Nick T. Thomopoulos

Fundamentals of Queuing Systems

Statistical Methods for Analyzing Queuing Models

Springer

Fundamentals of Queuing Systems

Nick T. Thomopoulos

Fundamentals of Queuing Systems

Statistical Methods for Analyzing Queuing Models

Springer

Nick T. Thomopoulos
Stuart School of Business
Illinois Institute of Technology
Chicago, IL 60661
USA

ISBN 978-1-4614-3712-3 e-ISBN 978-1-4614-3713-0
DOI 10.1007/978-1-4614-3713-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012934665

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

For my wife, my children and my grandchildren

Preface

I was fortunate to have a rich and diverse career in industry and academia. This included working at International Harvester as supervisor of operations research in the corporate headquarters; at IIT Research Institute (IITRI) as a senior scientist with applications that spanned world-wide in industry and government; as a professor in the Industrial Engineering Department at the Illinois Institute of Technology (IIT), in the Stuart School of Business at IIT; and the many years of consulting assignments with industry and government throughout the world. At IIT, I was fortunate to be assigned a broad array of courses, gaining a wide breadth with the variety of topics, and with the added knowledge I acquired from the students, and with every repeat of the course. I also was privileged to serve as the advisor to many bright Ph.D. students as they carried on their dissertation research. Bits of knowledge from the various courses and research helped me in the classroom, and also in my consulting assignments. I used my industry knowledge in classroom lectures so the students could see how some of the textbook methodologies actually are applied in industry. At the same time, the knowledge I gained from the classroom helped me to formulate and develop solutions to industry queuing applications as they unfolded. This variety of experience allowed me to view how queuing theory is and can be used in industry. This book is based on this total experience and also includes the quantitative methods that I found doable and useful.

Thanks especially to my wife, Elaine Thomopoulos, who encouraged me to write this book, and who gave consultation whenever needed. Thanks also to the many people who have helped and inspired me over the years and some are former IIT students from my queuing classes. I can name only a few here. Raida Abuizam (Purdue University—Calumet), Bob Allen (R. R. Donnelly), Deepak Bammi (Bammi Associates), Wayne Bancroft (Walgreens), Harry Bock (Florsheim Shoe Company), Debbie Cernauskas (Benedictine University), Edine Dahel (Monterey Institute), Ahmed El Melegy (Cairo University), Tom Galvin (Northern Illinois University), Ranko Glisic (IIT), John Garofalakis (Patras University), Tom Georginis (Lewis University), Shail Godambe (Motorola, Northern Illinois University), M. Zia Hassan (Illinois Institute of Technology), Willard Huson

(Navistar), Robert Janc (IIT Research Institute), Marsha Jance (Indiana University—Richmond), Chuck Jones (Illinois Institute of Technology), Arvid Johnson (Domenican University), Montira Jantaravareerat (IIT), Tom Knowles (Illinois Institute of Technology), Joachim Lauer (Northern Illinois University), Carol Lindee (Panduit), Nick Malham (FIC Inc.), Barry Marks (IIT Research Institute), Peter McManamon (IIT Research Institute), Fotis Mouzakis (Cass Business School of London), Pissanu Manaspiti (Rangsit University), Pricha Pantumsinchai (M-Focus), Noln Plumchitchom (IIT), Ted Prenting (Marist College), Athapol Ruangkanjanases (Marist College), Walter Ryder (University of Southern California), Hendrarto Supangkat (IIT), Ornlatcha Sivarak (Mahidol University), Spencer Smith (Illinois Institute of Technology), Mark Spieglan (FIC Inc.), Paul Spirakis (Patras University), Tongsakorn Vaivong (IIT), Reino Warren (University of Michigan—Flint) and Colleen Wilder (Valparaiso University).

Nick T. Thomopoulos

Fundamentals of Queuing Systems

Fundamentals of Queuing Systems describes the methods used to measure the probabilities and statistics for a wide variety of queuing systems. The material is timeless and the book will never become obsolete. The systems include infinite and finite arrival populations, single and multiple service facilities, and queues that are infinite, finite and none at all. Arrival times that are exponential and Erlang, and service times that are exponential, constant, Erlang and arbitrary. The book includes priority disciplines, 2 input populations, tandem systems, repeat service, waiting time densities for single and multi server systems, and matrix solution methods. The book introduces the concept of reusable inventory, service level, how to use reduced equations, and how to apply matrix solutions to approximate infinite queues. The book presents the basic topics that people want and should know in the work place. The presentation is easy to read for students and practitioners and there is little need to delve into difficult mathematical relationships. Numerical examples are presented to guide the reader on applications. Practitioners will be able to apply the methods learned to designing queuing systems in industry and government that even reach beyond this book. The typical worker will want the book on their bookshelf for reference when needed. The potential market is vast. It includes everyone in professional organizations like IEEE, DSI and INFORMS, people in industry, and students in management science, industrial engineering, electrical engineering and computer engineering.

Nick T. Thomopoulos has degrees in business (B.S.) and in mathematics (M.A.) from the University of Illinois, and in industrial engineering (Ph.D.) from Illinois Institute of Technology. He was supervisor of operations research at International Harvester, senior scientist at IIT Research Institute, and is a professor emeritus at Illinois Institute of Technology. He is the co-author of *Assembly Line Systems*, Hayden Books, (1974), author of *Applied Forecasting Methods*, Prentice Hall (1980), *Inventory Management and Planning,* Hitchcock Publishing Company (1990), and *Quantitative Methods along the Supply Chain*, Atlantic Publishers and Distributors (2011). He has published numerous papers, and for many years, he has consulted in a wide variety of industries in the United States, Europe and Asia. Nick has received honors over the years, such as the *Rist Prize* in

1972 from the Military Operations Research Society for new developments in queuing theory, the *Distinguished Professor Award* in Bangkok, Thailand in 2005 from the IIT Asian Alumni Association, and the *Professional Achievement Award* in 2009 from the IIT Alumni Association.

Nick T. Thomopoulos
Professor Emeritus
Illinois Institute of Technology

Contents

Chapter 1
Introduction

Abstract This chapter provides a quick summary of the contents in each of the remaining chapters. Also included is an early history on queuing theory, and a large list of examples.

1.1 Introduction

Queuing theory is a form of probability that pertains to the study of waiting lines (queues). This is for a system with a steady inflow of units (customers) and a specified number of servers (service facilities). The analyst wants to know if the number of service facilities in the system is adequate to handle the inflow of demands. The goal is to calculate various performance measures of the system. These include the probability a server is immediately available to a new arrival, the average number of units in the queue, in the system, and the corresponding times in the queue and system.

The word *queue* comes from the French interpretation of Latin c*auda*, meaning a tail. According the Funk and Wagnall's New International Dictionary, a queue is "a line of persons or vehicles waiting in the order of their arrival." The word queue is the common way to refer to a line in England.

1.2 The Queuing System

A typical queuing system includes the following components:

Input population = the source of units that become the customers to the system.
Arrivals = the units from the population that enter the system seeking service.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_1, © Springer Science+Business Media New York 2012

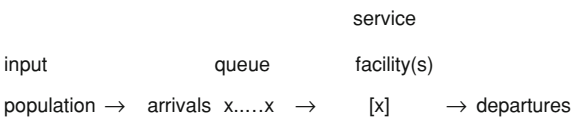

Fig. 1.1 A typical queuing system

Queue = the line that houses the units that are awaiting their turn to be serviced.
Service facilities = the place where the units are processed.
Departures = the units that have completed their service and leave the system.

A depiction of the queuing system is below in Fig. 1.1.

1.3 Early Literature

Agner Krarup Erlang (1878–1929), a Danish mathematician, invented the fields of traffic engineering and queuing theory starting in the 1900s. While working for the Copenhagen Telephone Company, he was confronted with the classic problem of determining how many circuits were needed to provide an acceptable telephone service. He formed the mathematical way of determining how many telephone operators were needed to handle a given volume of calls. He is the founder on the theory of telephone traffic and over his career, he published papers, starting in 1909, that became the foundation of queuing theory. He also developed the Erlang probability distribution, which plays a significant role in various queuing applications.

Queueing theory is now an important branch of operations research and has many applications. It measures the flow of demands into and out of the queuing system, and thereby is used to make decisions on the minimum number of resources needed. Queuing theory is used in business, engineering, public service, traffic, healthcare, finance and the military. A vast number of applications in all fields have been implemented and published since Erlang. Only a few are named here.

In 1953, David G. Kendall introduced Kendall's notation to describe the characteristics of a queuing system. This A/B/C notation is standard in queuing theory. The A/B/C code identifies a system where: A is the arrival time distribution, B is the service time distribution, and C is the number of servers.

In 1961, Thomas L. Saaty, authored one of the first comprehensive books on queuing theory. Another early and informative publication was by Phillip M. Morse in 1958.

In the 1960s, Leornard Kleinrock used queuing theory to applications on packed switching networks. His developments have evolved as the foundation in the birth of the Internet. In 1969, his Host computer became the first node of the Internet, and it was from there that he directed the transmission of the first message to pass over the Internet.

In 1990, Thomopoulos' book "Strategic Inventory Management and Planning" included many tables with measurements from a wide variety of queuing systems. The book introduced the concept of reusable inventory. The servers in the queuing systems can be thought of as reusable inventory that are used to fill the customer demands and do not leave the firm's possession once a demand is fulfilled. This type of inventory demand takes place when a demand occurs for use of the item and, and upon completion, the item remains to meet the next demand. When a demand cannot immediately enter the service facility, the demand is either in a backorder state or is a lost sale.

1.4 Some Applications

Applications in queuing theory are vast and vital. Through the use of queuing theory, management can design a system that runs smoothly and efficiently, with minimum waiting time for the customers and minimum idle time for the facility. Various applications in the use of queuing theory follow:

Backorder Applications: If all of the reusable inventory items are occupied when a demand arrives, and if the demand will or can wait, in essence the demand enters the queue and is in a backorder state. These are systems with an infinite queue length, or a finite queue length when space in the queue is still available. The common goal is to determine the number of service facilities to have available to efficiently service the customers with minimal waiting time. Some examples are:

- In a manufacturing plant, forklift trucks are used in running the daily operations. As each forklift truck need arises, the next available forklift truck performs the task and upon completion, awaits the next task, and in this way, the forklift trucks are the service facilities.
- In a similar way, specialized tools, fixtures and machines are needed to run the manufacturing operation and they then become the service facilities for the system.
- In a shoe factory, a large (and expensive) inventory of molds (called lasts) are needed in the manufacturing process. Each last is dedicated to a specific pair of shoes (by style and size). A pair of lasts remain in a pair of shoes for about two days in manufacturing, and thereby a large inventory of lasts are needed. An important decision for the management is to determine the composition of lasts (by style and size) to have in the plant inventory to allow the shoe scheduling to carryon in an efficient manner. The lasts become the service facility items in the plant.
- In distribution centers, examples of reusable inventory items (service facilites) are the binding machines, forklift trucks, receiving docks, shipping docks and order picking personnel.
- In a service repair shop, the service facilities (reusable items) are the specialized tools, fixtures and operators.

- In an office, examples of service facilities are fax machines, computer terminals, copy machines and printers, as well as the operators who repair and maintain these items.
- In retail locations (dealers and stores), the reusable items (service facilities) include the checkout counters, sales clerks, gas pumps in a gas station, push carts (in a grocery store) and tables (in a restaurant).

Lost Sales Applications: If a customer arrives to a system when all the service facilites are occupied, and if the customer cannot or will not wait, the system is classified as a lost sale state. This would be a system with no queue, or with a finite queue space when the queue is full. A common goal is to determine the number of service facilities to have available to minimize the number of lost customers. Various examples are listed below:

- In the event all the pump locations in a gas station are occurpied with cars, and when new customers will not wait for an empty pump, the potential customers to the station become lost customers. In this situation, the gas pumps are the service facilities.
- A restaurant will lose potential customers when all tables are occupied, and when new customers will not wait for an empty table and go elsewhere. The tables are the service facilities in this system.
- A rental agency loses a sale when a potential customer finds that all units of the item sought are leased out, and the customer goes elsewhere for the item. Each of the rental items in the agency become service facilities.
- A sales office with a limited number of phone lines may lose potential customers that call the office when all of the lines are busy. The office manager may wonder how many lines to have available to handle all the potential calls.
- In a car dealership, the typical customer insists on a loaner car in order to leave his/her auto at the facility for repair. The loaner cars become the service facilities at the dealership.

Other Applications: More uses of queuing theory are described below.

- A city is partitioned into a finite number of patrol beats that are designed from a contiguous set of blocks so that one patrol car can efficiently service the expected number of calls in the beat. The projected number of calls by block are summed to determine the projected calls for the patrol beat. The number of beats and the beat configurations usually vary by hour of day, day of week and month of year.
- On an expressway, the number of tollbooths to have open by day of week and hour of day are scheduled to minimize delays for the incoming traffic, and also to minimize idle time of the tollbooth operators.
- The scheduling frequency of aircraft in and out of an airport depends on the number of runways available. The concern at the airport is to minimize the wait time for the arriving (in the air) aircraft, and the departing (on the ground) aircraft. The runways of the airport are the service facilities, and so also are the airport controllers, and all the auxiliary crews that service arrival and/or departure flights.

- In a windshield manufacuting plant, an inventory of molds for each car model and year is used in the manufacturing process. One mold is needed to produce one windshield. The molds are expensive and take up much space. A forecast (on new cars and on cars that need replacements for damaged windshields) projects the demands (by car model and year) for the future time horizon. The forecast is needed to determine how many of each mold (by car model and year) to have in the plant inventory to efficiently run the windshield manufacturing operation.

More Applications: Below lists some more application of queuing theory.

- A bank wants to determine how many teller booths to have open to service the customers by hour of the day, and by day of week.
- Airport management wants to know how many crews to have available at an airport to maintain and clean the just-arrived aircraft for ready status as a departing aircraft.
- An ambulance service facility seeks to know how many ambulance crews to have in its district to meet the calls for service.
- The management of a distribution center wants to know how to allocate a fixed number of receiving docks by category of incoming trucks such as: truck-load, less-truck-load, UPS, local delivery, and so forth. The goal is to minimize the incoming trucks idle time in the yard awaiting their turn to the receiving docks.
- A military commander wants to know the number of medics to have available in a combat setting in order to reduce the time to service the wounded combatants.
- An architect inquires on the number and size of elevators in a multi-story office building to accommodate the day and hourly flow of people.
- An architect needs to know how many washrooms to include in the design for a ballpark, and further, how many stools, urinals and sinks should be available within each washroom facility.
- A manager of a large grocery store wants to know how many employees to assign to the delicatessen counter to service the customers.
- A postal manager seeks to determine the number of postal windows to have open in a post office by day and hour of the week, to minimize the wait time of the arriving customers.
- An architect inquires how many parking spaces to have available in a shopping center to efficiently service the arriving customers.
- A military logistics officer wants to know how many radio frequencies in their comunication system are needed for a fleet of ships to allow operators to send and receive messages with minimal delay time across a series of networks that share the radio frequencies. The radio frequencies become the service facilities.
- The military logistics officer also seeks how to allocate the radio frequencies to the various networks in the fleet of ships.
- A military logistics officer officer wants to know how many repair stations should be available to service the key equpment in a military operation.

