throughput of a computer system measured in transactions or interactions per unit time.
$\ln \cdot)$	Natural logarithm function (log to base e).
$\overline{N}_s$	Expected steady state number of customers receiving service, $\mathbb{E}[N_s]$.
LCFS	Last Come First Served queue discipline.
LIFO	Last In First Out queue discipline. Identical with LCFS.
M	Symbol for exponential interarrival or service time.
μ	Mean service rate per server, that is, the mean rate of service completions while the server is busy.
μ_a, μ_b	Parameters of the two-stage hyperexponential distribution of it? for the $M/H_2/1$ system.
$\overline{N}$	Expected steady state number of customers in the queueing system, $\mathbb{E}[N]$.
$N[t]$	Random variable describing the number of customers in the system at time t.
N	Random variable describing the steady state number of customers in the system.
N_b	Random variable describing the number of customers served by a server in one busy period.
$N_s[t]$	Random variable describing the number of customers receiving service at time t.
N_s	Random variable describing the steady state number of customers in the service facility.
O	Operating time of a machine in a machine repair queueing model. The time a machine remains in operation afler repair before repair is again necessary.
$P_n[t]$	Probability there arc n customers in the system at time t.
P_n	Steady state probability that there are n customer in the system.
PRI	Symbol for priority queueing discipline.
PS	Symbol for processor sharing queueing discipline.
p_i	A parameter of a hypoexponential random variable.
π_a, π_b	Parameters of the distribution function of w for the $M/H_2/1$ queueing system.
$\pi_X[r]$	The rth percentile for random variable X.
Q	Random variable describing the steady state number of customers in the queue.

Table 1. Basic Queueing Theory Notations and Definitions (continued)

$Q[t]$	Random variable describing the number of customers in the queue at time t.
ρ	$\rho = \lambda\overline{S}$ The traffic intensity or offered load. The international unit of this is erlang, named for A.K. Erlang, a queueing theory pioneer
RSS	Symbol for queueing discipline "random selection for service".
S	Random variable describing the service time. $\mathbb{E}[S] = \frac{1}{\mu}$.
$\overline{S}$	Expected customer service time, $\mathbb{E}[S] = \frac{1}{\mu}$.
SIRO	Symbol for service in random order, which is identical to RSS. It means each customer in queue has the same probability of being served next.
T	Random variable describing the total time a customer spends in the queueing system, $T = W + S$.
$\overline{T}$	Expected steady state time a customer spends in the system, $\overline{T} = \mathbb{E}[T] = \overline{W} + \overline{S}$.
τ	Random variable describing interarrival time. $\mathbb{E}[\tau] = \frac{1}{\lambda}$.
W	Random variable describing the time a customer spends in the queue before service begins.
W'	Random variable describing time a customer who must queue spends in the queue before receiving service. Also called conditional queueing time.
$\overline{W}$	Expected steady state time a customer spends in the queue, $\overline{W} = \mathbb{E}[W] = \overline{T} - \overline{S}$.

5.2 Relationships between random variables

Table 2. Relationships between Random Variables

a	Server utilization. The probability any particular server is busy.
$N = Q + N_s$	Number of customers in steady state system.
$\overline{N} = \overline{\lambda} \cdot \overline{T}$	Mean number of customers in steady state system. This formula is often called Little's law.
$\overline{N_s} = \overline{\lambda} \cdot \overline{S}$	Mean number of customers receiving service in steady state system. This formula sometimes called Little's law.
$\overline{Q} = \overline{\lambda} \cdot \overline{W}$	Mean number in steady state queue. Also called Little's law.
$\rho = \frac{\mathbb{E}[S]}{\mathbb{E}[\tau]} = \lambda\overline{S}$	Traffic intensity in erlangs.
$T = W + S$	Total waiting time in the system.
$\overline{T} = \overline{W} + \overline{S}$	Mean total waiting time in steady state system.

5.3 M/M/1 Formulas

Table 3. M/M/1 Queueing System

$\rho = \lambda \overline{S}, \quad P_n = P[N = n] = (1-\rho)\rho^n, \quad n = 0, 1, \ldots.$

$P[N \geq n] = \rho^n, \quad n = 0, 1, \ldots.$

$\overline{N} = \mathbb{E}[N] = \overline{\lambda} \cdot \overline{T} = \dfrac{\rho}{1-\rho}, \quad Var(N) = \dfrac{\rho}{(1-\rho)^2}.$

$\overline{Q} = \lambda \overline{W} = \dfrac{\rho^2}{1-\rho}, \quad Var(Q) = \dfrac{\rho^2(1+\rho-\rho^2)}{(1-\rho)^2}.$

$\mathbb{E}[Q|Q>0] = \dfrac{1}{1-\rho}, \quad Var[Q|Q>0] = \dfrac{\rho}{(1-\rho)^2}.$

$F_T(t) = P[T \leq t] = 1 - \exp\left(\dfrac{-t}{\overline{T}}\right), \quad P[T > t] = \exp\left(\dfrac{-t}{\overline{T}}\right).$

$\overline{T} = \mathbb{E}[T] = \dfrac{\overline{S}}{1-\rho} = \dfrac{1}{\mu(1-\rho)}, \quad Var(T) = \overline{T}^2.$

$\pi_T[r] = \overline{T} \ln\left(\dfrac{100}{100-r}\right), \quad \pi_T[90] = \overline{T} \ln 10, \quad \pi_T[95] = \overline{T} \ln 20$

$F_T(t) = P[W \leq t] = 1 - \rho \exp\left(\dfrac{-t}{\overline{T}}\right), \quad P[W > t] = \rho \exp\left(\dfrac{-t}{\overline{T}}\right).$

$\overline{W} = \dfrac{\rho \overline{S}}{1-\rho}, \quad Var(W) = \dfrac{(2-\rho)\rho \overline{S}^2}{(1-\rho)^2}.$

$\pi_W[r] = \max\left\{\overline{T} \ln\left(\dfrac{100\rho}{100-r}\right), 0\right\}.$

$\pi_W[90] = \max\{\overline{T} \ln(10\rho), 0\}, \quad \pi_W[95] = \max\{\overline{T} \ln(20\rho), 0\}.$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.4 M/M/1/K Formulas

Table 4. M/M/1/K Queueing System

$$P_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{(1-\rho^{K+1})} & \text{if } \lambda \neq \mu \\ \\ \dfrac{1}{K+1} & \text{if } \lambda = \mu \end{cases}$$

$n = 0, 1, \ldots, K$, where $\rho = \lambda \overline{S}$.

$\overline{\lambda} = (1 - P_K)\lambda$, Mean arrival rate into system.

$$\overline{N} = \begin{cases} \dfrac{\rho[1-(K+1)\rho^K + K\rho^{K+1}]}{(1-\rho)(1-\rho^{K+1})} & \text{ha } \lambda \neq \mu \\ \\ \dfrac{K}{2} & \text{ha } \lambda = \mu. \end{cases}$$

$\overline{Q} = \overline{N} - (1 - P_0)$, $\Pi_n = \dfrac{P_n}{1 - P_K}, n = 0, 1, \ldots, K-1.$

$$F_T(t) = 1 - \sum_{n=0}^{K-1} \Pi_n Q[n; \mu t],$$

where

$$Q[n; \mu t] = e^{-\mu t} \sum_{k=0}^{n} \frac{(\mu t)^k}{k!}.$$

$$\overline{T} = \frac{\overline{N}}{\overline{\lambda}}, \quad \overline{W} = \frac{\overline{Q}}{\overline{\lambda}}.$$

$$F_T(t) = 1 - \sum_{n=0}^{K-2} \Pi_{n+1} Q[n; \mu t].$$

$$\mathbb{E}[W | W > 0] = \frac{\overline{W}}{1 - \Pi_0}, \quad a = (1 - P_K)\rho.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.5 M/M/c Formulas

Table 5. M/M/c Queueing System

$$\rho = \lambda \overline{S}, \quad a = \frac{\rho}{c}$$

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!(1-a)} \right]^{-1} = \frac{c!(1-a)P[N \geq c]}{\rho^c}.$$

$$P_n = \begin{cases} \dfrac{\rho^n}{n!} P_0, & \text{if } n \leq c \\ \dfrac{\rho^n}{c! c^{n-c}} P_0, & \text{if } n \geq c. \end{cases}$$

$$P[N \geq n] = \begin{cases} P_0 \left[\displaystyle\sum_{k=n}^{c-1} \frac{\rho^k}{k!} + \frac{\rho^c}{c!(1-a)} \right] & \text{if } n < c, \\ P_0 \left[\dfrac{a^c a^{n-c}}{c!(1-a)} \right] = P[N \geq c] a^{n-c} & \text{if } n \geq c \end{cases}$$

$$\overline{Q} = \overline{\lambda} \cdot \overline{W} = \frac{\rho P[N \geq c]}{c(1-a)},$$

where

$$P[N \geq c] = C[c, \rho] = \frac{\dfrac{\rho^c}{c!}}{(1 - \dfrac{\rho}{c}) \displaystyle\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}}.$$

$$Var(Q) = \frac{aC[c,\rho][1 + a - aC[c,\rho]]}{(1-a)^2}.$$

$$\overline{N} = \overline{\lambda} \cdot \overline{T} = \overline{Q} + \rho.$$

$$Var(N) = Var(Q) + \rho(1 + P[N \geq c]).$$

$$\overline{W}[0] = 1 - P[N \geq c], \quad F_T(t) = 1 - P[N \geq c] \exp[-c\mu t(1-a)],$$

$$\overline{W} = \frac{P[N \geq c]\overline{S}}{c(1-a)}.$$

Table 5. M/M/c Queueing System (continued)

$$Var(W) = \frac{[2 - C[c,\rho]]C[c,\rho]\overline{S}^2}{c^2(1-a)^2}.$$

$$\pi_W[r] = \max\{0, \frac{\overline{S}}{c(1-a)} \ln\left(\frac{100C[c,\rho]}{100-r}\right)\}.$$

$$\pi_W[90] = \max\{0, \frac{\overline{S}}{c(1-a)} \ln(10C[c,\rho])\}.$$

$$\pi_W[95] = \max\{0, \frac{\overline{S}}{c(1-a)} \ln(20C[c,\rho])\}.$$

$$P[W \le t | W > 0] = 1 - \exp\left(\frac{-ct(1-a)}{\overline{S}}\right), \quad t > 0.$$

$$\mathbb{E}[W|W>0] = \mathbb{E}[W'] = \frac{\overline{S}}{c(1-a)}.$$

$$Var([W|W>0] = \left(\frac{\overline{S}}{c(1-a)}\right)^2.$$

$$F_T(t) = \begin{cases} 1 + C_1 e^{-\mu t} + C_2 e^{-c\mu t(1-a)} & \text{if } \rho \neq c-1 \\ 1 - \{1 + C[c,\rho]\mu t\}e^{-\mu t} & \text{if } \rho = c-1 \end{cases}$$

where

$$C_1 = \frac{P[N \ge c]}{1 - c(1-a)} - 1, \qquad \text{and} \qquad C_2 = \frac{P[N \ge c]}{c(1-a)-1}.$$

$$\overline{T} = \overline{W} + \overline{S}.$$

$$\mathbb{E}[T^2] = \begin{cases} \frac{2P[N \ge c][1 - c^2(1-a)^2]\overline{S}^2}{(\rho + 1 - c)c^2(1-a)^2} + 2\overline{S}^2 & \text{if } \rho \neq c-1 \\ 2\{2P[N \ge c] + 1\}\overline{S}^2 & \text{if } \rho = c-1 \end{cases}$$

$$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$$

$\pi_T[90] \approx \overline{T} + 1.3\mathbb{D}(T)$, $\pi_T[95] \approx \overline{T} + 2\mathbb{D}(T)$ (estimates due to James Martin).