1.5 Chapter Summaries

This book provides solutions to a wide variety of queuing systems. The following is a quick summary of the contents in each of the remaining chapters. By reviewing the queuing systems of this book, and following the methods of solution, the reader should be able to expand the methods to a wider spectrum of systems than are shown here.

Chapter 2 gives a summary on some of the key mathematical and probability concepts that are needed as a foundation for the remaining chapters. The chapter introduces the concepts that are used in the subsequent text so that they do not need to be repeated throughout the book. This includes a definition of the Poisson, Exponential and Erlang distributions and how they are related to each other. The chapter also lists the Postulates that are needed to define a queuing system. The postulates are used to identify a particular queuing system by way of difference equations. The difference equations yield the differential and equilibrium equations and finally the reduced equations. The equilibrium and/or the reduced equations are needed to generate the probability distribution on the number of units in the system, and then the various performance measures.

Chapters 3–5 describe systems where one service facility is in place. The input and output times are exponential. These chapters concern systems with an infinite queue, a finite queue and with no queue. An infinite queue example could be the airline passengers arriving to a security checkpoint in the airport. The checkpoint is the service facility and the passengers are the arrivals. A finite queue example is a one-man barbershop with three seats for the waiting customers. A no queue example is a rental store with one electric saw available for rental customers to check out. The saw is the service facility and the rental time becomes the service time. When the saw is out, future customers will not wait and go elsewhere.

Chapters 6–8 pertain to systems with a multiple number of service facilities. The input and output times are exponential. They are for systems with an infinite queue, a finite queue and no queue. An infinite queue example could be the cars on an expressway arriving to a toll center with three tollbooths. A finite queue example is a beauty shop with two hair stylists and with room for only five customers in the shop. A no queue example is a phone system in a real estate company with a five lines available to receive calls. When all lines are busy, any new call is lost.

Chapters 9 and 10 show how to analyze a one server system when the service times are from an arbitrary distribution. An example is a lift truck in a warehouse that hauls stock from the receiving dock to the storage area where the hauling time is normally distributed and not exponential. The lift truck is the service facility. Another example are the calls for service to a squad car in a one car patrol beat, where some calls are for minor scrapes and others are major incidents and the combined service times are not exponential.

Chapters 11 and 12 pertain to systems that have a limited number of units in the input population. These may be M machines in a shop that occasionally require service from one or R repairmen. This could be a firm with five copy machines and

one repairman. Another example is a taxi fleet of 100 cabs with four service mechanics on duty to maintain and repair the cabs as needed.

Chapters 13 and 14 describe systems where the service of a unit may have to be repeated. These are for systems that have one server, and that have multi servers, respectively. An example of the former may be a one operator machine shop fabrication of a fixture that is tested at the end to see if it passes a strength test. If not, another fixture must be fabricated. An example of the latter is a warehouse with several order pickers that receive customer orders. When an order is picked incorrectly it must be repeated.

Chapter 15 introduces a series of two systems where the arriving units goes from one system to another in a tandem way to receive processing. An example would be the patients arriving to a medical center where the first system is filling out the paper and insurance forms, and the second system is getting the medical attention.

Chapters 16 and 17 describe systems where the service discipline behaves in a preemptive priority way. The systems described are for exponential and arbitrary service times. An example could be a military unit using a one-frequency radio system where the top commander could interrupt any ongoing call whenever needed. Another example concerns the patients coming to an emergency clinic where some need immediate emergency treatment and others do not. The emergency patients override the non-emergency patients.

Chapter 18 shows how to analyze a system when the service time is constant. An example would be cars arriving to a carwash where the service time is always the same.

Chapters 19–21 describe systems when one or both of the arrival and service times are Erlang distributed. An example would be a jogging shoe manufacturer that uses a mold (called a last) to produce a shoe of a certain size and width. The arrival time between needs for the mold is exponential, and the time to use the mold on the shoe is Erlang. An example of Erlang arrivals and exponential service may be trucks that arrive to a receiving dock with one unloading crew. As the trucks come in, the crew unloads each truck in the order of arrival. Here, the crew is the service facility. An example of Erlang arrivals and service is a furniture store where, on each sale, the store has a stockman who fetches the item in the back storage area of the store, brings it to the customer's vehicle and helps to load the item in the vehicle. In this situation, the stockman is the service facility.

Chapters 22 and 23 show how to derive the waiting time density for a one server system and for a multi server system, respectively. An example with one service facility is when a moderate size city designs a beat for a squad car and wants to determine the probability that at least 90 percent of the calls received for the beat can begin service before 10 minutes. The squad car is the service facility and the calls within the beat are the arrivals. A multiple service facility example is when a package delivery service wants to determine the number of delivery vehicles to have in its fleet so the probability that a delivery begins within 20 minutes of the call. For a given number of vehicles, the probability is measured. If the probability is too high, another vehicle is added and the probability is again measured.

Bibliography

Erlang, A. K. (1909). Probability and Telephone Calls. Nyt Tidsskrift Matematik Series B, *20*, 33–39.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics*, *24*(3), 338–354.

Kleinrock, L. (1964). *Communication nets: stochastic message flow and design.* New York: McGraw-Hill.

Morse, P. M. (1958). Queues, inventories and maintenance. New York: Wiley.

Saaty, T. L. (1961). Elements of queueing theory with applications. New York: McGraw Hill.

Thomopoulos, N. T. (1990). Strategic inventory management and planning. Carol Stream: Hitchcock Publishing Co.

Chapter 2
Preliminary Concepts

Abstract This chapter describes the mathematical and probability concepts that are the foundation for the remaining chapters of the book. This includes the Poisson, Exponential and Erlang distributions, the Postulates that define a queuing system, and also the difference, differential, equilbrium and reduced equations. The equilibrium and/or the reduced equations are used to generate the probability distribution on the number of units in the system, and the various performance measures of the system.

2.1 Introduction

This chapter gives an overview on some of the key mathematical and probability concepts that are used in queuing theory. The chapter introduces the concepts that are used in the subsequent text so that they do not need to be repeated throughout the book. This includes a definition of the Poisson, Exponential and Erlang distributions and how they are related to each other. The chapter also lists the Postulates that are needed to define a queuing system. The postulates are used to identify a particular queuing system by way of difference equations. The difference equations yield the differential and equilibrium equations and finally the reduced equations. The equilibrium and/or the reduced equations are needed to generate the probability distribution of n units in the system, and then the various performance measures.

2.2 Some Useful Relations

Some of the identities that are used in developing the queuing models are listed here. Equations 2.1–2.3 are identities of infinite sums that apply when $0 < \theta < 1$. Equations 2.4–2.8 are identities that concern finite sums, and Eq. 2.9 is an identity that pertains to the exponent term.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_2, © Springer Science+Business Media New York 2012

$$\sum_{k \geq 0} \theta^k = 1/(1-\theta) \tag{2.1}$$

$$\sum_{k \geq 0} k\theta^k = \theta/(1-\theta)^2 \tag{2.2}$$

$$\sum_{k \geq 0} k^2\theta^k = \theta(1+\theta)/(1-\theta)^3 \tag{2.3}$$

$$\sum_{k=1}^{N} 1 = \mathrm{N} \tag{2.4}$$

$$\sum_{k=1}^{N} k = \mathrm{N}(\mathrm{N}+1)/2 \tag{2.5}$$

$$\sum_{k=1}^{N} k^2 = \mathrm{N}(\mathrm{N}+1)(2\mathrm{N}+1)/6 \tag{2.6}$$

$$\sum_{k=0}^{N} x^k = \left(1 - \mathrm{x}^{\mathrm{N}+1}\right)/(1-\mathrm{x}) \quad \mathrm{x} \neq 1 \tag{2.7}$$

$$\sum_{k=0}^{N} kx^k = \mathrm{x}\left[1 - (\mathrm{N}+1)\mathrm{x}^{\mathrm{N}} + \mathrm{N}\mathrm{x}^{\mathrm{N}+1}\right]/(1-\mathrm{x})^2 \quad \mathrm{x} \neq 1 \tag{2.8}$$

$$\mathrm{e}^{\mathrm{ax}} = \sum_{k \geq 0} (ax)^k/\mathrm{k}! \tag{2.9}$$

2.3 Exponential Distribution

Consider a random variable t that is continuous with $t \geq 0$ and follows the exponential distribution. The probability density of t is $f(t) = \theta e^{-\theta t,}$ and the corresponding cumulative distribution is $F(t) = 1 - e^{-\theta t}$. For the exponential variable t, the expected value and variance are $E(t) = 1/\theta$, and $V(t) = 1/\theta^2$, respectively.

2.4 Poisson Distribution

The Poisson probability distribution has a discrete variable n where $n = 0, 1, 2, \ldots$. The probability distribution of n is $P_n = \theta^n e^{-\theta}/n!$. The expected value and variance of n are $E(n) = \theta$, and $V(n) = \theta$, respectively.

The Poisson distribution can also be defined in units of time t. In this situation, the discrete variable n represents the number of occurrences in time t. The probability of n units in time t becomes,

$P(n,t) = (\theta t)^n e^{-(\theta t)}/n!$.

2.5 Relation Between the Exponential and Poisson Distributions

The Poisson distribution and the exponential distribution are related as shown here. Recall the exponential probability density is

$$f(t) = \theta e^{-\theta t}$$

Suppose τ is exponential with expected value $1/\theta$, and n is Poisson with mean θ. From the exponential,

$$\begin{aligned} P(\tau > t) &= 1 - F(t) \\ &= e^{-\theta t} \\ &= P(n = 0 \text{ in } t) \\ &= P(0, t) \end{aligned}$$

where the latter is Poisson. Also, note below, where the probability of n units in time t, P(n,t), becomes Poisson.

$$P(0, t) = e^{-\theta t}$$

$$P(1, t) = \int_{\tau=0}^{t} P(0, \tau)\ f(t - \tau) d\tau = \theta t e^{-\theta t}$$

$$P(2, t) = \int_{\tau=0}^{t} P(1, \tau)\ f(t - \tau) d\tau = (\theta t)^2 e^{-\theta t}/2!$$

$$\cdots$$

$$P(n, t) = \int_{\tau=0}^{t} P(n - 1, \tau)\ f(t - \tau) d\tau = (\theta t)^n e^{-\theta t}/n!$$

In the following discussion, θ is replaced with λ for arrival times. So when the arrivals to a system are exponential with an average time of $1/\lambda$ the number of units that arrive to the system in a unit of time is Poisson distributed with an average of λ. In the same way, if the number of units that arrive to a system is Poisson with parameter λ, the time between arrivals is exponential with an average of $1/\lambda$.

In the following discussion, θ is replaced with μ for service times. Hence, when the time to process the units is exponential and the average service time is $1/\mu$, the number of units that are serviced, during a continuously busy span of time, is Poisson distributed with an average of μ in a unit of time. If the units coming out of a continuously busy service facility is Poisson with a parameter of μ, the time to service the units are exponential with an average of $1/\mu$.

2.6 Convolution of Two Poisson Variables

Consider two Poisson random variables, x_1 and x_2, with parameters θ_1 and θ_2, respectively. Now assume another variable $x = x_1 + x_2$ is formed. Note the convolution below.

$$\begin{aligned} P(x) &= \sum_{x1=0}^{x} P(x_1)P(x - x_1) \\ &= \sum_{x1=0}^{x} \left[e^{-\theta 1}\theta_1^{x1}/x_1!\right]\left[e^{-\theta 2}\theta_2^{x-x1}/(x - x_1)!\right] \\ &= e^{-(\theta 1+\theta 2)}\theta_2^{x} \sum_{x1=0}^{x} (\theta_1/\theta_2)^{x1}/[x_1!(x - x_1)!] \\ &= e^{-(\theta 1+\theta 2)}(\theta_1 + \theta_2)^{x}/x! \end{aligned}$$

Thus, x is also Poisson with parameter $(\theta_1 + \theta_2)$.

2.7 Erlang Distribution

In some queuing systems, the time associated with arrivals and service times is assumed as an Erlang continuous random variable. The Erlang variable has two parameters, θ and k. The parameter k represents the number of exponential variables that are summed together to form the Erlang variable. Note, if y is an exponential variable with $E(y) = 1/\theta$, and x is the sum of k y's, then

$$x = (y_1 + \ldots + y_k),$$

and the expected value of x becomes,

$$E(x) = kE(y) = k/\theta.$$

Further, the variance of x, denoted as V(x), is derived from adding k variances of y, V(y), as below:

$$V(x) = kV(y) = k/\theta^2$$

Note, when $k = 1$, the Erlang variable is the same as an exponential variable where the mode is zero and the density is skewed to the right. As k increases, the mode moves further away from zero and becomes less skewed to the right. As k increases, the shape of the Erlang density starts to resemble a normal density, via the central limit theorem.

2.8 Memory-Less Property of the Exponential Distribution

Recall, when a random variable t is exponential, the probability density is

$$f(t) = \theta e^{-\theta t}$$

and the cumulative distribution is

$$F(t) = 1 - e^{-\theta t}$$

For a time increment h, the probability that t is larger than h becomes

$$P(t > h) = e^{-\theta h}$$

At $t = (t' + h)$, the probability t is larger than $(t' + h)$ is

$$P(t > t' + h) = e^{-\theta(t' + h)}$$

The conditional probability of $t > (t' + h)$ given $t > t'$ is

$$P(t > t' + h | t > t') = e^{-\theta(t' + h)}/e^{-\theta t'} = e^{-\theta h}$$

Note the probabilities $P(t > t' + h | t > t')$ and $P(t > h)$ are the same, i.e.,

$$P(t > t' + h | t > t') = P(t > h) = e^{-\theta h}$$

Because the two probabilities are the same, the exponential distribution is called a memory-less probability distribution.

2.9 Cumulative Distribution for a Small Increment h

Consider time t that follows the exponential distribution, and observe, for a particular time increment h, the cumulative distribution of h becomes $F(h) = 1 - e^{-\theta h}$. Note the expression for F(h) can be converted using Eq. (2.9) above in the following way.

$$\begin{aligned} F(h) = P(t<h) &= 1 - e^{-\theta h} \\ &= 1 - [(-\theta h)^0/0! + (-\theta h)^1/1! + (-\theta h)^2/2! + \ldots] \\ &= 1 - [1 + (-\theta h)^1/1! + (-\theta h)^2/2! + \ldots] \\ &= \theta h - [(-\theta h)^2/2! + (-\theta h)^3/3! + \ldots] \\ &= \theta h + o(h) \end{aligned}$$

where

$$o(h) = -[(-\theta h)^2/2! + (-\theta h)^3/3! + \ldots]$$

Note o(h) is a function that approaches zero faster than h. That is

$$\underset{h \to 0}{Lim}\{o(h)/h\} = 0$$

Thereby, as h approaches zero, o(h) also approaches zero. This expression concerning the probability distribution of h is applied subsequently to define the postulates in the queuing analysis.

2.10 Probability Postulates

Assume a queuing system where the arrivals follow an exponential distribution and the average time between arrivals is $1/\lambda$. As shown above, the probability that the time between two arrivals is h or less becomes $[\lambda h + o(h)]$. Also, we assume the service time follows an exponential distribution with an average service time of $1/\lambda$. Hence, the probability is $[\mu h + o(h)]$ that the service time is less than h. Also consider the two events: A = event of an arrival in time interval h, and D = event of a departure in time interval h.

Now note the probabilities listed below that concern the events of A and D during the time interval from t to t + h, and denoted here as (t, t + h). Recall, h approaches zero.

P[A in (t,t + h)] = $[\lambda h + o(h)]$
P[D in (t,t + h)] = $[\mu h + o(h)]$
P[neither A or D in (t,t + h)] = $[1 - \lambda h - o(h)][1 - \mu h - o(h)] = [1 - \lambda h - \mu h + o(h)]$
P[2 or more A and/or D in (t,t + h)] = $o(h)$

These four probabilities are the postulates that define most of the queuing systems that follow.

2.11 Difference Equations

Consider a queuing system with one service facility, infinite queue length, with exponential arrival times with an average of $1/\lambda$ and exponential service times with an average of $1/\mu$. The difference equations specify how the system operates. This is the first step to define a queuing system. The difference equations specify how the probability of n units in the system may change as time goes from t to (t + h), denoted as (t,t + h), and where h is a very small increment of time. The number of units n in the system at any time period are integers of $n \geq 0$. As described earlier, o(h) is a function that approaches zero faster than h. The difference equations are below:

$n = 0 \qquad P_0(t + h) = (1 - \lambda h)P_0(t) + \mu h P_1(t) + o(h)$
$n \geq 1 \qquad P_n(t + h) = (1 - \lambda h - \mu h)P_n(t) + \lambda h P_{n-1}(t) + \mu h P_{n+1}(t) + o(h)$

2.12 Differential Equations

Differential equations are obtained from the difference equations when the time increment h approaches zero. They are needed in an interim manner to subsequently yield the equilibrium equations. To convert, the three identities listed below are applied. The first shows how the derivative is formed. The second expresses the probability without the increment of h, and the third concerns the function o(h).

$$\underset{h\to 0}{Lim}\{[P_n(t+h) - P_n(t)]/h\} = P_n(t)'$$
$$\underset{h\to 0}{Lim}\{[\lambda h + \mu h)P_n(t)]/h\} = [(\lambda + \mu)P_n(t)]$$
$$\underset{h\to 0}{Lim}\{o(h)/h\} = 0$$

Thus, as h approaches zero in the difference equations, the following set of the differential equations evolve:

$$n = 0 \qquad P_0(t)' = (-\lambda)P_0(t) + \mu P_1(t)$$
$$n \geq 1 \qquad P_n(t)' = (-\lambda-\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)$$

2.13 Equilibrium Equations

The equilibrium equations are obtained by studying what happens to the differential equations when time t approaches infinity under equilibrium conditions. The two identities below are used in this process.

$$\underset{t\to\infty}{Lim}\{P_n(t)'\} = 0$$
$$\underset{t\to\infty}{Lim}\{P_n(t)\} = P_n$$

Applying the above identities to the differential equations yields the following equilibrium equations:

$$n = 0 \qquad 0 = -\lambda P_0 + \mu P_1$$
$$n \geq 1 \qquad 0 = -(\lambda + \mu)P_n + \lambda P_{n-1} + \mu P_{n+1}$$

2.14 Reduced Equations

Algebra is needed at this point to transform the equilibrium equations to reduced equations as is shown here. Note below where the equilibrium equations for n = 0, 1, 2, say, are listed on the left-hand-side, and the corresponding reduced equations are on the right-hand-side. When n = 0, the equilibrium equation and the associated reduced equation are the same. For n ≥ 1, the reduced equation for n is derived from the corresponding (n) equilibrium equation and the (n − 1) reduced equation.

$$
\begin{array}{lll}
n = 0 & 0 = -\lambda P_0 + \mu P_1 & \Rightarrow \quad 0 = -\lambda P_0 + \mu P_1 \\
n = 1 & 0 = -(\lambda + \mu)P_1 + \lambda P_0 + \mu P_2 & \Rightarrow \quad 0 = -\lambda P_1 + \mu P_2 \\
n = 2 & 0 = -(\lambda + \mu)P_2 + \lambda P_1 + \mu P_3 & \Rightarrow \quad 0 = -\lambda P_2 + \mu P_3
\end{array}
$$

The general form for the reduced equations becomes the following:

$$0 = -\lambda P_{n-1} + \mu P_n \qquad n \geq 1$$

2.15 Probability of n Units in the System (P_n)

The common notation in queuing is to use n as the number of units in the system at an arbitrary moment in time. In this way, n is discrete where n is zero or larger. One measure of interest in studying queuing systems is the probability of n units in the system, and this is denoted as P_n for $n \geq 0$.

2.16 Performance Measures

Some of the other measures of interest in queuing systems are listed below:

P_o = probability the system is empty
Ls = expected number of units in the service facility
Lq = expected number of units in the queue
L = expected number of units in the system
Ws = expected time in the service facility
Wq = expected time in the queue
W = expected time in the system
Wq' = expected time in the queue given the arrival is delayed
SL = service level = probability the arrival is not delayed in the queue
Ploss = probability an arrival is lost (does not enter the system)

2.17 Wait Time in Queue Given a Delay (Wq')

Using conditional expectation notation, $Wq' = W_{q|D}$ where D is the event that the arrival is delayed waiting in the queue before being serviced. Using the same notation, D' = event the arrival is not delayed. In general,

$W_{q|D}$ = wait time in queue given delay
$W_{q|D'}$ = wait time in queue given no delay
and
P(D) = probability of a delay
P(D′) = probability of no delay
The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:
$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$
Since, $W_{q|D'} = 0$,
$Wq' = W_{q|D} = Wq/P(D)$

2.18 Little's Law

In 1961, John Little published a paper showing that the expected number of units in the system, L, is related to the expected time in the system, W, by $L = \lambda W$, as long as the arrival rate λ is constant. In the same way, the following three relations are established:

$L = \lambda W$ = expected number of units in the system
$Ls = \lambda Ws$ = expected number of units in the service facility
$Lq = \lambda Wq$ = expected number of units in the queue

Using Little's Law,

$W = L/\lambda$ = expected time in the system
$Ws = Ls/\lambda$ = expected time in the service facility
$Wq = Lq/\lambda$ = expected time in the queue

2.19 Kendall's Notation

In queuing theory, Kendall's notation is the standard way to describe and classify the queuing systems. This method of classifying the systems was first suggested by D. G. Kendall in 1953 as a three-factor *A/B/C* notation system for identifying queues. It has since been extended to include up to six different factors.

The 3 factor notation (A/B/C) signifies the following:

A = arrival process
B = service time process
C = number servers

The six (6) factors (A/B/C/K/N/D) go even further where the latter three factors denote the following:

K = number places in system	(assume K = infinity unless specify other)
N = calling population	(assume N = infinity unless specify other)
D = service discipline	(assume non-priority unless specify other)

The arrival and service time factors (A,B)are denoted as below:

M = Markovian (Poisson or Exponential)
D = deterministic
Ek = Erlang with k stages
G = general

The service discipline (D)may take on the notation given below:

FIFO = first-in first-out
LIFO = last-in first-out
Random
Priority = preemptive or non-preemptive

Bibliography

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics*, *24*(3), 338–354.

Little, J. D. C. (1961). A proof of the queuing formula L = λW. *Operations Research*, *9*(3), 383–387.

Chapter 3
One Server, Infinite Queue (M/M/1)

Abstract This chapter pertains to a one-server, infinite capacity system with exponential inter-arrival and service times. Could be the airline passengers arriving to a security checkpoint in the airport. The checkpoint is the service facility and the passengers are the arrivals. The difference, equilibrium and reduced equations are listed. The probability on n units in the system, and the corresponding performance measures are developed. Examples are presented to guide the reader.

3.1 Introduction

Consider a system with one server and an infinite queue where the inter-arrival and the service times have exponential probability densities. The average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This could be cars arriving to a drive-through lane at a fast-food restaurant during the morning hours. The following notation applies here:

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$\rho < 1$ is needed to assure the system is in equilibrium
$n =$ number of units in the system $\quad (n \geq 0)$

Below is a list of the difference equations. Following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_3,

3.2 Difference Equations

$$n = 0 \; P_0(t+h) = (1-\lambda h)P_0(t) + \mu h P_1(t) + o(h)$$
$$n \geq 1 \quad P_n(t+h) = (1-\lambda h - \mu h)P_n(t) + \lambda h P_{n-1}(t) + \mu h P_{n+1}(t) + o(h)$$

3.3 Equilibrium Equations

$$n = 0 \quad 0 = -\lambda P_0 + \mu P_1$$
$$n \geq 1 \quad 0 = -(\lambda + \mu)P_n + \lambda P_{n-1} + \mu P_{n+1}$$

3.4 Reduced Equations

$$0 = -\lambda P_{n-1} + \mu P_n \quad n \geq 1$$

3.5 Probability on n Units in the System

Using the reduced equations and the notation $\rho = \lambda/\mu$, the probability of n units in the system becomes.

$$P_n = \lambda/\mu P_{n-1} = \rho P_{n-1} \quad n \geq 1$$

It is observed that

$$P_0 = \quad \rho^0 P_0$$
$$P_1 = \rho P_0 = \quad \rho^1 P_0$$
$$P_2 = \rho P_1 = \quad \rho^2 P_0$$

and so forth, whereby,

$$P_n = \rho^n P_o \quad n \geq 0$$

Because all the probabilities sum to unity,

$$\sum_{n \geq 0} P_n = P_o \sum_{n \geq 0} \rho^n = 1$$

Recall where $\rho < 1$ because of equilibrium. This allows applying (2.1) to the above relation to yield,

$$P_o \sum_{n \geq 0} \rho^n = P_o 1/(1-\rho) = 1$$

Thereby,

$$P_0 = (1-\rho).$$

Finally, the probability of n units in the system becomes

$$P_n = \rho^n(1-\rho) \quad n \geq 0$$

3.6 Probability the System is Idle

The probability the system is empty is merely the probability that n = 0, i.e.,

$$P_o = (1-\rho).$$

3.7 Expected Units in the Service Facility (Ls)

A handy relation concerns the expected number of arrivals A and departures D, that occur in a specified time interval T, is shown in the two probability expressions below.

$$E(\text{A in T}) = \lambda T[P_0 + P_1 + \ldots] = \lambda T$$
$$E(\text{D in T}) = \mu T[P_1 + P_2 + \ldots] = \mu T(1 - P_o) = \mu T \text{ Ls}$$

Note the latter expression is related to both P_0 and Ls. Further, since the system is in an equilibrium state, E(A in T) = E(D in T), and thereby

$$\lambda T = \mu T(1-P_o) = \mu T \text{ Ls}$$
$$P_o = (1-\lambda/\mu) = (1-\rho)$$

and

$$Ls = \rho.$$

3.8 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained with use of (2.2) as below,

$$Lq = \sum_{n \geq 1} (n-1)P_n = \sum_{n \geq 1} (n-1)\rho^n(1-\rho) = \rho^2/(1-\rho)$$

3.9 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$$L = Ls + Lq = \rho/(1-\rho)$$

3.10 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$$Ws = Ls/\lambda = 1/\mu$$
$$Wq = Lq/\lambda = \rho/[\mu(1-\rho)]$$
$$W = Ws + Wq = 1/[\mu(1-\rho)]$$

3.11 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. Note the probabilities for these events,

$$P(D') = P_o = 1 - \rho$$
$$P(D) = (1 - P_o) = \rho$$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$$

Since $W_{q|D'} = 0$,

$$Wq' = W_{q|D} = Wq/P(D) = Wq/(1 - P_0) = Wq/\rho$$

3.12 Service Level

The service level (SL) is the probability a new arrival does not wait for service. In this one server system, this is merely P_0, the probability the system is empty. Hence,

$SL = P_0$.