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.6 M/M/2 Formulas

Table 6. M/M/2 Queueing System

$$\rho = \lambda \overline{S}, \quad a = \frac{\rho}{2}$$

$$P_0 = \frac{1-a}{1+a}.$$

$$P_n = 2P_0 a^n, \quad n = 1, 2, 3, \ldots$$

$$P[N \geq n] = \frac{2a^n}{1+a}, \quad n = 1, 2, \ldots$$

$$\overline{Q} = \lambda \overline{W} = \frac{2a^3}{1-a^2},$$

$P[N \geq 2] = C[2, \rho]$ is the probability that ail arriving customer must queue for service. $P[N \geq 2]$ is given by

$$P[N \geq 2] = C[2, \rho] = \frac{2a^2}{1+a}.$$

$$Var(Q) = \frac{2a^3[(1+a)^2 - 2a^3]}{(1-a^2)^2}.$$

$$\overline{N} = \overline{\lambda} \cdot \overline{T} = \overline{Q} + \rho = \frac{2a}{1-a^2}.$$

$$Var(N) = Var(Q) + \frac{2a(1+a+2a^2)}{1+a}.$$

$$\overline{W}[0] = \frac{1+a-2a^2}{1+a}.$$

$$F_T(t) = 1 - \frac{2a^2}{1+a} \exp[-2\mu t(1-a)]$$

$$\overline{W} = \frac{a^2 \overline{S}}{1-a^2}.$$

$$Var(W) = \frac{a^2(1+a-a^2)\overline{S}^2}{(1-a^2)^2}.$$

$$\pi_W[r] = \max\{0, \frac{\overline{S}}{2(1-a)} \ln\left(\frac{200a^2}{(100-r)(1+a)}\right)\}.$$

Table 6. M/M/2 Queueing System (continued)

$$\pi_W[90] = \max\{0, \frac{\overline{S}}{2(1-a)} \ln\left(\frac{20a^2}{1+a}\right)\}.$$
$$\pi_W[95] = \max\{0, \frac{\overline{S}}{2(1-a)} \ln\left(\frac{40a^2}{1+a}\right)\}.$$

$$\overline{W}_{W'} = P[W \le t | W > 0] = 1 - \exp\left(\frac{-2t(1-a)}{\overline{S}}\right), \quad t > 0.$$
$$\mathbb{E}[W|W>0] = \mathbb{E}[W'] = \frac{\overline{S}}{2(1-a)}.$$
$$Var[W|W>0] = \left(\frac{\overline{S}}{2(1-a)}\right)^2.$$

$$F_T(t) = \begin{cases} 1 + \dfrac{1-a}{1-a^2-2a^2} e^{-\mu t} + \dfrac{2a^2}{1-a-2a^2} e^{-2\mu t(1-a)} & \text{where } \rho \neq 1 \\ 1 - \{1 + \dfrac{\mu t}{3}\} e^{-\mu t} & \text{where } \rho = 1 \end{cases}$$
$$\overline{T} = \overline{W} + \overline{S} = \frac{\overline{S}}{1-a^2}.$$

$$\mathbb{E}[T^2] = \begin{cases} \dfrac{a^2[1-4(1-a)^2]\overline{S}^2}{(2a-1)(1-a)(1-a^2)} + 2\overline{S}^2 & \text{if } \rho \neq 1 \\ \dfrac{10}{3}\overline{S}^2 & \text{if } \rho = 1 \end{cases}$$

$$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$$

$$\pi_T[90] \approx \overline{T} + 1.3\mathbb{D}(T), \quad \pi_T[95] \approx \overline{T} + 2\mathbb{D}(T)$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.7 M/M/c/c Formulas

Table 7. M/M/c/c Queueing System (M/M/c loss)

$\rho = \lambda \overline{S}$

$$P_n = \frac{\frac{\rho^n}{n!}}{1 + \rho + \frac{\rho^2}{2!} + \ldots + \frac{\rho^c}{c!}} \quad n = 0, 1, \ldots, c.$$

The probability that all servers are busy, P_c is called Erlang's B formula, $B[c, \rho]$, and thus

$$B[c, \rho] = \frac{\frac{\rho^c}{c!}}{1 + \rho + \frac{\rho^2}{2!} + \ldots + \frac{\rho^c}{c!}}.$$

$\overline{\lambda} = \lambda(1 - B[c, \rho])$ Is the average arrival rate of customers who actually enter the system. Thus, the true server utilization, a, is given by

$$a = \frac{\overline{\lambda}\ \overline{S}}{c}.$$

$$\overline{N} = \overline{\lambda}\ \overline{S}.$$

$$\overline{T} = \frac{\overline{N}}{\overline{\lambda}} = \overline{S}.$$

$$F_T(t) = 1 - \exp\left(\frac{-t}{\overline{S}}\right).$$

All of the formulas except the last one arc true for the M/G/c/c queueing system. For this system we have

$F_T(t) = F_S(t)$.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.8 M/M/c/K Formulas

Table 8. M/M/c/K Queueing System

$\rho = \lambda \overline{S}$.

$$P_0 = \left[\sum_{n=0}^{c} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \sum_{n=1}^{K-c} \left(\frac{\rho}{c} \right)^n \right]^{-1}.$$

$$P_n = \begin{cases} \dfrac{\rho^n}{n!} P_0 & \text{if } n = 1, 2, \ldots, c, \\ \dfrac{\rho^n}{c!} \left(\dfrac{\rho}{c} \right)^{n-c} P_0 & \text{if } n = c+1, \ldots, K. \end{cases}$$

The average arrival rate of customers who actually enter the system is $\overline{\lambda} = \lambda(1 - P_K)$.

The actual mean server utilization, a, is given by:

$$a = \frac{\overline{\lambda}\, \overline{S}}{c}.$$

$$\overline{Q} = \frac{\rho^c r P_0}{c!(1-r)^2} \left[1 + (K-c) r^{K-c+1} - (K-c+1) r^{K-c} \right],$$

where

$$r = \frac{\rho}{c}.$$

$$\overline{N} = \overline{Q} + \mathbb{E}[N_s] = \overline{Q} + \sum_{n=0}^{c-1} n P_n + c \left(1 - \sum_{n=0}^{c-1} P_n \right).$$

By Little's Law

$$\overline{W} = \frac{\overline{Q}}{\overline{\lambda}}, \qquad \overline{T} = \frac{\overline{N}}{\overline{\lambda}}.$$

$$\Pi_n = \frac{P_n}{1 - P_K}, \quad n = 0, 1, 2, \ldots, K-1,$$

Table 8. M/M/c/K Queueing System (continued)

where Π_n is the probability that an arriving customer who enters the system finds n customers already there.

$$\mathbb{E}[W|W>0] = \frac{\overline{W}}{1-\sum_{n=0}^{c-1}\Pi_n}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.9 M/M/∞ Formulas

Table 9. M/M/∞ Queueing System

$\rho = \lambda \overline{S}$.

$$P_n = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 0, 1, \ldots .$$

Since N has a Poisson distribution,

$$\overline{N} = \rho \quad \text{and} \quad Var(N) = \rho.$$

By Little's Law

$$\overline{T} = \frac{\overline{N}}{\lambda} = \overline{S}.$$

Since there is no queueing for service,

$$\overline{W} = \overline{Q} = 0,$$

and
$F_T(t) = P[T \leq t] = F_S(t) = P[S \leq t]$

That is, T has the same distribution as S.
All the above formulas arc true for the $M/G/\infty$ system, also.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.10 M/M/1/K/K Formulas

Table 10. M/M/1/K/K Queueing System

The mean operating time per machine (sometimes called the the mean time to failure, MTTF) is

$$\mathbb{E}[O] = \frac{1}{\alpha}.$$

The mean repair time per machine by one repairman is

$$\overline{S} = \frac{1}{\mu}.$$

The probability, P_0, that no machines arc out of service is given by

$$P_0 = \left[\sum_{k=0}^{K} \frac{K!}{(K-k)!}\left(\frac{\overline{S}}{\mathbb{E}[O]}\right)^k\right]^{-1} = B[K, z],$$

where $B[\cdot, \cdot]$ is Erlang's B formula and

$$z = \frac{\mathbb{E}[O]}{\overline{S}}.$$

Then, P_n, the probability that n machines arc out of service, is given by

$$P_n = \frac{K!}{(K-n)!} z^{-n} P_0, \quad n = 0, 1, \ldots, K,$$

The formula for P_n can also be written in the form

$$P_n = \frac{\dfrac{z^{K-n}}{(K-n)!}}{\displaystyle\sum_{k=0}^{K} \frac{z^k}{k!}}, \quad n = 0, 1, \ldots, K.$$

$$a = 1 - P_0.$$

$$\overline{\lambda} = \frac{a}{\overline{S}}.$$

$$\overline{T} = \frac{K}{\overline{\lambda}} - \mathbb{E}[O].$$

$$\overline{N} = \overline{\lambda} \cdot \overline{T}.$$

$$\overline{W} = \overline{T} - \overline{S}.$$

Table 10. M/M/1/K/K Queueing System (continued)

$$\Pi_n = \frac{(K-n)P_n}{K-\overline{N}} = \frac{\dfrac{z^{K-n-1}}{(K-n-1)!}}{\displaystyle\sum_{k=0}^{K-1}\frac{z^k}{k!}}, \quad n = 0, 1, 2, \ldots, K-1,$$

where Π_n is the probability that a machine that breaks down finds n machines in the repair facility.