Example 3.1
Suppose a one-service facility system with infinite capacity, and with exponential arrival and service times. The average time between arrivals is 10 min, and the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One server
Infinite capacity
Exponential input and output
$\tau_a = 10$ min = average time between arrivals
$\tau_s = 8$ min = average service time

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$ = utilization ratio
$P_n = (.20).80^n$ $n \geq 0$
$P_o = 0.2000$
$P_1 = 0.1600$
$P_2 = 0.1280$
$P_3 = 0.1024$
…
$Ls = 1 - P_o = 0.80$
$Lq = 3.20$
$L = Lq + Ls = 4.00$
$Ws = Ls/\lambda = 8$ min $= 0.1333$ h
$Wq = Lq/\lambda = 32$ min $= 0.5333$ h
$W = Wq + Ws = 40$ min $= 0.6666$ h
$Wq' = 40$ min $= 0.6666$ h
$SL = 0.20$ = service level

Example 3.2
Consider a one-man grease and oil shop that is open 10 h a day, where customers arrive on average every 10 min, and the average service time is 8 min. All is exponential. The average fee is $40 per car, and the average labor cost is $40 per hour. The owner wants to know what changes occur if he buys new equipment, at $50,000, that will allow him to reduce the average service time to 4 min. With the new equipment, he projects customers will come more often with an average arrival time of 8 min. Note the following measures below:

Input:
One-server
Infinite capacity
Exponential input and output
10-hour day
Average fee = $40 per car
Average labor cost = $40 per hour

	Current	Proposed
Capital cost		$50,000
Average arrival time (τ_a)	10 min	8 min
Average service time (τ_s)	8 min	4 min
Computations:		
$\rho(\tau_\sigma/\tau_a)$	0.80	0.50
λ (per hour)	6.0	7.5
10 h fee ($\lambda \times 10 \times 40$)	$2,400	$3,000
10 h labor (10×40)	$400	$400
10 h free minutes ($P_0 \times 10 \times 60$)	120	300
SL = $1-\rho$	0.20	0.50
Lq	3.20	0.50
Wq = Lq/$\lambda \times 60$ (min)	32.00	4.00
Payback days (50,000/600)		83.33

Example 3.3
The table below gives comparative results for $\rho = 0.1$–0.9 when one service facility, exponential arrival times, exponential service times and infinite queue capacity. The measures listed are P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL. For simplicity, the average service time is $\tau_s = 1.00$, and thereby Ws = 1.00 for all situations.

k	ρ	P_0	Lq	Ls	L	Wq	Ws	W	Wq′	SL
1	0.1	0.90	0.01	0.10	0.11	0.11	1.00	1.11	1.11	0.90
1	0.2	0.80	0.05	0.20	0.25	0.25	1.00	1.25	1.25	0.80
1	0.3	0.70	0.13	0.30	0.43	0.43	1.00	1.43	1.43	0.70
1	0.4	0.60	0.27	0.40	0.67	0.67	1.00	1.67	1.67	0.60
1	0.5	0.50	0.50	0.50	1.00	1.00	1.00	2.00	2.00	0.50
1	0.6	0.40	0.90	0.60	1.50	1.50	1.00	2.50	2.50	0.40
1	0.7	0.30	1.63	0.70	2.33	2.33	1.00	3.33	3.33	0.30
1	0.8	0.20	3.20	0.80	4.00	4.00	1.00	5.00	5.00	0.20
1	0.9	0.10	8.10	0.90	9.00	9.00	1.00	10.00	10.00	0.10

The table above can be used for any one server, infinite capacity queuing system with exponential arrival and service times. For example, if the average service time is $\tau_s = 8$ min, and the utilization ratio was $\rho = 0.80$, as in Example 3.1, all the measures listed above are the same, with a minor adjustment to the wait time measures. For this situation, $Wq = 4.00 \times \tau_s = 32.00$ min, $Ws = 1.00 \times \tau_s = 8.00$ min, $W = 5.00 \times \tau_s = 40.00$ min, and $Wq' = 5.00 \times \tau_s = 40.00$ min.

Chapter 4
One Server, Finite Queue (M/M/1/N)

Abstract This chapter explores the one-server, finite capacity system with exponential inter-arrival and service times. Could be, a one-man barbershop with three seats for the waiting customers. The difference, equilibrium and reduced equations are listed, and the probability on the number of units in the system are developed, along with the performance measures. Examples are presented.

4.1 Introduction

Suppose a system with one server and a finite queue where the maximum number of units allowed in the system is N, and where the inter-arrival and the service times have exponential probability densities. Further, the average time between arriving customers is $1/\lambda$ average service time is $1/\mu$. A finite queue example is a three line telephone service with one operator receiving information from a caller, while two lines are open for customers waiting to talk to the operator. When all three lines are full, potential new calls are lost.

The following notation applies here:

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$N =$ maximum units allowed in the system
$n =$ number of units in the system $\quad n = (0,N)$

Below is a list of the difference equations, the corresponding equilibrium equations and then the reduced equations. These are needed to develop the probability and statistical measures for the system.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_4, © Springer Science+Business Media New York 2012

4.2 Difference Equations

$$
\begin{array}{ll}
n = 0 & P_0(t + h) = (1-\lambda h)P_0(t) + \mu h P_1(t) + o(h) \\
n = (1,N-1) & P_n(t + h) = (1-\lambda h-\mu h)P_n(t) + \lambda h P_{n-1}(t) + \mu h P_{n+1}(t) + o(h) \\
n = N & P_N(t + h) = (1-\mu h)P_N(t) + \lambda h P_{N-1}(t) + o(h)
\end{array}
$$

4.3 Equilibrium Equations

$$
\begin{array}{ll}
n = 0 & 0 = -\lambda P_0 + \mu P_1 \\
n = (1,N-1) & 0 = -(\lambda + \mu)P_n + \lambda P_{n-1} + \mu P_{n+1} \\
n = N & 0 = -\mu P_N + \lambda P_{N-1}
\end{array}
$$

4.4 Reduced Equations

$$0 = -\lambda P_{n-1} + \mu P_n \quad n = (1, N)$$

4.5 Probability on n Units in the System

Using the reduced equations and the notation $\rho = \lambda/\mu$ the probability of n units in the system becomes,

$$P_n = \lambda/\mu P_{n-1} = \rho P_{n-1} \quad n = (1, N)$$

It is observed that

$$
\begin{array}{ll}
P_0 = & \rho^0 P_0 \\
P_1 = \rho P_0 = & \rho^1 P_0 \\
P_2 = \rho P_1 = & \rho^2 P_0 \\
\ldots & \\
P_N = \rho P_{N\text{-}1} = & \rho^N P_0
\end{array}
$$

and so forth, whereby,

$$P_n = \rho^n P_0 \quad n = (0, N)$$

Because all the probabilities sum to unity,

$$\sum_{n=0}^{N} P_n = P_0 \sum_{n=0}^{N} \rho^n = 1$$

In this system, the utilization ratio, ρ is greater than zero and could even be higher than one. To find P_0, the above relation is used with identity (2.7) as shown below.

$$P_o \sum_{n=0}^{N} \rho^n = P_o\left[(1-\rho^{N+1})/(1-\rho)\right]$$

and thereby,

$$P_0 = (1-\rho)/(1-\rho^{N+1})$$

Finally, the probability of n units in the system becomes

$$P_n = \rho^n(1-\rho)/(1-\rho^{N+1}) \quad n = (0, N)$$

4.6 Probability the System is Idle

The probability the system is empty is merely the probability that n = 0, i.e.,

$$P_o = (1-\rho)/(1-\rho^{N+1})$$

4.7 Expected Units in the Service Facility (Ls)

A handy set of relations concerns the expected number of arrivals A and departures D that occur in a specified time interval T. These are shown below.

$$E(A \text{ in } T) = \lambda T[P_0 + P_1 + \ldots + P_{N-1}] = \lambda[1 - P_N]T$$
$$E(D \text{ in } T) = \mu T[P_1 + P_2 + \ldots + P_N] = \mu T(1 - P_o) = \mu T \text{ Ls}$$

Note the latter expression is related to both P_0 and Ls. Further, since the system is in an equilibrium state, E(A in T) = E(D in T), and thereby

$$\lambda[1 - P_N] = \mu(1 - P_o) = \mu Ls$$

4.8 Lambda and Rho Effective

Because the queue size is finite, when the system is full, any new arrival is blocked from entering the system and becomes lost forever. Thereby, the average number of arrivals per unit of time that enter the system is less or equal to λ, and is here called lambda effective and labeled as λ_e. It is convenient to now define lambda effective and rho effective as below:

$\lambda_e = \lambda[1-P_N]$ = "lambda effective"
$\rho_e = \lambda_e/\mu$ = "rho effective"

In this context,

λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ = utilization rate, and could be greater than one,
ρ_e = effective utilization rate, and is less than one.

Since $\lambda[1 - P_N] = \mu(1 - P_o) = \mu Ls$.

$P_o = 1 - \lambda_e/\mu = 1 - \rho_e$

and

$Ls = \rho_e$.

4.9 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained, with use of (2.8), as below,

$$\begin{aligned} Lq &= \sum_{n=1}^{N} (n-1)P_n \\ &= \sum_{n=1}^{N} (n-1)\rho^n(1-\rho)/\left(1-\rho^{N+1}\right) \\ &= \rho^2\left[1 - N\rho^{N-1} + (N-1)\rho^N\right]/\left[(1-\rho)\left(1-\rho^{N+1}\right)\right] \end{aligned}$$

4.10 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$L = Ls + Lq$

4.11 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$$Ws = Ls/\lambda_e$$
$$Wq = Lq/\lambda_e$$
$$W = L/\lambda_e = Ws + Wq$$

4.12 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. The probabilities for these events are:

$$P(D') = P_o$$
$$P(D) = (1 - P_o)$$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given a delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$$

Since $W_{q|D'} = 0$,

$$Wq' = W_{q|D} = Wq/P(D) = Wq/(1 - P_0)$$

4.13 Service Level (SL) and Loss Probability (Ploss)

The service level (SL) is the probability an arrival to the system is not delayed in the queue, and this is simply P_0. The loss probability (Ploss) is the probability a new arrival is lost because the system capacity is too small. This is merely P_N, the probability the system is full, where any new arrival is blocked from entering. Hence,

$$SL = P_0$$
$$Ploss = P_N$$

Example 4.1
Suppose a one service facility system with finite capacity where $N = 5$ is the maximum number of units allowed in the system, and where the arrival and service times are exponential. The average time between arrivals is 10 min, and the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One-server
$N = 5$ = system capacity
Exponential input and output
$\tau_a = 10$ min = average time between arrivals
$\tau_s = 8$ min = average service time
Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 1/\tau_s = 0.125$ per minute
$\mu = 60/\tau_s = 7.5$ per hour
$P_n = (0.271).80^n$ n = (0,5)
$P_o = 0.2710$
$P_1 = 0.2168$
$P_2 = 0.1734$
$P_3 = 0.1387$
$P_4 = 0.1110$
$P_5 = 0.0888$
$\lambda_e = \lambda[1-P_5] = 0.0911$ per minute
$\lambda_e = 5.4672$ per hour
$\lambda_e = 0.7290$
$Ls = P_1 + P_2 + P_3 + P_4 + P_5 = 0.729$
$Lq = 1P_2 + 2P_3 + 3P_4 + 4P_5 = 1.139$
$L = 1.868$
$Ws = Ls/\lambda_e = 8$ min $= 0.1333$ h
$Wq = Lq/\lambda_e = 12.50$ min $= 0.2083$ h
$W = 20.52$ min $= 0.3417$ h
$Wq' = 17.14$ min $= 0.2857$ h
$SL = P_0 = 0.2710$
$Ploss = P_5 = 0.0888$

Example 4.2
A one-man barbershop is open 8 h a day and five days a week where customers arrive on average every 15 min and the average service time is 12 min. All is exponential. The average fee is $12 per cut. The shop has room for only two customers to wait ($N = 3$). When the shop is full, the potential customers do not enter. Note the following measures below:

Input:
One server
N = 3 = system capacity
Exponential input and output
Shop is open 8-hours a day and 5 days a week
Average fee is $12 per customer

Average inter-arrival time (τ_a)	15 min
Average service time (τ_s)	12 min

Computations:

ρ (τ_s/τ_a)	0.80
λ (per hour) = $60/\tau_a$	4.00
λ_e (per hour) = $\lambda[1-P_3]$	3.31
λ_e (per minute) = λ_e (per hour)/60	0.055
Expected customers per week ($40\lambda_e$)	132
Expected customers lost per week [$40(\lambda-\lambda_e)$]	28
Expected weekly fees (132 × 12)	$1584
Expected weekly fees lost (28 × 12)	$336

Note: $P_0 = 0.339$
$P_1 = 0.271$
$P_2 = 0.217$
$P_3 = 0.173$

$SL = P_0$	0.34
$Lq = 1P_2 + 2P_3$	0.563
Wq (minutes) = Lq/λe	10.2

Example 4.3
The table below lists values of Ploss, for $\rho = 0.1$–0.9 and N = 1–10, when one service facility and all times are exponential. Blanks are the same as 0.00.

Ploss

ρ/N	1	2	3	4	5	6	7	8	9	10
0.1	0.09	0.01	0.00							
0.2	0.17	0.03	0.01	0.00						
0.3	0.23	0.06	0.02	0.01	0.00					
0.4	0.29	0.10	0.04	0.02	0.01	0.00				
0.5	0.33	0.14	0.07	0.03	0.02	0.01	0.00			
0.6	0.38	0.18	0.10	0.06	0.03	0.02	0.01	0.01	0.00	
0.7	0.41	0.22	0.14	0.09	0.06	0.04	0.03	0.02	0.01	0.01
0.8	0.44	0.26	0.17	0.12	0.09	0.07	0.05	0.04	0.03	0.02
0.9	0.47	0.30	0.21	0.16	0.13	0.10	0.08	0.07	0.06	0.05

Note when $\rho = 0.8$ and N = 3. Ploss = 0.17.

Chapter 5
One Server, No Queue (M/M/1/1)

Abstract This chapter applies to a one-server, no queue system with exponential inter-arrival and service times. Could be a rental store with one electric saw available to rent. The saw is the service facility and the rental time becomes the service time. When the saw is out, future customers do not wait and go elsewhere. The difference, equilibrium and reduced equations are listed. The probability of n units in the system, and the performance measures are developed. Examples are presented.

5.1 Introduction

Suppose a system with one server and no queue where the maximum number of units allowed in the system is one, and where the inter-arrival and the service times have exponential probability densities. Further, the average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This could be a one person taxi service where customers are accepted only when the taxi is empty.

The following notation applies here:

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$n =$ number of units in the system $n = (0,1)$

Below is a list of the difference equations, the corresponding equilibrium equations and then the reduced equations. These are needed to develop the probability and statistical measures for the system.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_5, © Springer Science+Business Media New York 2012

5.2 Difference Equations

$n = 0 \qquad P_0(t + h) = (1 - \lambda h)P_0(t) + \mu h P_1(t) + o(h)$
$n = 1 \qquad P_1(t + h) = (1 - \mu h)P_1(t) + \lambda h P_0(t) + o(h)$

5.3 Equilibrium Equations

$n = 0 \qquad 0 = -\lambda P_0 + \mu P_1$
$n = 1 \qquad 0 = -\mu P_1 + \lambda P_0$

5.4 Reduced Equation

$0 = -\lambda P_0 + \mu P_1 \quad n = 1$

5.5 Probability on n Units in the System

Using the reduced equation and the notation $\rho = \lambda/\mu$ the probability of one unit in the system becomes,

$P_1 = \lambda/\mu P_0 = \rho P_0$

It is observed that

$P_0 = \qquad \rho^0 P_0$
$P_1 = \rho P_0 = \qquad \rho^1 P_0$

whereby,

$P_n = \rho^n \quad P_0 \quad n = (0, 1)$

Because the probabilities sum to unity,

$P_0 + P_1 = 1$

In this system, the utilization ratio, ρ is greater than zero and could even be higher than one. To find P_0 and P_1,the above relation is used as shown below.

$P_0[1 + \rho] = 1$

and thereby,

$P_0 = 1/(1 + \rho)$

$P_1 = \rho/(1 + \rho)$

Finally, the probability of n units in the system becomes

$P_n = \rho^n/(1 + \rho)$ n = (0,1)

5.6 Probability the System is Empty

The probability the system is empty is merely the probability that n = 0, i.e.,

$P_o = 1/(1 + \rho)$

5.7 Expected Units in the Service Facility (Ls)

The expected number of units in the system is

$Ls = 0P_0 + 1P_1 = \rho/(1 + \rho)$

5.8 Lambda and Rho Effective

It is convenient to now define lambda effective and rho effective as below:

$\lambda_e = \lambda[1 - P_1] = \lambda/(1 + \rho) =$ "lambda effective"

$\rho e = \lambda e/\mu = \rho(1 + \rho) =$ "rho effective"

In this context,
λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ = utilization ratio, and could be greater than one,
ρ_e = effective utilization ratio, and will be less than one.

$P_o = 1 - \lambda_e/\mu = 1 - \rho_e$

5.9 Expected Units in the Queue (Lq)

Since this system has no queue,

$Lq = 0$

5.10 Expected Units in the System (L)

The expected number of units in the system is the same as the expected number of units in the service facility, i.e.,

$L = Ls$

5.11 Expected Time in Service (Ws), Queue (Wq) and System (W)

$Ws = 1/\mu$

$Wq = 0$

$W = Ws$

5.12 Service Level and Loss Probability

The service level (SL) is the probability an arrival to the system is not delayed in the queue, and this is simply P_0. The loss probability (Ploss) is the probability a new arrival is lost because the system capacity is too small. This is P_1, the probability the system is full, because when $n = 1$, any new arrival is blocked from entering. Hence,

$SL = P_0$
$Ploss = P_1$

Example 5.1

Suppose a one-service facility system with no queue, and where the arrival and service times are exponential. The average time between arrivals is 10 min, and

the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One-server
No queue
Exponential input and output
τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$P_o = 1/[1 + \rho] = 0.5556$
$P_1 = \rho/[1 + \rho] = 0.4444$
$\lambda_e = \lambda[1 - P_1] = 0.0556$ per minute
$\lambda_e = 3.3333$ per hour
$\rho_e = 0.4444$
Ls = P1 = 0.444
Lq = 0
L = 0.444
Ws = 8 min = 0.1333 h
Wq = 0
W = 8 min = 0.1333 h
$Wq' = 0$
$SL = P_0 = 0.5556$
$Ploss = P_1 = 0.4444$

Example 5.2

A rental agency is open 6 days a week. They have one trailer for rent at $100 per day. The average customer arrivals are one per two days and the average rental time is 2.5 days. Customers will not wait for the trailer if it is out. Below are some of the statistics for the agency.

Input:
One-server
No queue
Agency is open 6 days a week
Rental rate is $100 per day

Average inter-arrival time (days) = τ_a	2.00
Average service time (days) = τ_s	2.50

Computations:

$\rho = \tau_s/\tau_a$	1.25
$P_0 = 1/(1 + \rho)$	0.444
$P_1 = \rho/(1 + \rho)$	0.556
λ (per day) $= 1/\tau_a$	0.500
μ (per day) $= 1/\tau_s$	0.400
λ_e (per day) $= \lambda(1 - \rho)$	0.222
$\rho_\varepsilon = \lambda_e/\mu$	0.555
Expected customers per week ($\lambda_e \times 6$)	1.332
Expected customers lost per week ($\lambda - \lambda_e$) $\times$ 6	1.668
Expected fees per week ($) (1.332 $\times$ 100 $\times$ 2.5)	333
Expected fees lost per week ($) (1.668 $\times$ 100 $\times$ 2.5)	417

Example 5.3

The table below gives comparative results for $\rho = 0.1$–10.0 when one service facility, exponential arrival times, exponential service times and no queue capacity. The measures listed are:Ploss, ρ_e and SL. Note, the higher the utilization ratio, ρ, the greater the portion of lost customers, Ploss. Also note how rho effective, ρ_e, is always less than one. This is necessary to have equilibrium when only one service facility.

K	ρ	Ploss	ρ_e	SL
1	0.1	0.09	0.09	0.91
1	0.5	0.33	0.33	0.67
1	1.0	0.50	0.50	0.50
1	2.0	0.67	0.67	0.33
1	5.0	0.83	0.83	0.17
1	10.0	0.91	0.91	0.09

Chapter 6
Multi Servers, Infinite Queue (M/M/k)

Abstract This chapter pertains to a multi-server, infinite capacity system with exponential inter-arrival and service times. Could be cars on an expressway arriving to a toll center with three tollbooths. The probability on n units, and the performance measures of the system are developed. The difference, equilibrium and reduced equations are listed, and examples are presented.

6.1 Introduction

Consider a system with k servers and an infinite queue where the inter-arrival and the service times have exponential probability densities. The average time between arriving units (customers) is $1/\lambda$, and the average service time is $1/\mu$. This could be the customers arriving to six checkout counters in a Wall Mart store. The following notation applies here:

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility
$k =$ number of service facilities
$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$\rho/k < 1$ is needed to ensure the system is in equilibrium
$n =$ number of units in the system $\quad (n \geq 0)$

Below is a list of the difference equations. Following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_6, © Springer Science+Business Media New York 2012

6.2 Difference Equations

$$
\begin{array}{ll}
n = 0 & P_0(t+h) = (1-\lambda h)P_0(t) + \mu h P_1(t) + o(h) \\
n = (1,k-1) & P_n(t+h) = (1-\lambda h - n\mu h)P_n(t) + \lambda h P_{n-1}(t) + \\
 & \quad (n+1)\mu h P_{n+1}(t) + o(h) \\
n \geq k & P_n(t+h) = (1-\lambda h - k\mu h)P_n(t) + \lambda h P_{n-1}(t) + \\
 & \quad k\mu h P_{n+1}(t) + o(h)
\end{array}
$$

6.3 Equilibrium Equations

$$
\begin{array}{ll}
n = 0 & 0 = -\lambda P_0 + \mu P_1 \\
n = (1,k-1) & 0 = -(\lambda + n\mu)P_n + \lambda P_{n-1} + (n+1)\mu P_{n+1} \\
n \geq k & 0 = -(\lambda + k\mu)P_n + \lambda P_{n-1} + k\mu P_{n+1}
\end{array}
$$

6.4 Reduced Equations

$$
\begin{array}{ll}
0 = -\lambda P_{n-1} + n\mu P_n & n = (1,k) \\
0 = -\lambda P_{n-1} + k\mu P_n & n > k
\end{array}
$$

6.5 Probability on n Units in the System

Using the reduced equations and the notation $\rho = \lambda/\mu$, the probability of n units in the system are as below. For n = 0 to k, the reduced equations yield the following;

$$
\begin{array}{llll}
P_0 & & = \rho^0 P_0 & \\
P_1 & = \rho P_0 & = \rho^1 P_0 & \\
P_2 & = \rho/2 P_1 & = \rho^2/2! P_0 & \\
P_3 & = \rho/3 P_2 & = \rho^3/3! P_0 & \\
\ldots & & & \\
P_n & = \rho/n P_{n-1} & = \rho^n/n! P_0 & n = (0,k)
\end{array}
$$

When n is k + 1 and larger, the reduced equations yield the relations listed below.

$$
\begin{array}{llll}
P_{k+1} & = \rho/k P_k & = \rho^{k+1}/[k!k] P_0 & \\
P_{k+2} & = \rho/k P_{k+1} & = \rho^{k+2}/[k!k^2] P_0 & \\
\ldots & & & \\
P_n & = \rho/k P_{n-1} & = \rho^n/[k!k^{n-k}] P_0 & n > k
\end{array}
$$

Summarizing,

$$\begin{aligned} P_n &= \rho^n/n!P_0 && n = (0,k) \\ P_n &= \rho^n/[k!k^{n-k}]P_0 && n > k \end{aligned}$$

At n = k, both of the above equations are the same; and because probabilities across all values of n sum to unity, the relation below applies.

$$\begin{aligned} \Sigma_{n\geq 0}P_n &= P_0\{\sum_{n=0}^{k-1}\rho^n/n! + \sum_{n\geq k}\rho^n/[k!k^{n-k}]\} \\ &= P_0\{\sum_{n=0}^{k-1}\rho^n/n! + \rho^k/k!\sum_{n\geq k}\rho^{n-k}/k^{n-k}]\} \end{aligned}$$

For equilibrium, $\rho/k < 1$. Applying (2.2) on the above right-hand term yields,

$$\Sigma_{n\geq 0}P_n = P_0\{\sum_{n=0}^{k-1}\rho^n/n! + \rho^k/[(k-1)!(k-\rho)]\}$$

So now, the probability of n = 0 is:

$$P_0 = 1/\{\sum_{n=0}^{k-1}\rho^n/n! + \rho^k/[(k-1)!(k-\rho)]\}$$

Finally, the probability of n units in the system becomes

$$P_n = \begin{cases} \rho^n/n!P_0 & n = (0, k-1) \\ \rho^n/[k!k^{n-k}]P_0 & n \geq k \end{cases}$$

6.6 Expected Units in the Service Facility (Ls)

Below lists the relations for the expected number of arrivals A, and expected number of departures D in a specified time interval T.

$$E(A \text{ in } T) = \lambda T[P_0 + P_1 + \ldots] = \lambda T$$
$$E(D \text{ in } T) = \mu T[P_1 + 2P_2 + \ldots + kP_k + kP_{K+1} + \ldots] = \mu TLs$$

Note the latter expression is related to Ls. Further, since the system is in an equilibrium state, E(A in T) = E(D in T), and thereby

$$\lambda = \mu Ls$$

and

$$Ls = \rho.$$

6.7 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained, with use of (2.2), as below,

$$Lq = \sum_{n>k} (n-k)P_n = \sum_{n>k} (n-k)\rho^n/[k!k^{n-k}]P_0$$
$$= P_0\rho^k/k! \sum_{n>k} (n-k)\rho^{n-k}/k^{n-k}]$$

Now using (2.3), yields,

$$Lq = P_0\rho^{k+1}/[(k-1)!(k-\rho)^2]$$

6.8 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$$L = Ls + Lq$$

6.9 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$$Ws = Ls/\lambda = 1/\mu$$
$$Wq = Lq/\lambda$$
$$W = Ws + Wq$$

6.10 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. Note the probabilities for these events,

$$P(D') = P_{n<k}$$
$$P(D) = P_{n\geq k}$$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given a delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$
Since $W_{q|D'} = 0$,
$Wq' = W_{q|D} = Wq/P(D) = Wq/P_{n\geq k}$

6.11 Service Level

The service level (SL) is the probability a new arrival does not wait for service. This is the probability that n is less than k, $P_{n<k.}$ Hence,

$SL = P_{n<k}$

Example 6.1

Suppose a two-service facility system with infinite capacity, and with exponential arrival and service times. The average time between arrivals is 10 min, and the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
Two-servers
Infinite capacity
Exponential input and output
τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min
Computations:

$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$P_n = (.4286).80^n/n!$ n = (0,2)
$P_n = (.2143).80^n/2^{n-2}$ $n \geq 3$
$P_o = 0.429$
$P_1 = 0.343$
$P_2 = 0.137$
$P_3 = 0.055$
…

$Ls = \rho = 0.800$
$Lq = 0.152$
$L = Lq + Ls = 0.952$
$Ws = 8$ min $= 0.133$ h
$Wq = Lq/\lambda$ (per minute) $= 1.52$ min $= 0.025$ h
$W = L/\lambda$ (per minute) $= 9.52$ min $= 0.159$ h
$P_{n\geq 2} = 1 - (P_0 + P_1) = 0.228$
$Wq' = 6.67$ min $= 0.111$ h
$SL = P_o + P_1 = 0.772$

Example 6.2

Consider a one-man (operator) grease and oil shop that is open 10 h a day, where customers arrive on average every 10 min and the average service time is 8 min. All is exponential. The average fee is $40 per car and the average labor cost is $40 per hour. The owner wants to know what changes occur if he has two operators. With the two operators, he projects customers will come more often with an average arrival time of 8 min. Note the following measures below:

Input:
Infinite capacity
Exponential input and output
Shop is open 10 h a day
Average fee = $40 per car
Average labor cost = $40 per hour

Number operators	1	2
Average arrival time (minutes) $= \tau_a$	10	8
Average service time (minutes) $= \tau_s$	8	8

Computations:

$\rho = \tau_s/\tau_a$	0.80	1.00
λ (per minute) $= 1/\tau_a$	0.100	0.125
μ (per minute) $= 1/\tau_s$	0.125	0.125
λ (per hour) $= 60/\tau_a$	6.0	7.5
Expected 10-hour fees ($\lambda \times 10 \times 40$)	$2,400	$3,000
Expected 10-hour labor cost (10×40)	$400	$800
P_0	0.200	0.333
P_1	……	0.333
…		
Lq	3.200	0.333
Wq (minutes) (Lq/λ)	32.00	2.67
1. operator ($P_0 \times 10 \times 60$)	120	
2. operators ($2P_0 + 1P_1$) $\times$ (10×60)		600
SL:		
1. operator (P_0)	0.200	
2. operators ($P_0 + P_1$)		0.666

Example 6.3

The table below gives comparative results for service facilities of 2, 5 and 10, and each with three levels of utilization ratios, ρ, and all with exponential arrival times, exponential service times and infinite queue capacity. The measures listed are P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL. For simplicity, the average service time is $\tau_s = 1.00$, and thereby Ws = 1.00 for all situations.

k	ρ	P_0	Lq	Ls	L	Wq	Ws	W	Wq′	SL
2	0.5	0.60	0.03	0.50	0.53	0.07	1.00	1.07	0.67	0.90
2	1.0	0.33	0.33	1.00	1.33	0.33	1.00	1.33	1.00	0.67
2	1.8	0.05	7.67	1.80	9.47	4.26	1.00	5.26	5.00	0.15
5	1.0	0.37	0.00	1.00	1.00	0.00	1.00	1.00	0.25	1.00
5	2.0	0.13	0.04	2.00	2.04	0.02	1.00	1.02	0.33	0.94
5	3.0	0.05	0.35	3.00	3.35	0.12	1.00	1.12	0.50	0.76
10	5.0	0.01	0.04	5.00	5.04	0.01	1.00	1.01	0.20	0.96
10	7.0	0.00	0.52	7.00	7.52	0.07	1.00	1.07	0.33	0.78
10	9.0	0.00	6.02	9.00	15.02	0.67	1.00	1.67	1.00	0.33

The table above can be used for any queuing system with a corresponding number of servers and exponential arrival and service times, and with infinite queue. For example, if five servers where the utilization ratio is ρ = 2.0, and the average service time is $\tau_s = 8$ min, then Wq = 0.02 × τ_s = 0.16 min, Ws = 1.00 × τ_s = 8.00 min, W = 1.02 × τ_s = 8.16 min, and Wq′ = 0.33 × τ_s = 2.64 min.

Example 6.4

The table below gives the minimum number of service facilities (k) needed to achieve the SL in an infinite queue capacity system with selected values of the utilization ratio (ρ) ranging from 0.1 to 700.