$$F_T(t) = P[T \le t] = 1 - \frac{Q(K-1; z+t\mu)}{Q(K-1; z)}, \quad t \ge 0,$$

where

$$Q(n; x) = e^{-x}\sum_{k=0}^{n}\frac{x^k}{k!}.$$

$$F_T(t) = P[W \le t] = 1 - \frac{Q(K-2; z+t\mu)}{Q(K-1; z)}, \quad t \ge 0,$$

$$\mathbb{E}[W|W > 0] = \frac{\overline{W}}{1-\Pi_0}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.11 M/G/1/K/K Formulas

Table 11. M/G/1/K/K Queueing System

The mean operating time per machine (sometimes called the the mean time to failure, MTTF) is

$$\mathbb{E}[O] = \frac{1}{\alpha}.$$

The mean repair time per machine by one repairman is

$$\overline{S} = \frac{1}{\mu}.$$

The probability, P_0, that no machines arc out of service is given by

$$P_0 = \left[1 + \frac{K\overline{S}}{\mathbb{E}[O]} \sum_{n=0}^{K-1} \binom{K-1}{n} B_n\right]^{-1},$$

where

$$B_n = \begin{cases} 1 & n = 0, \\ \prod_{i=1}^{n} \left(\frac{1-\overline{S}^*[i\alpha]}{\overline{S}^*[i\alpha]}\right) & n = 1, 2, \ldots, K-1. \end{cases}$$

and $\overline{S}^*[\theta]$ is the Laplace-Stieltjes transform of s.

$a = 1 - P_0.$

$\overline{\lambda} = \frac{a}{\overline{S}}.$

$\overline{T} = \frac{K}{\overline{\lambda}} - \mathbb{E}[O].$

$\overline{N} = \overline{\lambda} \cdot \overline{T}.$

$\overline{W} = \overline{T} - \overline{S}.$

$\overline{Q} = \overline{\lambda} \cdot \overline{W}.$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.12 M/M/c/K/K Formulas

Table 12. M/M/c/K/K Queueing System

The mean operating time per machine (sometimes called the the mean time to failure, MTTF) is

$$\mathbb{E}[O] = \frac{1}{\alpha}.$$

The mean repair time per machine by one repairman is

$$\overline{S} = \frac{1}{\mu}.$$

The probability, P_0, that no machines are out of service is given by

$$P_0 = \left[\sum_{k=0}^{c} \binom{K}{k} z^{-k} + \sum_{k=c+1}^{K} \frac{k!}{c! c^{k-c}} \binom{K}{k} z^{-k} \right]^{-1},$$

where

$$z = \frac{\mathbb{E}[O]}{\overline{S}}.$$

Then,P_n, the probability that n machines are out of service is given by

$$P_n = \begin{cases} \binom{K}{n} z^{-n} P_0 & n = 0, 1, \ldots, c, \\ \frac{n!}{c! c^{n-c}} \binom{K}{n} z^{-n} P_0 & n = c+1, \ldots, K. \end{cases}$$

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

$$\overline{W} = \frac{\overline{Q}(\mathbb{E}[O] + \overline{S})}{K - \overline{Q}}.$$

$$\overline{\overline{\lambda}} = \frac{K}{\mathbb{E}[O] + \overline{W} + \overline{S}}.$$

$$\overline{T} = \frac{K}{\overline{\overline{\lambda}}} - \mathbb{E}[O].$$

$$\overline{N} = \overline{\overline{\lambda}}\, \overline{T}.$$

$$\Pi_n = \frac{(K-n)P_n}{K-\overline{N}},$$

where Π_n is the probability that a machine which breaks down finds n inoperable machines already in the repair facility. We denote Π_n by $\Pi_n[K]$ to emphasize the fact that there are K machines. It can be shown that

$$\Pi_n[K] = P_n[K-1], \quad n = 0, 1, \ldots, K-1.$$

$$P_n[K-1] = \frac{c^c}{c!}\frac{p(K-n-1;cz)}{p(K-1;cz)}P_0[K-1],$$

where, of course,

$$p(k;\alpha) = \frac{\alpha^k}{k!}e^{-\alpha}.$$

$$F_T(t) = P[W \le t] = 1 - \frac{c^c Q(K-c-1;cz)P_0[K-1]}{c!p(K-1;cz)}, t \ge 0,$$

where

$$Q(k;\alpha) = e^{-\alpha}\sum_{n=0}^{k}\frac{\alpha^n}{n!}.$$

$$F_T(t) = P[T \le t] = 1 - C_1\exp(-t/\overline{S}) + C_2\frac{Q(K-c-1;c(z+t\mu))}{Q(K-c-1;cz)},$$

$t \ge 0$,

where $C_1 = 1 + C_2$ d'ż˝s

$$C_2 = \frac{c^c Q(K-c-1;cz)}{c!(c-1)(K-c-1)!p(K-1;cz)}P_0[K-1].$$

The probability that a machine that breaks down must wait for repair is given by

$$D = \sum_{n=c}^{K-1}\Pi_n = 1 - \sum_{n=0}^{c-1}\Pi_n.$$

$$\mathbb{E}[W|W>0] = \frac{\overline{W}}{D}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.13 D/D/c/K/K Formulas

Table 13. D/D/c/K/K Queueing System

The mean operating time per machine (sometimes called the the mean time to failure, MTTF) is

$$\mathbb{E}[O] = \frac{1}{\alpha}.$$

The mean repair time per machine by one repairman is

$$\overline{S} = \frac{1}{\mu}.$$

$$a = \min\{1, \frac{K}{c(1+z)}\},$$

where

$$z = \frac{\mathbb{E}[O]}{\overline{S}}.$$

$$\overline{\lambda} = ca\mu = \frac{ca}{\overline{S}}.$$

$$\overline{T} = \frac{K}{\overline{\lambda}} - \mathbb{E}[O].$$

$$\overline{N} = \overline{\lambda}\,\overline{T}.$$

$$\overline{W} = \overline{T} - \overline{S}.$$

$$\overline{Q} = \overline{\lambda}\,\overline{W}.$$

The equations for this model are derived in "A straightforward model of computer performance prediction" by John W. Boyse es David R. Warn in ACM Comput. Surveys, 7(2), (June 1972).

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.14 M/G/1 Formulas

Table 14. M/G/1 Queueing System

The z-transform of N, the steady-state number of customers in the system is given by:

$$G_N(z) = \sum_{n=0}^{\infty} P_n z^n = \frac{(1-\rho)(1-z)\overline{S}^*[\lambda(1-z)]}{\overline{S}^*[\lambda(1-z)] - z},$$

where $\overline{S}^*$ is the Laplace-Stieltjes transform of the servide time S. The Laplace-Stieltjes transforms of T and W are given by

$$W^*[\theta] = \frac{(1-\rho)\theta\overline{S}^*[\theta]}{\theta - \lambda + \lambda\overline{S}^*[\theta]},$$

dż˝s

$$\overline{W}^*[\theta] = \frac{(1-\rho)\theta}{\theta - \lambda + \lambda\overline{S}^*[\theta]}.$$

Each of the three transforms above is called the Pollaczek-Khintchine transform equation by various authors. The probability, P_0, of no customers in the system has the simple and intuitive equation $P_0 = 1 - \rho$, where the server utilization $\rho = \lambda\overline{S}$. The probability that the server is busy is $P[N \geq 1] = \rho$.

$\overline{W} = \dfrac{\lambda\mathbb{E}[S^2]}{2(1-\rho)} = \dfrac{\rho\overline{S}}{1-\rho}\left(\dfrac{1+C_S^2}{2}\right)$ (Pollaczek formula).

$\overline{Q} = \lambda\overline{W}$.

$$Var(Q) = \frac{\lambda^3\mathbb{E}[S^3]}{3(1-\rho)} + \left(\frac{\lambda^2\mathbb{E}[S^2]}{2(1-\rho)}\right)^2 + \frac{\lambda^2\mathbb{E}[S^2]}{2(1-\rho)}.$$

$$\mathbb{E}[W|W>0] = \frac{\overline{S}}{1-\rho}\left(\frac{1+C_S^2}{2}\right).$$

$$\mathbb{E}[W^2] = 2\overline{W}^2 + \frac{\lambda\mathbb{E}[S^3]}{3(1-\rho)}.$$

$Var(W) = \mathbb{E}[W^2] - \overline{W}^2$.

$\overline{T} = \overline{W} + \overline{S}$.

$\overline{N} = \overline{\lambda}\cdot\overline{T} = \overline{Q} + \rho$.

$$Var(N) = \frac{\lambda^3\mathbb{E}[S^3]}{3(1-\rho)} + \left(\frac{\lambda^2\mathbb{E}[S^2]}{2(1-\rho)}\right)^2 + \frac{\lambda^2(3-2\rho)\mathbb{E}[S^2]}{2(1-\rho)} + \rho(1-\rho).$$

$$\mathbb{E}[T^2] = \mathbb{E}[W^2] + \frac{\mathbb{E}[S^2]}{1-\rho}.$$

Table 14. M/G/1 Queueing System (continued)

$Var(T) = \mathbb{E}[T^2] - \overline{T}^2$.

$\pi_T[90] \approx \overline{T} + 1.3\mathbb{D}(T), \quad \pi_T[95] \approx \overline{T} + 2\mathbb{D}(T)$.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Table 15. M/H_2/1 Queueing System

The z-transform of the steady-state number in the system, N, is given by

$$G_N(z) = \sum_{n=0}^{\infty} P_n z^n = C_1 \frac{z_1}{z_1 - z} + C_2 \frac{z_2}{z_2 - z},$$

where z_1 and z_2 are the roots of the next equation

$$a_1 a_2 z^2 - (a_1 + a_2 + a_1 a_2)z + 1 + a_1 + a_2 - a = 0,$$

where

$$a = \lambda \overline{S},$$

$$a_i = \frac{\lambda}{\mu_i}, \quad i = 1, 2,$$

$$C_1 = \frac{(z_1 - 1)(1 - a z_2)}{z_1 - z_2},$$

and

$$C_2 = \frac{(z_2 - 1)(1 - a z_1)}{z_2 - z_1}.$$

From $G_N(z)$ we get

$$P_n = C_1 z_1^{-n} + C_2 z_2^{-n}, \quad n = 0, 1, \ldots.$$

Specifically, $P_0 = 1 - a$.

$$P[N \geq n] = C_1 \frac{z_1^{-n+1}}{z_1 - 1} - C_2 \frac{z_2^{-n+1}}{z_2 - 1}.$$

Additionally,

$$P[N \geq 1] = a.$$

$$F_T(t) = P[W \leq t] = 1 - C_5 e^{-\rho t} - C_6 e^{-bt}, \ t \geq 0,$$

where $\rho = -\zeta_1, \ b = -\zeta_2, \ \zeta_1, \zeta_2$ are the solutions of the

$$\theta^2 + (\mu_1 + \mu_2 - \lambda)\theta + \mu_1 \mu_2 (1 - a) = 0,$$

equation,

$$C_5 = \frac{\lambda(1-a)\zeta_1 + a(1-a)\mu_1\mu_2}{\rho(\zeta_1 - \zeta_2)}$$

and

$$C_6 = \frac{\lambda(1-a)\zeta_2 + a(1-a)\mu_1\mu_2}{\rho(\zeta_2 - \zeta_1)}.$$

$$\overline{W} = \frac{\lambda\mathbb{E}[S^2]}{2(1-a)} = \frac{a\overline{S}}{1-a}\left(\frac{1+C_S^2}{2}\right).$$ (Pollaczek-formula)

$$\mathbb{E}[W|W>0] = \frac{\overline{S}}{1-a}\left(\frac{1+C_S^2}{2}\right).$$

$$\mathbb{E}[W^2] = 2\overline{W}^2 + \frac{\lambda\mathbb{E}[S^3]}{3(1-a)}.$$

In this formula we substitute

$$\mathbb{E}[S^3] = \frac{6p_1}{\mu_1^3} + \frac{6p_2}{\mu_2^3},$$

then

$$Var(W) = \mathbb{E}[W^2] - \overline{W}^2.$$

$$F_T(t) = P[T \le t] = 1 - \pi_a e^{-\mu_a t} - \pi_b e^{-\mu_b t}, \quad t \ge 0,$$

where

$$\pi_a = C_1\frac{z_1}{z_1 - 1},$$

$$\pi_b = C_2\frac{z_2}{z_2 - 1},$$

Table 15. M/H_2/1 Queueing System (continued)

$\mu_a = \lambda(z_1 - 1),$

and

$\mu_b = \lambda(z_2 - 1).$

$\overline{T} = \overline{W} + \overline{S}.$

$\mathbb{E}[T^2] = \mathbb{E}[W^2] + \dfrac{\mathbb{E}[S^2]}{1-a},$

where of course

$\mathbb{E}[S^2] = \dfrac{2p_1}{\mu_1^2} + \dfrac{2p_2}{\mu_2^2}.$

$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$

$C_T^2 = \dfrac{\mathbb{E}[T^2]}{\overline{T}^2} - 1.$

$\overline{Q} = \overline{\lambda} \cdot \overline{W} = \dfrac{a^2}{1-a}\left(\dfrac{1+C_S^2}{2}\right).$

$Var(Q) = \dfrac{\lambda^3\mathbb{E}[S^3]}{3(1-a)} + \left(\dfrac{\lambda^2\mathbb{E}[S^2]}{2(1-a)}\right)^2 + \dfrac{\lambda^2\mathbb{E}[S^2]}{2(1-a)}.$

$\overline{N} = \lambda\overline{T} = \overline{Q} + a.$

$Var(N) = \dfrac{\lambda^3\mathbb{E}[S^3]}{3(1-a)} + \left(\dfrac{\lambda^2\mathbb{E}[S^2]}{2(1-a)}\right)^2 + \dfrac{\lambda^2(3-2a)\mathbb{E}[S^2]}{2(1-a)} + a(1-a).$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Since S has Gamma-distribution

$$\mathbb{E}[S^n] = \frac{\beta(\beta+1)\dots(\beta+n-1)}{\alpha^n}, \quad n = 1, 2, \dots.$$

Since

$$C_S^2 = \frac{1}{\beta},$$

so

$$\mathbb{E}[S^2] = \overline{S}^2(1 + C_S^2),$$

$$\mathbb{E}[S^3] = \overline{S}^3(1 + C_S^2)(1 + 2C_S^2),$$

and

$$\mathbb{E}[s^n] = \overline{S}^n \prod_{k=1}^{n-1}(1 + kC_S^2), \quad n = 1, 2, \dots.$$

This time

$$\overline{W} = \frac{\lambda\mathbb{E}[S^2]}{2(1-a)} = \frac{a\overline{S}}{1-a}\left(\frac{1+C_S^2}{2}\right),$$

$$\overline{Q} = \overline{\lambda} \cdot \overline{W},$$

$$Var(Q) = \frac{a^2(1+C_S^2)}{2(1-a)}\left[1 + \frac{a^2(1+C_S^2)}{2(1-a)} + \frac{2a(1+2C_S^2)}{3}\right],$$

$$\mathbb{E}[W|W > 0] = \frac{\overline{S}}{1-a}\left(\frac{1+C_S^2}{2}\right),$$

$$\mathbb{E}[W^2] = 2\overline{W}^2 + \frac{a\overline{S}^2(1+C_S^2)(1+2C_S^2)}{3(1-a)},$$

$$Var(W) = \mathbb{E}[W^2] - \overline{W}^2,$$

$$\overline{T} = \overline{W} + \overline{S}, \qquad \overline{N} = \overline{\lambda} \cdot \overline{T} = \overline{Q} + a,$$

$$Var(N) = \frac{a^3(1+C_S^2)(1+2C_S^2)}{3(1-a)} + \left(\frac{a^2(1+C_S^2)}{2(1-a)}\right)^2 + \frac{a^2(3-2a)(1+C_S^2)}{2(1-a)} + a(1-a).$$

$$\mathbb{E}[T^2] = \mathbb{E}[W^2] + \frac{\overline{S}^2(1+C_S^2)}{1-a}.$$

$$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$$

$$\pi_T[90] \approx \overline{T} + 1.3\mathbb{D}(T), \quad \pi_T[95] \approx \overline{T} + 2\mathbb{D}(T).$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Table 17. M/E_k/1 Queueing System

Mivel S Since S has Erlang -k distribution, hence

$$\mathbb{E}[S^n] = \left(1+\frac{1}{k}\right)\left(1+\frac{2}{k}\right)\ldots\left(1+\frac{n-1}{k}\right)\overline{S}^n, \quad n=1,2,\ldots.$$

so

$\mathbb{E}[S^2] = \overline{S}^2\left(1+\frac{1}{k}\right)$,

and

$\mathbb{E}[s^3] = \overline{S}^3\left(1+\frac{1}{k}\right)\left(1+\frac{2}{k}\right)$.

This time

$$\overline{W} = \frac{\lambda\mathbb{E}[s^2]}{2(1-a)} = \frac{a\overline{S}}{1-a}\left(\frac{1+\frac{1}{k}}{2}\right).$$ (Pollaczek's formula)

$$\overline{Q} = \overline{\lambda}\cdot\overline{W}.$$

$$Var(Q) = \frac{a^2(1+k)}{2k(1-a)}\left[1+\frac{a^2(1+k)}{2k(1-a)}+\frac{2a(k+2)}{3k}\right].$$

$$\mathbb{E}[W|W>0] = \frac{\overline{S}}{1-a}\left(\frac{1+\frac{1}{k}}{2}\right).$$

$$\mathbb{E}[W^2] = 2\overline{W}^2 + \frac{a\overline{S}^2(k+1)(k+2)}{3k^2(1-a)}.$$

$$Var(W) = \mathbb{E}[W^2] - \overline{W}^2.$$

$$\overline{T} = \overline{W}+\overline{S}, \qquad \overline{N} = \overline{\lambda}\cdot\overline{T} = \overline{Q}+a.$$

$$Var(N) = \frac{a^3(k+1)(k+2)}{3k^2(1-a)} + \left(\frac{a^2(1+\frac{1}{k})}{2(1-a)}\right)^2 + \frac{a^2(3-2a)(1+\frac{1}{k})}{2(1-a)} + a(1-a).$$

$$\mathbb{E}[T^2] = \mathbb{E}[W^2] + \frac{\overline{S}^2(1+\frac{1}{k})}{1-a}.$$

$$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$$

$$\pi_T[90] \approx \overline{T} + 1.3\mathbb{D}(T), \quad \pi_T[95] \approx \overline{T} + 2\mathbb{D}(T).$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Since S has a constant distribution

$$\mathbb{E}[S^n] = \overline{S}^n, \quad n = 1, 2, \ldots,$$

so

$$G_N(z) = \frac{(1-a)(1-z)}{1 - ze^{a(1-z)}}.$$

We suppose that

$$|ze^{a(1-z)}| < 1,$$

we can expand $G_N(z)$ in the geometric series

$$g_N(z) = (1-a)(1-z)\sum_{j=0}^{\infty} \left[ze^{a(1-z)}\right]^j.$$

This thime we can show that,

$$P_1 = (1-a)(e^a - 1),$$

and

$$P_n = (1-a)\sum_{j=1}^{n} \frac{(-1)^{n-j}(ja)^{n-j-1}(ja+n-j)e^{ja}}{(n-j)!} \quad n = 2, 3, \ldots.$$

Additionally

$$F_T(t) = \sum_{n=0}^{k-1} P_n + P_k \left(\frac{t - (k-1)\overline{S}}{\overline{S}} \right),$$

where $(k-1)\overline{S} \leq t \leq k\overline{S}, \quad k = 1, 2, \ldots.$

Table 18. M/D/1 Queueing System (continued)