	SL			
	0.85	0.90	0.95	0.99
ρ	k			
0.1	1	1	2	2
0.2	2	2	2	3
0.3	2	2	2	3
0.4	2	2	3	3
0.5	2	2	3	4
0.6	2	3	3	4
0.7	3	3	3	4
0.8	3	3	4	4
0.9	3	3	4	5
1	3	3	4	5
2	5	5	6	7
3	6	6	7	9

4	7	8	9	10
5	9	9	10	12
10	15	16	17	19
20	26	27	29	32
30	38	39	41	45
40	49	50	52	57
50	60	61	64	66
60	70	72	75	80
70	81	83	86	92
80	92	94	97	103
90	102	105	108	114
100	113	115	119	125
200	218	221	226	235
300	322	326	331	343
400	425	429	436	449
500	528	533	540	555
600	631	636	644	660
700	733	739	747	765

Example 6.5
Consider a sports jogging shoe manufacturer using a mold (called a 'last') to produce a certain style shoe. The forecast calls for 250 pair for a five day week and size 9 traditionally gets 20 percent of the orders. For size 9, ten percent of the orders are for a narrow (9 N) width, 50 percent for medium (9 M), and 40 percent for wide (9 W). On average, the last stays in the shoe for one day in the manufacturing process. The plant management wants to know how many lasts to have in the plant inventory by width in size 9 to achieve SLs between 85 to 99 percent. Note the table below. The forecast for size 9 is $0.20 \times 250 = 50$ per week.

							SL			
							0.85	0.90	0.95	0.99
Size	%	f5	f1	λ	μ	ρ	k			
9 N	10	5	1	1	1	1.0	3	3	4	5
9 M	50	25	5	5	1	5.0	9	9	10	12
9 W	40	20	4	4	1	4.0	7	8	9	10
Sum							19	20	23	27

Note the average service time is $\tau_s = 1$ day, and therefore, $\mu = 1$ per day for all sizes. The five day forecast for size 9 N is $f5 = .10 \times 50 = 5$, and for one day it is $f1 = f5/5 = 1.0$. Hence, $\lambda = 1$ per day, and because $\mu = 1$, $\rho = 1.0$. Using the results from Example 6.4, the table lists the minimum number of lasts needed (k) by SL as 3, 3, 4 and 5. In a corresponding way, the minimum number of lasts needed for sizes 9 M and 9 W by SL are shown. The sum of the three lasts needed range from 19 to 27.

Chapter 7
Multi Servers, Finite Queue (M/M/k/N)

Abstract This chapter explores a multi-server, finite capacity system with exponential inter-arrival and service times. An example is a beauty shop with two hair stylists and with room for only five customers in the shop. The difference, equilibrium and reduced equations are listed. The probability on n units, and the performance measures are developed. Examples are presented to guide the reader.

7.1 Introduction

Consider a system with k servers and a finite queue where N is the maximum number of units allowed in the system. The inter-arrival and the service times have exponential probability densities, where the average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. Could be cars at a gas station with four pumps and room for only six cars in the station. The following notation applies here:

$\tau_a = 1/\lambda$ = average time between arrivals
$\tau_s = 1/\mu$ = average time to service a unit
λ = average number of arrivals per unit of time
μ = average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
k = number of service facilities
N = maximum units allowed in the system
n = number of units in the system n = (0,N)

Below is a list of the difference equations. Following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_7, © Springer Science+Business Media New York 2012

7.2 Difference Equations

$$
\begin{array}{ll}
n = 0 & P_0(t+h) = (1-\lambda h)P_0(t) + \mu h P_1(t) + o(h) \\
n = (1,k-1) & P_n(t+h) = (1-\lambda h - n\mu h)P_n(t) + \lambda h P_{n-1}(t) \\
 & \quad + (n+1)\mu h P_{n+1}(t) + o(h) \\
n = (k,N-1) & P_n(t+h) = (1-\lambda h - k\mu h)P_n(t) + \lambda h P_{n-1}(t) \\
 & \quad + k\mu h P_{n+1}(t) + o(h) \\
n = N & P_N(t+h) = (1-k\mu h)P_N(t) + \lambda h P_{N-1}(t) + o(h)
\end{array}
$$

7.3 Equilibrium Equations

$$
\begin{array}{ll}
n = 0 & 0 = -\lambda P_0 + \mu P_1 \\
n = (1,k-1) & 0 = -(\lambda + n\mu)P_n + \lambda P_{n-1} + (n+1)\mu P_{n+1} \\
n = (k,N-1) & 0 = -(\lambda + k\mu)P_n + \lambda P_{n-1} + k\mu P_{n+1} \\
n = N & 0 = -(k\mu)P_N + \lambda P_{N-1}
\end{array}
$$

7.4 Reduced Equations

$$
\begin{array}{ll}
0 = -\lambda P_{n-1} + n\mu P_n & n = (1,k) \\
0 = -\lambda P_{n-1} + k\mu P_n & n = (k+1,N)
\end{array}
$$

7.5 Probability on n Units in the System

Using the reduced equations and the notation $\rho = \lambda/\mu$, the probability of n units in the system are as below. For n = 0 to k, the reduced equations yield the following;

$$
\begin{array}{llll}
P_0 & & = \rho^0 P_0 & \\
P_1 & = \rho P_0 & = \rho^1 P_0 & \\
P_2 & = \rho/2 P_1 & = \rho^2/2! P_0 & \\
P_3 & = \rho/3 P_2 & = \rho^3/3! P_0 & \\
\ldots & & & \\
P_n & = \rho/n P_{n-1} & = \rho^n/n! P_0 & n = (0,k)
\end{array}
$$

When n is k + 1 and larger, the reduced equations yield the relations listed below.

$$\begin{aligned} P_{k+1} &= \rho/kP_k &&= \rho^{k+1}/[k!k]P_0 \\ P_{k+2} &= \rho/kP_{k+1} &&= \rho^{k+2}/[k!k^2]P_0 \\ &\ldots \\ P_n &= \rho/kP_{n-1} &&= \rho^n/[k!k^{n-k}]P_0 \quad n= (k+1,N) \end{aligned}$$

Summarizing,

$$\begin{aligned} P_n &= \rho^n/n!P_0 && n = (0,k) \\ P_n &= \rho^n/[k!k^{n-k}]P_0 && n = (k+1,N) \end{aligned}$$

At n = k, both of the above equations are the same; and because the probabilities across all values of n sum to unity, the relation below applies.

$$\begin{aligned} \Sigma_{n\geq 0}P_n &= P_0\{\sum_{n=0}^{k-1}\rho^n/n! + \sum_{n=k}^{N}\rho^n/[k!k^{n-k}]\} \\ &= P_0\{\sum_{n=0}^{k-1}\rho^n/n! + \rho^k/k!\sum_{n=k}^{N}\rho^{n-k}/k^{n-k}]\} \end{aligned}$$

Applying (2.7) on the above right-hand term yields,

$$\Sigma_{n\geq 0}P_n = P_0\{\sum_{n=0}^{k-1}\rho^n/n! + \rho^k/k![(k^{N-k+1} - \rho^{N-k+1})/(k-\rho)k^{N-k}]\}$$

So now, the probability of n = 0 becomes:

$$P_0 = 1/\{\sum_{n=0}^{k-1}\rho^n/n! + \rho^k/k![(k^{N-k+1} - \rho^{N-k+1})/(k-\rho)k^{N-k}]\}$$

Finally, the probability of n units in the system is

$$P_n = \begin{cases} \rho^n/n!P_0 & n = (0,k) \\ \rho^n/[k!k^{n-k}]P_0 & n = (k+1,N) \end{cases}$$

7.6 Expected Units in the Service Facility (Ls)

Below lists the relations for the expected number of arrivals A, and expected number of departures D in a specified time interval T.

$$E(A \text{ in } T) = \lambda T[P_0 + P_1 + \ldots + P_{N-1}] = \lambda T\,[1 - P_N]$$
$$E(D \text{ in } T) = \mu T[P_1 + 2P_2 + \ldots + kP_k + kP_{K+1} + \ldots + kP_N] = \mu TLs$$

Note the latter expression includes Ls. Further, since the system is in an equilibrium state, E(A in T) = E(D in T), and thereby

$\lambda[1 - P_N] = \mu Ls$

7.7 Lambda and Rho Effective

It is convenient to now define lambda effective and rho effective as below:

$\lambda_e = \lambda[1 - P_N]$ = "lambda effective"
$\rho_e = \lambda_e/\mu$ = "rho effective"

In this context,

λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ = utilization rate,
ρ_e = effective utilization rate,
ρ/k = might be larger than one,
ρ_e/k = will be less than one.

Since $\lambda[1 - P_N] = \mu$ Ls.
$Ls = \rho_e$.

7.8 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained as below,

$$Lq = \sum_{n=k}^{N} (n-k)P_n = \sum_{n=k}^{N} (n-k)\rho^n/[k!k^{n-k}]P_0$$

7.9 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$L = Ls + Lq$

7.10 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$Ws = Ls/\lambda_e = 1/\mu$
$Wq = Lq/\lambda_e$
$W = L/\lambda_e = Ws + Wq$

7.11 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. The probabilities for these events are,

$P(D') = P_{n<k}$
$P(D) = P_{n\geq k}$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given a delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$
Since $W_{q|D'} = 0$,
$Wq' = W_{q|D} = Wq/P(D) = Wq/P_{n\geq k}$

7.12 Service Level and Loss Probability

The service level (SL) is the probability an arrival to the system is not delayed in the queue, and this is simply $P_{n<k}$. The loss probability (Ploss) is the probability a new arrival is lost because the system capacity is too small. This is merely P_N. It is the probability the system is full, where any new arrival is blocked from entering. Hence,

$SL = P_{n<k}$
$Ploss = P_N$

Example 7.1

Suppose a two-service facility system with finite capacity, where the maximum number of units allowed in the system is five, and where the arrival and service times are exponential. The average time between arrivals is 10 min, and the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
$k = 2$ servers
$N = 5$ is system capacity
$\tau_a = 10$ min
$\tau_s = 8$ min
Input and output are exponential

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$P_n = (.431).80^n/n!$ n = (0, 2)
$P_n = (.431).80^n/(2!2^{n-2})$ n = (3, 5)
$P_o = 0.431$
$P_1 = 0.345$
$P_2 = 0.138$
$P_3 = 0.055$
$P_4 = 0.022$
$P_5 = 0.009$
$\lambda_e = \lambda$(per minute)$[1 - P_5] = 0.0991$ min
$\lambda_e = \lambda$(per hour)$[1 - P_5] = 5.946$ per hour
$\rho_e = \lambda_e/\mu = 0.793$
$Ls = 1P_1 + 2[P_2 + P_3 + P_4 + P_5] = 0.793$
$Lq = 1P_3 + 2P_4 + 3P_5 = 0.126$
$L = Lq+Ls = 0.919$
$Ws = Ls/\lambda_e = 8$ min $= 0.133$ h
$Wq = Lq/\lambda_e = 1.27$ min $= 0.021$ h
$W = Wq+Ws = 9.27$ min $= 0.154$ h
$Wq' = 5.77$ min $= 0.096$ h
$SL = P_0 + P_1 = 0.78$
$Ploss = P_5 = 0.009$

Example 7.2

A one-man barbershop is open 8 h a day and five days a week where customers arrive on average every 15 min and the average service time is 12 min. All is exponential. The average fee is $12 per cut. The shop has room for only two customers to wait (N = 3), if the shop is full, the potential customers do not enter. The owner wants to know how much his weekly fees will grow if he has two barbers. Note the following measures below:

Number of barbers (k)	1	2
Input:		
Number operators is k		
System capacity is N = 3		
Shop is open 40 h per week		
Input and output are exponential		
Average fee = $12		
τ_a = average arrival time (minutes)	15	15
τ_s = average service time (minutes)	12	12
Computations:		
ρ	0.80	0.80
P_0	0.339	0.445
P_1	0.271	0.356
P_2	0.217	0.142
P_3	0.173	0.056
λ (per hour) = $60/\tau_a$	4.00	4.00
λ_e (per hour) = λ [1 − P_3]	3.31	3.78
λ_e (per minute) = λ [1 − P_3]/60	0.055	0.063
Expected customers per week ($40\lambda_e$)	132	151
Expected customers lost per week [$40(\lambda - \lambda_e)$]	28	9
Expected weekly fees 12 × ($40\lambda_e$)	$1584	$1812
Expected weekly fees lost 12 × [$40(\lambda - \lambda_e)$]	$336	$108
SL (P_0)	0.34	
SL ($P_0 + P_1$)		0.80
Lq ($1P_2 + 2P_3$)	0.563	
Lq ($1P_3$)		0.056
Wq (minutes)	10.2	0.9

Example 7.3

The table below lists the loss probability, Ploss, for selected parameter values of: k, the number of service facilities, ρ, the utilization ratio, and N, the system capacity. The parameter ranges are k = (1, 2, 3), ρ = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0), and N = (1 to 10). Note, for example, when k = 2, ρ = 1.5 and N = 4,

Ploss = 0.12, indicating that twelve percent of the potential arrivals to the system are lost due to limited space in the capacity. Also if k = 3, Ploss = 0.06.

		N								
		1	2	3	4	5	6	7	8	9	10
k	ρ	Ploss									
1	0.5	0.33	0.14	0.07	0.03	0.02	0.01	0.00			
1	1.0	0.50	0.33	0.25	0.20	0.17	0.14	0.13	0.11	0.10	0.09
1	1.5	0.60	0.47	0.42	0.38	0.37	0.35	0.35	0.34	0.34	0.34
1	2.0	0.67	0.57	0.53	0.52	0.51	0.50	0.50	0.50	0.50	0.50
1	2.5	0.71	0.64	0.62	0.61	0.60	0.60	0.60	0.60	0.60	0.60
1	3.0	0.75	0.69	0.68	0.67	0.67	0.67	0.67	0.67	0.67	0.67
2	0.5		0.08	0.02	0.00						
2	1.0		0.20	0.09	0.04	0.02	0.01	0.01	0.00		
2	1.5		0.31	0.19	0.12	0.09	0.06	0.04	0.03	0.02	0.02
2	2.0		0.40	0.29	0.22	0.18	0.15	0.13	0.12	0.11	0.10
2	2.5		0.47	0.37	0.32	0.28	0.26	0.25	0.24	0.23	0.22
2	3.0		0.53	0.44	0.40	0.37	0.36	0.35	0.34	0.34	0.34
3	0.5			0.01	0.00						
3	1.0			0.06	0.02	0.01	0.00				
3	1.5			0.13	0.06	0.03	0.02	0.01	0.00		
3	2.0			0.21	0.12	0.08	0.05	0.03	0.02	0.01	0.01
3	2.5			0.28	0.19	0.14	0.10	0.08	0.06	0.05	0.04
3	3.0			0.35	0.26	0.20	0.17	0.15	0.13	0.11	0.10

Example 7.4

A fast food restaurant has room for only six cars (N = 6) in its drive-through lot. During the busy hours, cars arrive on average every 1.5 min. The average time to service a customer is 3.0 min. All times are exponential. Note, $\lambda = 60/1.5 = 40$ per hour, $\mu = 60/3.0 = 20$ per hour, and $\rho = 40/20 = 2.0$. When the lot is full, new arrivals do not enter. The average profit per car is \$10, and the service facility cost is \$50 per hour. If all of the potential customers are serviced, the average profit per hour would be $\lambda \times 10 = \$400$.