So,

$\overline{W}[0] = P_0.$

$\overline{W} = \dfrac{a\overline{S}}{2(1-a)}.$

$F_W[W|W>0] = \dfrac{\overline{S}}{2(1-a)}.$

$\mathbb{E}[W^2] = 2\overline{W}^2 + \dfrac{a\overline{S}^2}{3(1-a)}.$

$Var(W) = \mathbb{E}[W^2] - \overline{W}^2.$

$\overline{Q} = \lambda\overline{W} = \dfrac{a^2}{2(1-a)}.$

$Var(Q) = \dfrac{a^3}{3(1-a)} + \left[\dfrac{a^2}{2(1-a)}\right]^2 + \dfrac{a^2}{2(1-a)}.$

$$F_T(t) = \begin{cases} 0 & if \quad t < \overline{S}, \\ \sum\limits_{n=0}^{k-1} P_n + P_k\left(\frac{t-k\overline{S}}{\overline{S}}\right) & if \quad t \geq \overline{S}. \end{cases}$$

where

$k\overline{S} \leq t < (k+1)\overline{S}, \quad k = 1, 2, \ldots.$

$\overline{T} = \overline{W} + \overline{S}.$

$\mathbb{E}[T^2] = \mathbb{E}[W^2] + \dfrac{\overline{S}^2}{1-a}.$

$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$

$\overline{N} = \overline{\lambda} \cdot \overline{T} = \overline{Q} + a.$

$Var(N) = \dfrac{a^3}{3(1-a)} + \left(\dfrac{a^2}{2(1-a)}\right)^2 + \dfrac{a^2(3-2a)}{2(1-a)} + a(1-a).$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.15 GI/M/1 Formulas

Table 19. GI/M/1 Queueing System

The steady-state probability that an arriving customer will find the system empty, is the unique solution of the equation $1-\Pi_0 = A^*[\mu\Pi_0]$ such that $0 < \Pi_0 < 1$, where $A^*[\Theta]$ is the Laplace-Stieltjes transform of r. The steady-state number of customers in the system, N has the distribution $\{P_n\}$, where $P_0 = P[N = 0] = 1 - a,\ P_n = a\Pi_0(1 - \Pi_0)^{n-1},\ n = 1, 2, \ldots,$

$\overline{N} = \dfrac{a}{\Pi_0}$, and $Var(N) = \frac{a(2-\Pi_0-a)}{\Pi_0^2}$.

$\overline{Q} = \dfrac{(1-\Pi_0)a}{\Pi_0}$.

$Var(Q) = \dfrac{a(1-\Pi_0)(2-\Pi_0-a(1-\Pi_0))}{\Pi_0^2}$.

$\mathbb{E}[Q|Q > 0] = \dfrac{1}{\Pi_0}$.

$\overline{T} = \dfrac{\overline{S}}{\Pi_0}$.

$F_T(t) = P[T \le t] = 1 - \exp(-t/\overline{T})$.

$\Pi_T[r] = \overline{T} \ln\left[\dfrac{100}{100-r}\right]$.

$\Pi_T[90] = \overline{T}\ln 10, \quad \Pi_T[95] = \overline{T}\ln 20$.

$\overline{W} = (1-\Pi_0)\dfrac{\overline{S}}{\Pi_0}$.

$Var(W) = (1-\Pi_0^2)\left(\dfrac{\overline{S}}{\Pi_0}\right)^2$.

$F_T(t) = P[W \le t] = 1 - (1-\Pi_0)\exp(-t/\overline{T})$.

$\Pi_W[r] = \max\left\{0, \overline{T}\ln\left(\dfrac{100(1-\Pi_0)}{100-r}\right)\right\}$.

W', the queueing time for those who must, has the same distribution as T.

Table 20. Π_0 versus a for GI/M/1 Queueing System [1]

a	E_2	E_3	U	D	H_2	H_2
0.100	0.970820	0.987344	0.947214	0.999955	0.815535	0.810575
0.200	0.906226	0.940970	0.887316	0.993023	0.662348	0.624404
0.300	0.821954	0.868115	0.817247	0.959118	0.536805	0.444949
0.400	0.724695	0.776051	0.734687	0.892645	0.432456	0.281265
0.500	0.618034	0.669467	0.639232	0.796812	0.343070	0.154303
0.600	0.504159	0.551451	0.531597	0.675757	0.263941	0.081265
0.700	0.384523	0.626137	0.412839	0.533004	0.191856	0.044949
0.800	0.260147	0.289066	0.284028	0.371370	0.124695	0.024404
0.900	0.131782	0.147390	0.146133	0.193100	0.061057	0.010495
0.950	0.066288	0.074362	0.074048	0.098305	0.030252	0.004999
0.980	0.026607	0.029899	0.029849	0.039732	0.012039	0.001941
0.999	0.001333	0.001500	0.001500	0.001999	0.000600	0.000095

[1] At the first H_2 distribution $p_1 = 0.4,\ \mu_1 = 0.5\lambda,\ \mu_2 = 3\lambda$. At the second H_2 distribution $p_1 = 0.024405,\ \mu_1 = 2p_1\lambda,\ \mu_2 = 2p_2\lambda$.

5.16 GI/M/c Formulas

Table 21. GI/M/c Queueing System

Let $\Pi_n, n = 0, 1, 2, \ldots$ be the steady state number of customers that an arriving customer finds in the system. Then

$$\Pi_n = \begin{cases} \sum_{i=n}^{c-1} (-1)^{i-n} \binom{i}{n} U_i, & n = 0, 1, \ldots, c-2, \\ \\ D\omega^{n-c}, & n = c-1, c, \ldots, \end{cases}$$

where ω is the unique solution of the equation $\omega = A^*[c\mu(1-\omega)]$ such that $0 < \omega < 1$, where $A^*[\theta]$ is the Laplace-Stieltjes transform of r,

$g_j = A^*[j\mu], \quad j = 1, 2, \ldots, c,$

$$C_j = \begin{cases} 1, & j = 0, \\ \prod_{i=1}^{j} \left(\frac{g_i}{1-g_i} \right), & j = 1, 2, \ldots, c, \end{cases}$$

$$D = \left[\frac{1}{1-\omega} + \sum_{j=1}^{c} \frac{\binom{c}{j}}{C_j(1-g_j)} \left(\frac{c(1-g_j)-j)}{c(1-\omega)-j} \right) \right]^{-1}$$

and

$$U_n = DC_n \sum_{j=n+1}^{c} \frac{\binom{c}{j}}{C_j(1-g_j)} \left(\frac{c(1-g_j)-j)}{c(1-\omega)-j} \right), \quad n = 0, 1, \ldots, c-1.$$

$$F_T(t) = P[W \le t] = 1 - P[W > 0]e^{-c\mu(1-\omega)t},\ t \ge 0,$$

where

$$P[W > 0] = \frac{D}{1-\omega}. \quad \overline{W} = \frac{D\overline{S}}{c(1-\omega)^2}. \quad \mathbb{E}[W|W > 0] = \frac{\overline{S}}{c(1-\omega)}.$$

If $c(1-\omega) \neq 1$, then

$$F_T(t) = P[\omega \le t] = 1 + (G-1)e^{-\mu t} - Ge^{-c\mu(1-\omega)t}, \quad t \ge 0,$$

where

$$G = \frac{D}{(1-\omega)[1-c(1-\omega)]}.$$

When $c(1-\omega) = 1$, then

$$F_T(t) = P[\omega \le t] = 1 - \left[1 + \frac{D\mu t}{1-\omega}\right] e^{-\mu t}, \quad t \ge 0.$$

We also have

$$\overline{T} = \overline{W} + \overline{S}.$$

$$P_0 = 1 - \frac{\lambda\overline{S}}{c} - \lambda\overline{S}\sum_{j=1}^{c-1} \Pi_{j-1}\left(\frac{1}{j} - \frac{1}{c}\right).$$

$$P_n = \begin{cases} \frac{\lambda\overline{S}\Pi_{n-1}}{n,} & n = 1, 2, \ldots, c-1, \\ \frac{\lambda\overline{S}\Pi_{n-1}}{c,} & n = c, c+1, \ldots. \end{cases}$$

5.17 M/G/1 Priority queueing system

Table 22. M/G/1 Queueing System (classes, no priority)

There are n customer classes. Customers from class i arrive in a Poisson pattern with mean arrival rate $\lambda_i, i = 1, 2, \ldots, n$. Each class has its own general service time with $\mathbb{E}[S_i] = 1/\mu_i, \quad \mathbb{E}[S_i^2], \quad \mathbb{E}[S_i^3]$. All customers served on a FCFS basis with no consideration for class. The total arrival stream to the system has a Poisson arrival pattern with

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

The first three moments of service time arc given by

$$\overline{S} = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n],$$

$$\mathbb{E}[S^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^2],$$

and

$$\mathbb{E}[S^3] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^3] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^3] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^3],$$

By Pollaczek's formula,

$$\overline{W} = \frac{\lambda\mathbb{E}[S^2]}{2(1-a)}.$$

The mean time in the system for each class is given by

$$\overline{T}_i = \overline{W} + \mathbb{E}[S_i], \quad i = 1, 2, \ldots, n.$$

The overall mean customer time in the system,

$$\overline{T} = \frac{\lambda_1}{\lambda}\overline{T}_1 + \frac{\lambda_2}{\lambda}\overline{T}_2 + \ldots + \frac{\lambda_n}{\lambda}\overline{T}_n.$$

The variance of the waiting time

$$Var(W) = \frac{\lambda\mathbb{E}[S^3]}{3(1-a)} + \frac{\lambda^2(\mathbb{E}[S^2])^2}{4(1-a)^2}.$$

The variance of T is given by

$$Var(T_i) = Var(W) + Var(S_i), \quad i = 1, 2, \ldots, n.$$

The second moment of T by class is

$$\mathbb{E}[T_i^2] = Var(T_i) + \overline{T}_i^2, \quad i = 1, 2, \ldots, n.$$

Table 22. M/G/1 Queueing System (classes, no priority)
(continued)

Thus, the overall second moment of T is given by

$$\mathbb{E}[T^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[T_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[T_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[T_n^2],$$

and

$$Var(T) = \mathbb{E}[T^2] - \overline{T}^2.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Table 23. M/G/1 Nonpreemptive (HOL) Queueing System

There are n priority classes with each class having a Poisson arrival pattern with mean arrival rate λ_i. Each customer has the same exponential service time requirement. Then the overall arrival pattern is Poiisson with mean:

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

The server utilization

$$\overline{S} = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n],$$

$$\mathbb{E}[S^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^2],$$

and

$$\mathbb{E}[S^3] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^3] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^3] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^3],$$

Let

$$\rho_j = \lambda_1\mathbb{E}[S_1] + \lambda_2\mathbb{E}[S_2] + \ldots + \lambda_j\mathbb{E}[S_j], \quad j = 1, 2, \ldots, n,$$

and notice that

$$\rho_n = \rho = \lambda\overline{S}.$$

The mean times in the queues:

$$\overline{W}_j = \mathbb{E}[W_j] = \frac{\lambda\mathbb{E}[S^2]}{2(1-\rho_{j-1})(1-\rho_j)},$$

$$j = 1, 2, \ldots, n, \quad \rho_0 = 0.$$

Table 23. M/G/1 Nonpreemptive (HOL) Queueing System
(continued)

The mean queue lengths are

$$\overline{Q}_j = \overline{\lambda}_j \cdot \overline{W}_j, \quad j = 1, 2, \ldots, n.$$

The unified time in the queue

$$\overline{W} = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n].$$

The mean times of staying in the system

$$\overline{T}_j = \mathbb{E}[T_j] = \mathbb{E}[W_j] + \mathbb{E}[S_j], \quad j = 1, 2, \ldots, n,$$

and the average of the customers staying at the system is

$$\overline{N}_j = \overline{\lambda}_j \cdot \overline{T}_j, \quad j = 1, 2, \ldots, n.$$

The total time in the system

$$\overline{T} = \overline{W} + \overline{S}.$$

The total queue length

$$\overline{Q} = \overline{\lambda} \cdot \overline{W},$$

and the average of the customers staying at the system

$$\overline{N} = \overline{\lambda} \cdot \overline{T}.$$

The variance of the total time stayed in the system by class

$$Var(T_j) = Var(S_j) + \frac{\lambda\mathbb{E}[S^3]}{3(1-\rho_{j-1})^2(1-\rho_j)}$$
$$+\frac{\lambda\mathbb{E}[S^2]\left(2\sum_{i=1}^{j}\lambda_i\mathbb{E}[S_i^2] - \lambda\mathbb{E}[S^2]\right)}{4(1-\rho_{j-1})^2(1-\rho_j)^2}$$
$$+\frac{\lambda\mathbb{E}[S^2]\sum_{i=1}^{j-1}\lambda_i\mathbb{E}[S_i^2]}{2(1-\rho_{j-1})^3(1-\rho_j)}, \quad j = 1, 2, \ldots, n.$$

Table 23. M/G/1 Nonpreemptive (HOL) Queueing System
(continued)

The variance of the total time stayed in the system

$$Var(T) = \frac{\lambda_1}{\lambda}[Var(T_1) + \overline{T}_1^2] + \frac{\lambda_2}{\lambda}[Var(T_2) + \overline{T}_2^2]$$
$$+ \ldots + \frac{\lambda_n}{\lambda}[Var(T_n) + \overline{T}_n^2] - \overline{T}^2.$$

The variance of the waiting time by class

$$Var(W_j) = Var(T_j) - Var(S_j), \quad j = 1, 2, \ldots, n.$$

We know that $\mathbb{E}[W_j^2] = Var(W_j) + \overline{W}_j^2, \quad j = 1, 2, \ldots, n,$

so

$$\mathbb{E}[W^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n^2].$$

Finally

$$Var(W) = \mathbb{E}[W^2] - \overline{W}^2.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Table 24. M/G/1 absolute priority Queueing System

There are n customer classes. Class 1 customers receive the most favorable treatment; class n customers receive the least favorable treatment. Customers from class i arrive in a Poisson pattern with mean arrival rate λ_i,$t = 1, 2, \ldots, n$. Each class has its own general service time with $\mathbb{E}[S_i] = 1/\mu_i$, and finite second and third moments $\mathbb{E}[S_i^2]$, $\mathbb{E}[S_i^3]$. The priority system is preemptive resume, which means that if a customer of class j is receiving service when a customer of class $i < j$ arrives, the arriving customer preempts the server and the customer who was preempted returns to the head of the line for class j customers. The preempted customer resumes service at the point of interruption upon reentering the service facility. The total arrival stream to the system has a Poisson arrival pattern with

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

The first three moment of service time are given by:

$$\overline{S} = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n],$$

$$\mathbb{E}[S^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^2],$$

$$\mathbb{E}[S^3] = \frac{\lambda_1}{\lambda}\mathbb{E}[S_1^3] + \frac{\lambda_2}{\lambda}\mathbb{E}[S_2^3] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[S_n^3].$$

Let

$$\rho_j = \lambda_1\mathbb{E}[S_1] + \lambda_2\mathbb{E}[S_2] + \ldots + \lambda_j\mathbb{E}[S_j], \quad j = 1, 2, \ldots, n,$$

and notice that

$$\rho_n = \rho = \lambda\overline{S}.$$

The mean time in the system for each class is

$$\overline{T}_j = \mathbb{E}[T_j] = \frac{1}{1-\rho_{j-1}}\left[\mathbb{E}[S_j] + \frac{\sum_{i=1}^{j}\lambda_i\mathbb{E}[S_i^2]}{2(1-\rho_j)}\right],$$

$$\rho_0 = 0, \quad j = 1, 2, \ldots, n.$$

Table 24. M/G/1 absolute priority Queueing System

(continued)

Waiting times

$$\overline{W}_j = \mathbb{E}[T_j] - \mathbb{E}[S_j], \quad j = 1, 2, \ldots, n.$$

The mean length of the queue number j:

$$\overline{Q}_j = \lambda_j \overline{W}_j, \quad j = 1, 2, \ldots, n.$$

The total waiting time, $\overline{W}$, is given by:

$$\overline{W} = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n].$$

The mean number of customers staying in the system for each class is

$$\overline{N}_j = \lambda_j \overline{W}_j, \quad j = 1, 2, \ldots, n.$$

The mean total time is

$$\overline{T} = \frac{\lambda_1}{\lambda}\overline{T}_1 + \frac{\lambda_2}{\lambda}\overline{T}_2 + \ldots + \frac{\lambda_n}{\lambda}\overline{T}_n = \overline{W} + \overline{S}.$$

The mean number of customers waiting in the queue is

$$\overline{Q} = \overline{\lambda} \cdot \overline{W},$$

and the average number of customers staying in the system

$$\overline{N} = \overline{\lambda} \cdot \overline{T}.$$

Table 24. M/G/1 absolute priority Queueing System
(continued)

The variance of the total time of staying in the system for each class is

$$Var(T_j) = \frac{Var(S_j)}{(1-\rho_{j-1})^2} + \frac{\mathbb{E}[S_j]\sum_{i=1}^{j-1}\lambda_i\mathbb{E}[S_i^2]}{(1-\rho_{j-1})^3}$$
$$+\frac{\sum_{i=1}^{j}\lambda_i\mathbb{E}[S_i^3]}{3(1-\rho_{j-1})^2(1-\rho_j)} + \frac{\left(\sum_{i=1}^{j}\lambda_i\mathbb{E}[S_i^2]\right)^2}{4(1-\rho_{j-1})^2(1-\rho_j)^2}$$
$$+\frac{\left(\sum_{i=1}^{j}\lambda_i\mathbb{E}[S_i^2]\right)\left(\sum_{i=1}^{j-1}\lambda_i\mathbb{E}[S_i^2]\right)}{2(1-\rho_{j-1})^3(1-\rho_j)}, \quad \rho_0 = 0, \quad j = 1,2,\ldots,n.$$

The overall variance

$$Var(T) = \frac{\lambda_1}{\lambda}[Var(T_1) + \overline{T}_1^2] + \frac{\lambda_2}{\lambda}[Var(T_2) + \overline{T}_2^2]$$
$$+\ldots+\frac{\lambda_n}{\lambda}[Var(T_n) + \overline{T}_n^2] - \overline{T}^2.$$

The variance of waiting times for each class is

$$Var(W_j) = Var(T_j) - Var(S_j), \quad j = 1,2,\ldots,n.$$

Because,

$$\mathbb{E}[W_j^2] = Var(W_j) + \overline{W}_j^2, \quad j = 1,2,\ldots,n,$$

so

$$\mathbb{E}[W^2] = \frac{\lambda_1}{\lambda}\mathbb{E}[W_1^2] + \frac{\lambda_2}{\lambda}\mathbb{E}[W_2^2] + \ldots + \frac{\lambda_n}{\lambda}\mathbb{E}[W_n^2].$$

Finally

$$Var(W) = \mathbb{E}[W^2] - \overline{W}^2.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.18 M/G/c Processor Sharing system

Table 25. M/G/1 processor sharing Queueing System

The Poisson arrival stream has an average arrival rate of λ and the average service rate is μ. The service time distribution is general with the restriction that its Laplace transform is rational, with the denominator having degree at least one higher than the numerator. Equivalently. the service time, s, is Coxian. The priority system is processor-sharing, which means that if a customer arrives when there are already $n-1$ customers in the system, the arriving customer (and all the others) receive service at the average rate μ/n. Then $P_n = \rho^n(1-\rho),\ n = 0, 1, \ldots,$ where $\rho = \lambda/\mu$. We also have

$$\overline{N} = \frac{\rho}{1-\rho}, \quad \mathbb{E}[T|S=t] = \frac{t}{1-\rho}, \text{ and } \overline{T} = \frac{\overline{S}}{1-\rho}.$$

Finally

$$\mathbb{E}[W|S=t] = \frac{\rho t}{1-\rho}, \text{ and } \overline{W} = \frac{\rho\overline{S}}{1-\rho}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Table 26. M/G/c processor sharing Queueing System

The Poisson arrival stream has an average arrival rate of λ. The service time distribution is general with the restriction that its Laplace transform is rational, with the denominator having degree at least one higher than the numerator. Equivalently, the service time, s, is Coxian. The priority system is processor-sharing, which works as follows. When the number of customers in the service center, is less than c, then each customers is served simultaneously by one server; that is, each receives service at the rate μ. When $N > c$. each customer simultaneously receives service at the rate $c\mu/N$. We find that just as for the M/G/l processor-sharing system.