Using the results from Example 7.3, when one service facility is open, k = 1, Ploss = 0.50, and the average profit per hour is $(1 - .50) \times 400 = \$200$. If two servers are open, k = 2, Ploss = 0.15 and the average profit per hour is $(1 - .15) \times 400 = \$340$. When three servers are open, k = 3, Ploss = 0.05 and the average profit per hour is $(1 - .05) \times 400 = \$380$. Note, the average service cost per hour when one server (k = 1) is \$50, when two servers (k = 2), it is \$100, and when three servers (k = 3), \$150. These results should help management decide how many servers to have on duty during the busy hours.

Chapter 8
Multi Servers, No Queue (M/M/k/k)

Abstract This chapter pertains to a multi-server, no queue system with exponential inter-arrival and service times. An example is a phone system in a real estate company with five lines available to receive calls. When all lines are busy, new calls are lost. The corresponding difference, equilibrium and reduced equations are listed. The probability on n units, and the performance measures are developed. Examples are presented to guide the reader.

8.1 Introduction

Consider a system with k servers and no queue. The inter-arrival and the service times have exponential probability densities, where the average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This could be a rental agency with only two small cargo trailers available for let. When both trailers are let out, potential customers go elsewhere. The following notation applies here:

$\tau_a = 1/\lambda$ = average time between arrivals
$\tau_s = 1/\mu$ = average time to service a unit
λ = average number of arrivals per unit of time
μ = average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
k = number of service facilities
n = number of units in the system $\quad n = (0,k)$

Below is a list of the difference equations: following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_8, © Springer Science+Business Media New York 2012

8.2 Difference Equations

$n = 0$	$P_0(t + h) = (1 - \lambda h)P_0(t) + \mu h P_1(t) + o(h)$
$n = (1,k - 1)$	$P_n(t + h) = (1 - \lambda h - n\mu h)P_n(t) + \lambda h P_{n-1}(t) + (n + 1)\mu h P_{n+1}(t) + o(h)$
$n = k$	$P_k(t + h) = (1 - k\mu h)P_k(t) + \lambda h P_{k-1}(t) + o(h)$

8.3 Equilibrium Equations

$n = 0$	$0 = -\lambda P_0 + \mu P_1$
$n = (1,k - 1)$	$0 = -(\lambda + n\mu)P_n + \lambda P_{n-1} + (n + 1)\mu P_{n+1}$
$n = k$	$0 = -(k\mu)P_k + \lambda P_{k-1}$

8.4 Reduced Equations

$0 = -\lambda P_{n-1} + n\mu P_n \qquad n = (1,k)$

8.5 Probability of n Units in the System

Using the reduced equations and the notation $\rho = \lambda/\mu$, the probability of n units in the system are as below. For $n = 0$ to k, the reduced equations yield the following;

P_0		$= \rho^0 P_0$	
P_1	$= \rho P_0$	$= \rho^1 P_0$	
P_2	$= \rho/2 P_1$	$= \rho^2/2! P_0$	
P_3	$= \rho/3 P_2$	$= \rho^3/3! P_0$	
...			
P_n	$= \rho/n P_{n-1}$	$= \rho^n/n! P_0$	$n = (0,k)$

Summarizing,

$P_n \qquad = \rho^n/n! P_0 \qquad n = (0,k)$

Because the probabilities across all values of n sum to unity, the relation below applies.

$$P_n = P_0 \sum_{n=0}^{k} \rho^n/n!$$

So now, the probability of n = 0 becomes:

$$P_0 = 1/\sum_{n=0}^{k} \rho^n/n!$$

Finally, the probability of n units in the system is

$$P_n = \rho^n/n![1/\sum_{n=0}^{k} \rho^n/n!] \qquad n = (0,k)$$

8.6 Expected Units in the Service Facility (Ls)

Below lists the relations for the expected number of arrivals A, and expected number of departures D in a specified time interval T.

$E(A \text{ in } T) = \lambda T[P_0 + P_1 + \ldots + P_{k-1}] = \lambda T\,[1 - P_k]$
$E(D \text{ in } T) = \mu T[P_1 + 2P_2 + \ldots + kP_k] = \mu TLs$

Note the latter expression includes Ls. Further, since the system is in an equilibrium state, E(A in T) = E(D in T), and thereby,

$\lambda[1 - P_k] = \mu Ls$

8.7 Lambda and Rho Effective

It is convenient to now define lambda effective and rho effective as below:

$\lambda_e = \lambda[1 - P_k]$ = "lambda effective"
$\rho_e = \lambda_e/\mu$ = "rho effective"

In this context,

λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ = utilization ratio,
ρ_e = effective utilization ratio,
ρ/k = might be larger than one,
ρ_e/k = will be less than one.

Since $\lambda[1 - P_k] = \mu Ls$.

$Ls = \rho_e$.

8.8 Expected Units in the Queue (Lq)

There is no queue in this system, and thereby,

$Lq = 0$

8.9 Expected Units in the System (L)

The expected number of units in the system is

$L = Ls$

8.10 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$Ws = Ls/\lambda_e = 1/\mu$
$Wq = 0$
$W = L/\lambda_e = Ws$

8.11 Loss Probability

The loss probability (Ploss) is the probability a new arrival is lost because the system is full. This is merely P_k. Hence,

$Ploss = P_k$

Example 8.1
Suppose a two-service facility system with no queue, and where the arrival and service times are exponential. The average time between arrivals is 10 min, and the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
k = 2-servers
No queue
Input and output are exponential
τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour

$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$P_n = (.4717).80^n/n!$ n = (0, 2)
$P_o = 0.472$
$P_1 = 0.377$
$P_2 = 0.151$
$\lambda_e = \lambda[1 - P_2] = 5.094$ per hour
$\rho_e = \lambda_e/\mu = 0.680$
$Ls = (1P_1 + 2P_2) = 0.680$
$Lq = 0$
$L = 0.680$
$Ws = Ls/\lambda_e = 0.133$ h
$Wq = 0$
$W = 0.133$ h
$Wq' = 0$
$Ploss = P_2 = 0.151$

Example 8.2
A rental agency is open 6 days a week. They have one trailer for rent at $100 per day. The average arrivals are one per two days and the average rental time is 2.5 days. Customers will not wait for the trailer if it is out. The owner can buy another trailer for $10,000 and wants to know the financial changes to the agency if he does so. Below are some of the statistics for the agency, including the number of weeks to payback the $10,000 investment.

Input:		
Number of trailers (k = N)	1	2
No queue		
Input and output are exponential		
Agency open 6 days a week		
Rent is $100 per day		
Investment ($)		10,000
τ_a = average inter-arrival time (days)	2.00	2.00
τ_s = average service time (days)	2.50	2.50
Computations:		
P_0	0.444	0.472
P_1	0.556	0.377
P_2		0.151
λ (per day) = $(1/\tau_a)$	0.500	0.500
λ_e (per day) = $\lambda[1 - P_N]$	0.222	0.424
Expected customers per week = $(\lambda_e \times 6)$	1.332	2.544
Expected customers lost per week = $(\lambda - \lambda_e)6$	1.668	0.456
Expected fees per week ($) = $2.5 \times 100(\lambda_e \times 6)$	333	636
Expected fees lost per week ($) $2.5 \times 100(\lambda - \lambda_e)$	417	114
Weeks for payback [10,000/(636 − 333]		33

Example 8.3
The table below gives comparative results when 2, 5 and 10 service facilities (k) are available and each with three levels of the utilization ratio, ρ. The arrival times and the service times are exponential and there is no queue capacity. That is, the maximum number of units allowed in the system is k, the same as the number of service facilities. The measures listed are: P_0, L, Ploss, ρ_e and ρ_e/k. Note, as the utilization ratio, ρ, increases, the measures L and Ploss also increase. Note also, the measure ρ_e/k is always less than one. This is necessary to have equilibrium in the system.

k	ρ	P_0	L	Ploss	ρ_e	ρ_e/k
2	0.5	0.615	0.462	0.077	0.46	0.23
2	1.0	0.400	0.800	0.200	0.80	0.40
2	1.8	0.226	1.140	0.367	1.14	0.57
5	1.0	0.368	0.997	0.003	0.99	0.20
5	2.0	0.138	1.927	0.037	1.93	0.39
5	3.0	0.054	2.670	0.110	2.67	0.53
10	5.0	0.007	4.908	0.018	4.91	0.49
10	7.0	0.001	6.449	0.079	6.45	0.65
10	9.0	0.000	7.488	0.168	7.49	0.75

Example 8.4
The table below gives the minimum number of service facilities (k) needed in a no queue capacity system to achieve the service level (SL) with selected values of the utilization ratio (ρ) ranging from 0.1 to 700.

	SL			
	0.85	0.90	0.95	0.99
ρ	k			
0.1	1	1	2	2
0.2	2	2	2	3
0.3	2	2	2	3
0.4	2	2	3	3
0.5	2	2	3	4
0.6	2	3	3	4
0.7	2	3	3	4
0.8	3	3	3	4
0.9	3	3	4	5
1	3	3	4	5

2	4	4	5	7
3	5	6	7	8
4	6	7	8	10
5	7	8	9	11
10	12	13	15	18
20	21	23	26	30
30	30	32	36	42
40	38	42	46	53
50	47	51	56	64
60	56	60	66	75
70	64	69	76	85
80	73	78	86	96
90	81	88	95	107
100	90	97	105	117
200	175	188	202	221
300	261	278	298	324
400	346	368	394	426
500	431	458	489	527
600	516	549	585	628
700	601	639	680	728

Example 8.5

Consider a rental agency, open seven days a week, and suppose for items A, B, C, D, E, the average rental time is one day. Assume, the calls per week for each item are 7, 14, 21, 35 and 70, respectively. The customer will not wait if the item is out of stock. The management wants to know how many of the items to have in the store to achieve a service level of: SL = 0.85, 0.90, 0.95 and 0.99. Note the table below gives the seven day forecast (f7) by item. Also listed are the one day forecasts (f1) and the associated queuing parameters by item, λ, μ and ρ. Using the results from Example 8.4, the table lists the minimum number of units (k), by item, to have in stock to achieve the service level.

						SL			
						0.85	0.90	0.95	0.99
Item	f7	f1	λ	μ	ρ	k			
A	7	1	1	1	1	3	3	4	5
B	14	2	2	1	2	4	4	5	7
C	21	3	3	1	3	5	6	7	8
D	35	5	5	1	5	7	8	9	11
E	70	10	10	1	10	12	13	15	18

Chapter 9
One Server, Arbitrary Service (M/G/1)

Abstract This chapter considers a one-server, infinite capacity system with exponential inter-arrival times, and arbitrary service times with the average and standard deviation known. Could be a lift truck in a warehouse that hauls stock from the receiving dock to the storage area where the hauling time is normally distributed (not exponential). The lift truck is the service facility. The performance measures are developed and examples are presented.

9.1 Introduction

This chapter is sometimes referred as the Pollaczek–Khintchin formula, named after the authors of this important development. Their method considers a system with one server and an infinite queue where the inter-arrival time is exponential, and the service time t_s has an arbitrary distribution that could be either continuous or discrete. The average time between arriving customers is $1/\lambda$. The average service time is $1/\mu$ and the associated variance is σ^2. An example is the calls for service to a squad car in a one car patrol beat. The arrivals are Poisson distributed and the service times are not exponential. The following notation applies here:

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\sigma^2 =$ variance of the time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$\rho < 1$ is needed to assure the system is in equilibrium
$n =$ number of units in the system
$P_n =$ probability of n units in the system $\quad (n \geq 0)$

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_9, © Springer Science+Business Media New York 2012

9.2 Expected Units in the Service Facility (Ls) and Probability the System is Empty (P_0)

Consider the expected number of arrivals A, and departures D, that occur in a specified time interval T. The two probability expressions concerning A and D in T are listed below.

$E(A \text{ in } T) = \lambda T[P_0 + P_1 + \ldots] = \lambda T$
$E(D \text{ in } T) = \mu T[P_1 + P_2 + \ldots] = \mu T(1 - P_o) = \mu T \, Ls$

Note the latter expression is related to both P_0 and Ls. Further, since the system is in an equilibrium state, E(A in T) = E(D in T), and thereby

$\lambda = \mu(1 - P_o)$
$P_o = (1 - \lambda/\mu) = (1 - \rho)$

and

$Ls = \rho$.

9.3 Three Events

Consider the three events in the system that concern the time of departure from two units.

(1) To begin, suppose n units are in the queue, just prior to the departure of a unit in service.
(2) Just after the departure, there are (n − 1) units in the queue and one unit starting service.
(3) When the unit in service departs, there are now $n' = [(n - 1) + r + d]$ units in the system. The r units are those that entered the system while the just departed unit was being serviced. The variable d is defined below.