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

5.19 M/M/c Priority system

Table 27. M/M/c relative priority (HOL) Queueing System

There are n priority classes with each class having a Poisson arrival pattern with mean arrival rate λ_i. Each customer has the same exponential service time requirement. Then the overall arrival pattern is Poisson with mean $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$. The server utilization

$$a = \frac{\lambda \overline{S}}{c} = \frac{\lambda}{c\mu},$$

$$\overline{W}_1 = \frac{C[c,\rho]\overline{S}}{c(1-\lambda_1 \overline{S}/c)},$$

and these equations are also true:

$$\overline{W}_j = \frac{C[c,\rho]\overline{S}}{c\left[1-\left(\overline{S}\sum_{i=1}^{j-1}\lambda_i\right)/c\right]\left[1-\left(\overline{S}\sum_{i=1}^{j}\lambda_i\right)/c\right]}, \quad j = 2,\ldots,n.$$

$$\overline{T}_j = \overline{W}_j + \overline{S}, \quad j = 1,2,\ldots,n.$$

$$\overline{Q}_j = \overline{\lambda}_j \cdot \overline{W}_j, \quad j = 1,2,\ldots,n.$$

$$\overline{N}_j = \overline{\lambda}_j \cdot \overline{T}_j, \quad j = 1,2,\ldots,n.$$

$$\overline{W} = \frac{\lambda_1}{\lambda}\overline{W}_1 + \frac{\lambda_2}{\lambda}\overline{W}_2 + \ldots + \frac{\lambda_n}{\lambda}\overline{W}_n.$$

$$\overline{Q} = \overline{\lambda} \cdot \overline{W}.$$

$$\overline{T} = \overline{W} + \overline{S}.$$

$$\overline{N} = \overline{\lambda} \cdot \overline{T}.$$

Java applets for direct calculations can be found at
`https://qsa.inf.unideb.hu`

Appendix and Bibliography

In this Appendix some properties of the generating function, sometimes called as z-transform, and the Laplace-transform are listed. More properties can be found, for example in Kleinrock [62].

Some properties of the generating function

	Sequence $\Longleftrightarrow$	Generating function
1.	$f_n,\ n = 0, 1, 2, \ldots$	$G(z) = \sum_{n=0}^{\infty} f_n z^n$
2.	$af_n + bg_n$	$aG(z) + bH(z)$
3.	$a^n f_n$	$f(az)$
4.	$f_{\frac{n}{k}},\ n = 0, k, 2k, \ldots$	$G(z^k)$
5.	$f_{n+k},\ k > 0$	$\frac{G(z)}{z^k} - \sum_{i=1}^{k} z^{i-k-1} f_{i-1}$
6.	$f_{n-k},\ k > 0$	$z^k G(z)$
7.	$n(n-1)\cdots(n-m+1)f_n$	$z^m \frac{d^m}{dz^m} G(z), \quad m \geq 1$
8.	$f_n * g_n := \sum_{k=0}^{\infty} f_{n-k} g_k$	$G(z)H(z)$
9.	$f_n - f_{n-1}$	$(1-z)G(z)$
10.	$\sum_{k=0}^{n} f_k,\ n = 0, 1, 2, \ldots$	$\frac{G(z)}{1-z}$
11.	$\sum \frac{\partial}{\partial a} f_n$	$\frac{\partial}{\partial a} G(z)$
12.	Series sum property	$G(1) = \sum_{n=0}^{\infty} f_n$
13.	Alternating sum property	$G(-1) = \sum_{n=0}^{\infty} (-1)^n f_n$
14.	Initial value theorem	$G(0) = f_0$
15.	Intermediate value theorem	$\frac{1}{n!} \frac{d^n G(z)}{dz^n}\Big\|_{z=0} = f_n$
16.	Final value theorem	$\lim_{z \to 1}(1-z)G(z) = \lim_{n \to \infty} f_n$

Some properties of the Laplace-transform

	Function $\Longleftrightarrow$	Transform
1.	$f(t),\ t \geq 0$	$f^*(s) = \int_0^\infty f(t)e^{-st}dt$
2.	$af(t) + bg(t)$	$af^*(s) + bg^*(s)$
3.	$f\left(\frac{t}{a}\right),\quad (a > 0)$	$af^*(as)$
4.	$f(t-a)$	$e^{-as}f^*(s)$
5.	$e^{-at}f(t)$	$f^*(s+a)$
6.	$t^n f(t)$	$(-1)^n \frac{d^n f^*(s)}{ds^n}$
7.	$\frac{f(t)}{t}$	$\int_{s_1=s}^\infty f^*(s_1)ds_1$
8.	$\frac{f(t)}{t^n}$	$\int_{s_1=s}^\infty ds_1 \int_{s_2=s_1}^\infty ds_2 \ldots \int_{s_n=s_{n-1}}^\infty ds_n f^*(s_n)$
9.	$f(t) * g(t) = \int_0^t f(t-x)g(x)dx$	$f^*(s)g^*(s)$
10.	$\frac{df(t)}{dt}$	$sf^*(s) - f(0)$
11.	$\frac{d^n f(t)}{dt^n} := f^{(n)}(t)$	$s^n f^*(s) - s^{n-1}f(0) - s^{n-2}f'(0) - ... - f^{(n-1)}(0)$
12.	$\frac{\partial}{\partial a}f(t)$ a is parameter a	$\frac{\partial}{\partial a}F(s)$
13.	Integral property	$f^*(0) = \int_0^\infty f(t)dt$
14.	Initial value theorem	$\lim_{s\to\infty} sf^*(s) = \lim_{t\to 0} f(t)$
15.	Final value theorem	$\lim_{s\to 0} sf^*(s) = \lim_{t\to\infty} f(t)$ if $sf^*(s)$ is analytic for $Re(s) \geq 0$

Bibliography

[1] Adan, I., and Resing, J. Queueing Theory . `http://www.win.tue.nl/~iadan/queueing.pdf`, 2015.

[2] Alfa, A. S. *Applied discrete-time queues*. Springer, 2016.

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

[4] Anisimov, V., Zakusilo, O., and Donchenko, V. *Elements of queueing theory and asymptotic analysis of systems*. Visha Skola, Kiev, 1987.

[5] Artalejo, J., and Gómez-Corral, A. *Retrial queueing systems*. Springer, Berlin, 2008.

[6] Asztalos, D. Optimal control of finite source priority queues with computer system applications. *Computers & Mathematics with Applications 6* (1980), 425–431.

[7] Baron, M. *Probability and statistics for computer scientists*. CRC Press, 2019.

[8] Begain, K., Bolch, G., and Herold, H. *Practical perfromance modeling, Application of the MOSEL language*. Wiley & Sons, New York, 2001.

[9] Bhat, U. N. *An introduction to queueing theory: modeling and analysis in applications*. Birkhäuser, 2015.

[10] Bocharov, P. P., D'Apice, C., and Pechinkin, A. *Queueing theory*. Walter de Gruyter, 2011.

[11] Böhm, W. *A Course on Queueing Models* . Chapman and Hall/CRC, 2016.

[12] Bolch, G., Greiner, S., de Meer, H., and Trivedi, K. *Queueing networks and Markov chains, 2nd ed.* Wiley & Sons, New York, 2006.

[13] Borovkov, A. *Stochastic processes in queueing theory*, vol. 4. Springer Science & Business Media, 2012.

[14] Bose, S. *An introduction to queueing systems*. Kluwer Academic/Plenum Publishers, New York, 2002.

[15] Breuer, L., and Baum, D. *An introduction to queueing theory and matrix-analytic methods*. Springer, 2005.

[16] Brockmeyer, E., Halstrom, H., and Jensen, A. The life and works of a.k. erlang. *Academy of Technical Sciences, Copenhagen* (1948).

[17] Bunday, B., and Scraton, R. The G/M/r machine interference model. *European Journal of Operational Research 4* (1980), 399–402.

[18] Chan, W. C. *An elementary introduction to queueing systems.* World Scientific, 2014.

[19] Chee-Hock, N., and Boon-He, S. *Queueing modelling fundamentals, 2nd ed.* Wiley & Son, Chichester, 2002.

[20] Chen, H., and Yao, D. D. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, vol. 46. Springer Science & Business Media, 2013.

[21] Chun, Y., et al. *Fair queueing.* Springer, 2016.

[22] Cohen, J. The multiple phase service network with generalized processor sharing. *Acta Informatica 12* (1979), 245–284.

[23] Cooper, R. *Introduction to Queueing Theory, 3-rd Edition.* CEE Press, Washington, 1990.
`http://web2.uwindsor.ca/math/hlynka/qonline.html`.

[24] Csige, L., and Tomkó, J. Machine interference problem with exponential distributions (in Hungarian). *Alkalmazott Matematikai Lapok* (1982), 107–124.

[25] Daigle, J. *Queueing theory with applications to packet telecommunication.* Springer, New York, 2005.

[26] Daigle, J. N. *Queueing theory for telecommunications.* Addison-Wesley, Reading, MA, 1992.

[27] Dattatreya, G. *Performance analysis of queuing and computer networks.* CRC Press, Boca Raton, 2008.

[28] Dshalalow, J. H. *Frontiers in queueing : Models and applications in science and engineering.* CRC Press., Boca Raton, 1997.

[29] Erlang, A. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B 20* (1909), 33–39.

[30] Erlang, A. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineers' Journal 10* (1918), 189–197.

[31] Falin, G., and Templeton, J. *Retrial queues.* Chapman and Hall, London, 1997.

[32] Franken, P., Konig, D., and Arndt, U. Schmidt, V. *Queues and point processes.* Academie Verlag, Berlin, 1981.

[33] Gautam, N. *Analysis of queues: methods and applications.* CRC Press, 2012.

[34] Gebali, F. *Analysis of computer and communication networks.* Springer, New York, 2008.

[35] Gelenbe, E., and Mitrani, I. *Analysis and synthesis of computer systems.* Academic Press, London, 1980.

[36] Gelenbe, E., and Pujolle, G. *Introduction to queueing networks.* Wiley & Sons, Chichester, 1987.

[37] Gnedenko, B., Belyayev, J., and Solovyev, A. *Mathematical methods of reliability theory (in Hungarian).* Műszaki Könyvkiadó, Budapest, 1970.

[38] Gnedenko, B., Belyayev, Y., and Solovyev, A. *Mathematical methods of reliability theory.* Academic Press, New York, London, 1969.

[39] Gnedenko, B., and Kovalenko, I. *Introduction to queueing theory.* Birkhaeuser, Boston, MA, 1991.

[40] Gross, D., Shortle, J., Thompson, J., and Harris, C. *Fundamentals of queueing theory, 4th edition.* John Wiley & Sons, New York, 2008. .

[41] Györfi, L., and Páli, I. *Queueing theory in informatics systems (in Hungarian).* Műegyetemi Kiadó, Budapest, 1996.

[42] Haghighi, A., and Mishev, D. *Queueing models in industry and business.* Nova Science Publishers, Inc., New York, 2008.

[43] Haghighi, A. M., and Mishev, D. P. *Difference and differential equations with applications in Queueing theory.* John Wiley & Sons, 2013.

[44] Haghighi, A. M., and Mishev, D. P. *Delayed and network queues.* John Wiley & Sons, 2016.

[45] Hall, R. W. *Queueing methods for services and manufacturing.* Prentice Hall, Englewood Cliffs, NJ, 1991.

[46] Harchol-Balter, M. *Performance modeling and design of computer systems: queueing theory in action.* Cambridge University Press, 2013.

[47] Haribaskaran, G. *Probability, queueing theory and reliability engineering.* Laxmi Publications, Bangalore, 2006.

[48] Hassin, R. *Rational queueing.* CRC press, 2016.

[49] Haverkort, B. *Performance of computer communication systems: A model-based approach.* Wiley & Sons, New York, 1998.

[50] Haviv, M. A course in queueing theory, 2013.

[51] Hillier, F. S., and Lieberman, G. J. *Introduction to Operations Research,*. Irwin/McGraw-Hill, 2010.

[52] Hlynka, M. Queueing theory page. `http://web2.uwindsor.ca/math/hlynka/queue.html`.

[53] Ivcsenko, G., Kastanov, V., and Kovalenko, I. *Theory of queueing systems*. Nauka, Moscow, 1982.

[54] Iversen, V. *Teletraffic Engineering Handbook*. ITC in Cooperation with ITU-D SG2, 2005. `http://web2.uwindsor.ca/math/hlynka/queue.html`.

[55] Jain, R. *The art of computer systems performance analysis*. Wiley & Sons, New York, 1991.

[56] Jaiswal, N. *Priority queues*. Academic Press, New York, 1969.

[57] Jereb, L., and Telek, M. Queueing systems (in Hungarian). teaching material, BME Department of Telecommunication.

[58] Karlin, S., and Taylor, H. *Stochastic process (in Hungarian)*. Gondolat Kiadó, Budapest, 1985.

[59] Karlin, S., and Taylor, H. *An introduction to stochastic modeling*. Harcourt, New York, 1998.

[60] Khintchine, A. *Mathematical methods in the theory of queueing*. Hafner, New York, 1969.

[61] Kitaev, M. Y., and Rykov, V. V. *Controlled queueing systems*. CRC press, 1995.

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

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

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

[65] Kobayashi, H., and Mark, B. *System modeling and analysis: Foundations of system performance evaluation*. Pearson Education Inc., Upper Sadle River, 2008.

[66] Korolyuk, V., and Korolyuk, V. *Stochastic models of systems*. Kluwer Academic Publishers, Dordrecht, London, 1999.

[67] Kovalenko, I., Pegg, P., and Kuznetzov, N. *Mathematical theory of reliability of time dependent systems with practical applications*. Wiley & Sons, New York, 1997.

[68] Kulkarni, V. *Modeling, analysis, design, and control of stochastic systems.* Springer, New York, 1999.

[69] Lakatos, L., Szeidl, L., and Telek, M. *Algorithms in informatics, Vol. II (in Hungarian).* ELTE Eötvös Kiadó, 2005, ch. Queueing theory (in Hungarian), pp. 1298–1347.

[70] Lakatos, L., Szeidl, L., and Telek, M. *Introduction to queueing systems with telecommunication applications.* Springer, 2019.

[71] Lavenberg, S., e. *Computer performance modeling handbook.* Academic Press, New York, 1983.

[72] Lee, A. M. *Applied queueing theory.* Macmillan International Higher Education, 2016.

[73] Lefebvre, M. *Basic probability theory with applications.* Springer, 2009.

[74] Leon-Garcia, A. *Probability, statistics, and random processes for electrical engineering.* Pearson Education, 2017.

[75] Medhi, J. *Stochastic models in queueing theory.* Elsevier, 2002.

[76] Mieghem, P. *Performance analysis of communications networks and systems.* Cambridge University Press, Cambridge, 2006.

[77] Nelson, R. *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling.* Springer Science & Business Media, 2013.

[78] Newell, C. *Applications of queueing theory*, vol. 4. Springer Science & Business Media, 2013.

[79] Ovcharov, L., and Wentzel, E. *Applied Problems in Probability Theory.* Mir Publishers, Moscow, 1986.

[80] Palaniammal, S. *Probability and Queueing Theory.* PHI Learning Pvt. Ltd., 2011.

[81] Prabhu, N. U. *Foundations of queueing theory*, vol. 7. Springer Science & Business Media, 2012.

[82] Prékopa, A. *Probability theory (in Hungarian).* Műszaki Könyvkiadó, Budapest, 1962.

[83] Pósafalvi, A., and Sztrik, J. On the heterogeneous machine interference with limited server's availability. *European Journal of Operational Research 28* (1987), 321–328.

[84] Pósafalvi, A., and Sztrik, J. A numerical approach to the repairman problem with two different types of machines. *Journal of Operational Reseach Society 40* (1989), 797–803.

[85] Ravichandran, N. *Stochastic Methods in Reliability Theory.* John Wiley and Sons, 1990.

[86] Reimann, J. *Probability theory and statistics for engineers (in Hungarian) .* Tankönyvkiadó, Budapest, 1992.

[87] Rényi, A. *Probability theory (in Hungarian).* Tankönyvkiadó, Budapest, 1973.

[88] Ross, S. M. *Introduction to Probability Models.* Academic Press, Boston, 1989.

[89] Saaty, T. *Elements of Queueing Theory with Applications.* McGraw-Hill, 1961.

[90] 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.

[91] Sauer, C., and Chandy, K. *Computer systems performance modelling.* Prentice Hall, Englewood Cliffs, N.J., 1981.

[92] Schatte, P. On the finite population G]/M/l queue and its application to multiprogrammed computers. *Journal of lnformation Processing and Cybernetics 16* (1980), 433–441.

[93] Shortle, J. F., Thompson, J. M., Gross, D., and Harris, C. M. *Fundamentals of queueing theory*, vol. 399. John Wiley & Sons, 2018.

[94] Smith, J. M. *Introduction to Queueing Networks: Theory and Practice.* Springer, 2018.

[95] Smith, W. L. *Queues.* Chapman and Hall/CRC, 2020.

[96] Stewart, W. *Introduction to the numerical solution of Markov chains.* Princeton University Press, Princeton, 1995.

[97] Stewart, W. *Probability, Markov chains, queues, and simulation.* Princeton University Press, Princeton, 2009.

[98] Stidham, S. *Optimal design of queueing systems.* CRC Press/Taylor & Francis, 2009.

[99] Syski, R. *Introduction to Congestion Theory in Telephone Systems, 2nd Edition.* North Holland, 2005.

[100] Sztrik, J. On the finite-source $\vec{G}/m/r$ queues. *European Journal of Operational Research 20* (1985), 261–268.

[101] Sztrik, J. On the $n/G/M/1$ queue and Erlang's loss formulas. *Serdica 12* (1986), 321–331.

[102] Sztrik, J. On the $\vec{G}/M/r/FIFO$ machine interference model with state-dependent speeds. *Journal of Operational Researc Society 39* (1988), 201–201.

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

[104] Sztrik, J. *An introduction to queueing theory and its applications (in Hungarian).* Kossuth Egyetemi Kiadó, Debrecen, 2000.
http://irh.inf.unideb.hu/user/jsztrik/education/eNotes.htm.

[105] Sztrik, J. *A key to queueing theory with applications (in Hungarian).* Kossuth Egyetemi Kiadó, Debrecen, 2004.
http://irh.inf.unideb.hu/user/jsztrik/education/eNotes.htm.

[106] Sztrik, J. Practical queueing theory (in Hungarian). Teaching material, Debrecen University Egyetem, Faculty if Informatics, 2005.
http://irh.inf.unideb.hu/user/jsztrik/education/09/index.html.

[107] Sztrik, J. *Performance modeling of informatics systems (in Hungarian).* EKF Líceum Kiadó, Eger, 2007.

[108] Sztrik, J. Modeling and analyisis of information technology sytems, 2012.
http://irh.inf.unideb.hu/user/jsztrik/education/eNotes.htm.

[109] Sztrik, J. Basic Queuing Theory, Foundation of System Performance Modeling, Globe Edit. *Omni Scriptum GmbH & Co. KG* (2016).

[110] Sztrik, J. *Modeling and Analysis of Information Technology Systems.* GlobeEdit, 2016.

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

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

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

[114] Takács, L. *Introduction to the theory of queues.* Oxford University Press, New York, 1962.

[115] Takács, L. *Combinatorial Methods in the Theory of Stochastic Processes.* John Wiley & Sons, 1977.

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

[117] Tijms, H. *A first course in stochastic models.* Wiley & Son, Chichester, 2003.

[118] Tomkó, J. On sojourn times for semi-Markov processes. *Proceeding of the 14th European Meeting of Statisticians, Wroclaw* (1981).

[119] Tomkó, J. Sojourn time problems for Markov chains (in Hungarian). *Alkalmazott Matematikai Lapok* (1982), 91–106.

[120] Trivedi, K. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications, 2-nd edition.* Wiley & Son, New York, 2002.

[121] Ushakov, I. A., and Harrison, R. A. *Handbook of reliability engineering. Transl. from the Russian. Updated ed.* John Wiley & Sons, New York, NY, 1994.

[122] van Hoorn, M. *Algorithms and approximations for queueing systems.* Centrum voor Wiskunde en Informatica, Amsterdam, 1984.

[123] Virtamo, J. *Queueing Theory.* Helsinki University of Technology, 2015. `http://www.netlab.tkk.fi/opetus/s383143/kalvot/english.shtml`.

[124] Weber, T. *Solving Performance Models Based on Basic Queueing Theory Formulas.* Grin Publishing, 2017.

[125] Wentzel, E., and Ovcharov, L. *Applied problems in probabbility theory.* Mir Publisher, Moscow, 1986.

[126] White, J. *Analysis of queueing systems.* Academic Press, New York, 1975.

[127] Wolf, R. *Stochastic Modeling and the Theory of Queues.* Prentice-Hall, 1989.

[128] Yashkov, S. Processor-sharing queues: some progress in analysis. *Queueing Systems: Theory and Applications 2* (1987), 1–17.

[129] Zukerman, M. Introduction to queueing theory and stochastic teletraffic models. *arXiv preprint arXiv:1307.2968* (2013).