$d = 0$ if $n \geq 1$
$= 1$ if $n = 0$

9.4 Expected Value of n′, E(n′)

Below lists the expected value of n′.

$$E(n') = E(r + n - 1 + d)$$
$$= E(r) + E(n) - 1 + E(d)$$

Since the system is in equilibrium,

$$E(n') = E(n)$$

and thereby,

$$0 = E(r) - 1 + E(d)$$

and

$$E(r) = 1 - E(d)$$

Note,

$$\begin{aligned} E(d) &= 0P_{n \geq 1} + 1P_{n=0} \\ &= 0 + P_0 \\ &= 1 - \rho \end{aligned}$$

Since $d = d^2,\ E(d) = E\left(d^2\right)$

and thereby,

$$E(d^2) = 1 - \rho$$

Further,

$$E(r) = 1 - (1 - \rho) = \rho$$

9.5 Expected Value of n'^2, $E(n'^2)$

$$\begin{aligned} E(n'^2) &= E[(r + n - 1 + d)^2] \\ &= E\left[r^2\right] + E\left[n^2\right] + 1 + E\left[d^2\right] + 2E[rn] - 2E[r] + 2E[rd] \\ &\quad - 2E[n] + 2E[nd] - 2E[d] \end{aligned}$$

Now since the system is in equilibrium,

$$E(n'^2) = E\left[n^2\right]$$

and now,

$$0 = E\left[r^2\right] + 1 + E\left[d^2\right] + 2E[rn] - 2E[r\} + 2E[rd] - 2E[n] + 2E[nd] - 2E[d]$$

Because r and n are independent,

$$E[rn] = E[r]E[n] = \rho E[n]$$

Also, because r and d are independent,

$$E[rd] = E[r]E[d] = \rho(1-\rho)$$

Note, the relation between n and d gives

$$E[nd] = 0$$

Recall that the r arrivals occur during the time for service, t_s, and since r is Poisson, this leads to the relation below.

$$\begin{aligned} E\left[r^2\right] &= \int E\left[r^2|t_s\right] f(t_s) dt_s \\ &= \int [\lambda t_s + \lambda^2 t_s^2] f(t_s) dt_s \\ &= \rho + \lambda^2 E\left[t_s^2\right] \\ &= \rho + \lambda^2 \sigma^2 + \rho^2 \end{aligned}$$

Finally, the prior relation becomes.

$$0 = \rho + \lambda^2\sigma^2 + \rho^2 + 1 + 1 - \rho + 2\rho E[n] - 2\rho + 2\rho(1-\rho) - 2E[n] + 0 - 2(1-\rho)$$

Combining terms,

$$0 = 2\rho - \rho^2 + \lambda^2\sigma^2 + 2(\rho - 1)E[n\}$$

9.6 Expected Number of Units in the System (L)

Applying more algebra, and noting $E[n] = L$, the expected number of units in the system, yields,

$$L = [\lambda^2\sigma^2 + 2\rho - \rho^2]/[2(1-\rho)]$$

9.7 Expected Number of Units in the Queue (Lq)

The associated expression, Lq, for the expected number of units in the queue becomes,

$$Lq = [\lambda^2\sigma^2 + \rho^2]/[2(1-\rho)]$$

9.8 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$Ws = Ls/\lambda = 1/\mu$
$Wq = Lq/\lambda$
$W = Ws + Wq$

9.9 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. Note the probabilities for these events,

$P(D') = P_o = 1 - \rho$
$P(D) = (1 - P_o) = \rho$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$
Since $W_{q|D'} = 0$,
$Wq' = W_{q|D} = Wq/P(D) = Wq/(1 - P_0) = Wq/\rho$

9.10 Service Level

The service level (SL) is the probability a new arrival does not wait for service. This is merely P_0, the probability the system is empty. Hence,

$SL = P_0$.

9.11 Summary of the Statistical Measures

Below is a summary of the statistical measures for this system.

$P_o = 1 - \lambda/\mu = (1 - \rho)$
$Ls = \rho$
$Lq = [\lambda^2\sigma^2 + \rho^2]/[2(1 - \rho)]$
$L = [\lambda^2\sigma^2 + 2\rho - \rho^2]/[2(1 - \rho)]$
$Ws = Ls/\lambda = 1/\mu$
$Wq = Lq/\lambda$
$W = Ws + Wq$
$Wq' = Wq/\rho$
$SL = P_0 = 1 - \rho$

Example 9.1
Suppose a one service facility system with infinite capacity, and with exponential arrivals. The average time between arrivals is 10 min. The average time per service is 8 min, and the standard deviation of service is $\sigma = 1$ min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One-server
Infinite queue
Arrival times are exponential
Service times are not exponential
τ_a = average arrival times = 10 min
τ_s = average service times = 8 min
σ^2 = variance of service times = 1 minute^2

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$P_o = 0.2000$
$Ls = 0.80$
$Lq = 1.625$
$L = 2.425$
$Ws = 8$ min
$Wq = 16.25$ min
$W = 24.25$ min
$Wq' = 20.31$ min

Example 9.2

The table below gives comparative results for a queuing system with one service facility, and with three levels of utilization ratios, ($\rho = 0.1, 0.5, 0.9$), where the arrival times are exponential and the service times are arbitrary with coefficient of variation, (cov = 0.0, 0.4, 1.0, 2.0), and the queue capacity is infinite. The measures listed are P_0, Lq, Wq, Ws, Wq′ and SL. For simplicity, the average service time is $\tau_s = 1.00$, and thereby Ws = 1.00 for all situations. At cov = 0, the service time is a constant; at cov = 0.4, the service time is like a normal distribution; at cov = 1.0, the service time is exponential; and at cov = 2.0, the service time is highly tilted towards zero. Note how Lq, Wq and Wq′ increase directly with increases in cov, even when k and ρ remain unchanged.

k	ρ	cov	P_0	Lq	Wq	Ws	Wq′	SL
1	0.1	0.0	0.90	0.01	0.06	1.00	0.56	0.90
1	0.1	0.4	0.90	0.01	0.06	1.00	0.64	0.90
1	0.1	1.0	0.90	0.01	0.11	1.00	1.11	0.90
1	0.1	2.0	0.90	0.03	0.28	1.00	2.78	0.90
1	0.5	0.0	0.50	0.25	0.50	1.00	1.00	0.50
1	0.5	0.4	0.50	0.29	0.58	1.00	1.16	0.50
1	0.5	1.0	0.50	0.50	1.00	1.00	2.00	0.50
1	0.5	2.0	0.50	1.25	2.50	1.00	5.00	0.50
1	0.9	0.0	0.10	4.05	4.50	1.00	5.00	0.10
1	0.9	0.4	0.10	4.70	5.22	1.00	5.80	0.10
1	0.9	1.0	0.10	8.10	9.00	1.00	10.00	0.10
1	0.9	2.0	0.10	20.25	22.50	1.00	25.00	0.10

The table above can be used for any corresponding one server, infinite capacity queuing system with exponential arrival and arbitrary times. For example, if the average service time is $\tau_s = 8$ min, and the utilization ratio was $\rho = 0.50$, and the coefficient of variation is cov = 0.4, all the measures listed above are the same, with a minor adjustment to the wait time measures. For this situation, $Wq = 0.58 \times \tau_s = 4.64$ min, $Ws = 1.00 \times \tau_s = 8.00$ min, $W = 1.58 \times \tau_s = 12.64$ min, and $Wq' = 1.16 \times \tau_s = 9.28$ min.

Chapter 10
2 Populations, One Server, Arbitrary Service (M/G/1/2)

Abstract This chapter pertains when arrivals from two populations come to a system with one server and infinite capacity. The inter-arrival times are exponential and the services times from each population are arbitrary. An example are the calls for service to a squad car in a one-car patrol beat, where some calls are for minor scrapes and others are major incidents, and the combined service times are not exponential. The performance measures of the system are developed, and examples are presented.

10.1 Introduction

This chapter concerns two populations of customers, where the inter-arrivals times follow the exponential probability density and the service times are arbitrary. The chapter shows how to measure the statistics for the total system and for each of the individual populations. It is noted where the results from the chapter could easily be extended to include three or more arrival populations. This chapter is also an extension to Chap. 9 where the Pollaczek–Khintchin formula was developed. The average time between arriving customers is $1/\lambda_1$ and $1/\lambda_2$ for populations 1 and 2, respectively. The average service times are $1/\mu_1$ and $1/\mu_2$, and the associated variances are σ_1^2 and σ_2^2 for populations 1 and 2, respectively. This could be a situation where cars enter a one-car service garage, and some cars are for maintenance only (e.g., grease and oil), and others are for repair (e.g., transmission fault). The following notation applies here:

τ_{a1}, τ_{a2} = average inter-arrival time for units from populations 1 and 2
λ_1, λ_2 = average number of arrivals per unit of time from populations 1 and 2
τ_{s1}, τ_{s2} = average service time for units from populations 1 and 2
σ_1^2, σ_1^2 = variance of service times from populations 1 and 2

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_10, © Springer Science+Business Media New York 2012

μ_1, μ_2 = average number of units of populations 1 and 2 that are processed in a unit of time for a continuously busy service facility
$\lambda = \lambda_1 + \lambda_2$
$1/\mu$ = average service time for an arbitrary unit
$\rho = \lambda/\mu$ = utilization ratio
$\rho < 1$ is needed to assure the system is in equilibrium
n_1, n_2 = number of units in the system from populations 1 and 2
$n = n_1 + n_2$ = total number of units in the system $(n \geq 0)$
P_n = probability of n units in the system

10.2 Expected Time for an Arbitrary Arrival ($1/\lambda$)

Recall from Chap. 2, the convolution of two Poisson variables gives yet another Poisson variable, and thereby the expected arrival rate for an arbitrary unit to the system is

$$\lambda = \lambda_1 + \lambda_2$$

So now, the time between arrivals (t_a) to the system is exponential with an average time of $E(t_a) = 1/\lambda$

Note also where (λ_1/λ) is the probability a unit in the system is from population 1, and (λ_2/λ) is the corresponding probability the unit is from population 2.

10.3 Expected Time and Variance of Time in Service ($1/\mu$) and σ^2

For an arbitrary unit, the expected service time is the following,

$$\begin{aligned} E(t_s) &= 1/\mu \\ &= (1/\mu_1)(\lambda_1/\lambda) + (1/\mu_2)(\lambda_2/\lambda) \\ &= (\rho_1 + \rho_2)/\lambda \end{aligned}$$

and thereby the corresponding service rate is

$$\mu = 1/E(t_s)$$

To compute the associated variance, $E(t_s^2)$ is needed. This requires using the variances for populations 1 and 2 in the following way. For population 1, the variance of the service time (t_{s1}) is computed by the following expression,

$$V(t_{s1}) = E(t_{s1}^2) - E(t_{s1})^2$$

Hence,

$$E(t_{s1}^2) = V(t_{s1}) + E(t_{s1})^2$$

In the same way, the corresponding relation for a unit from population 2 becomes,

$$E(t_{s2}^2) = V(t_{s2}) + E(t_{s2})^2$$

Thereby, for an arbitrary unit in the system, the expression below applies,

$$E(t_s^2) = (\sigma_1^2 + 1/\mu_1^2)(\lambda_1/\lambda) + (\sigma_2^2 + 1/\mu_2^2)(\lambda_2/\lambda)$$

So now the variance for an arbitrary unit in the system is computed by,

$$V(t_s) = \sigma^2 = E(t_s^2) - E(t_s)^2$$

10.4 Statistics for an Arbitrary Unit in the System

Note, the parameters λ, μ and σ^2 are now known for the total system. Further, the arrivals are Poisson distributed and the service times are arbitrary and this allows using the results from Chap. 9 to measure the statistics for the total system. Hence, for an arbitrary unit, the following statistics are readily obtained:

$P_o = 1 - \lambda/\mu = (1 - \rho)$
$Ls = \rho$
$Lq = [\lambda^2\sigma^2 + \rho^2]/[2(1 - \rho)]$
$L = [\lambda^2\sigma^2 + 2\rho - \rho^2]/[2(1 - \rho)]$
$Ws = Ls/\lambda = 1/\mu$
$Wq = Lq/\lambda$
$W = Ws + Wq$
$Wq' = Wq/\rho$
$SL = P_0 = 1 - \rho$

10.5 Expected Number of Units in Service (Ls, Ls_1, Ls_2)

The expected number of units in the service facility for the system is denoted as Ls. Ls can be written as below.

$$\begin{aligned} Ls &= \lambda/\mu = \rho \\ &= (\lambda_1 + \lambda_2)(\rho_1 + \rho_2)/\lambda \\ &= \rho_1 + \rho_2 \end{aligned}$$

Thereby, the expected number of units in the service facility for populations 1 and 2 are listed below.

$Ls_1 = \lambda_1/\mu_1 = \rho_1$
$Ls_2 = \lambda_2/\mu_2 = \rho_2$

10.6 Expected Number of Units in Queue (Lq, Lq_1, Lq_2)

The expected number of units in the queue for the system is Lq. The corresponding number by population is determined as follows:

$$Lq_1 = (\lambda_1/\lambda)Lq$$
$$Lq_2 = (\lambda_2/\lambda)Lq$$

10.7 Expected Number of Units in the System (L, L_1, L_2)

The expected number of units in the system is L. The corresponding number by population is determined as follows:

$$L_1 = Ls_1 + Lq_1$$
$$L_2 = Ls_2 + Lq_2$$

10.8 Expected Time in Service (Ws, Ws_1, Ws_2)

The expected time in the service facility for the system is Ws. The corresponding time by population is determined as follows:

$$Ws_1 = 1/\mu_1$$
$$Ws_2 = 1/\mu_2$$

10.9 Expected Time in Queue (Wq, Wq_1, Wq_2)

The expected time in the queue for the system is Wq. The corresponding time by population is determined as follows:

$$Wq_1 = Wq$$
$$Wq_2 = Wq$$

10.10 Expected Time in the System (W, W_1, W_2)

The expected time in the system is W. The corresponding time by population is determined as follows:

$W_1 = Ws_1 + Wq_1$
$W_2 = Ws_2 + Wq_2$

10.11 Expected Time in Queue Given a Delay (Wq′, Wq'_1, Wq'_2)

The expected wait time in the queue given a delay for the system is Wq′. The corresponding time by population is determined as follows:

$Wq'_1 = Wq'$
$Wq'_2 = Wq'$

10.12 Service Level (SL, SL_1, SL_2)

The service level is the probability that an arrival to the system does not wait to enter the service facility. For the system, this is $SL = P_0$. In the same way, the service level for each population is also the same as SL, i.e., the probability the system is empty when the new arrival enters the system, yielding,

$SL_1 = SL_2 = SL = P_0$

Example 10.1
Consider a one-service facility system with infinite capacity, and with two input populations with exponential arrival times. The average time between arrivals is 30 min for population 1, and 10 min for population 2. The service times are not exponential and the average service times are 10 and 4 min for populations 1 and 2, respectively. The variances of the service times are 2.0 and 1.0 min^2 for populations 1 and 2, respectively. Some of the key probabilities and statistics associated with this system are listed below.

Input:
Two arrival populations (1,2)
One server
Infinite queue

Population inter-arrival times are exponential

τ_{a1} = average inter-arrival time for population 1 = 30 min
τ_{a2} = average inter-arrival time for population 2 = 10 min

Population service times are not exponential

τ_{s1} = average service time for population 1 = 10 min
τ_{s2} = average service time for population 2 = 4 min
σ_1^2 = variance of service time for population 1 = 2
σ_2^2 = variance of service time for population 2 = 1

Computations:
Population arrival and service rates

$\lambda_1 = 1/\tau_{a1} = 0.033/\text{min} = 2/\text{hour}$
$\mu_1 = 1/\tau_{s1} = 0.100/\text{min} = 6/\text{hour}$
$\lambda_2 = 1/\tau_{a2} = 0.100/\text{min} = 6/\text{hour}$
$\mu_2 = 1/\tau_{s2} = 0.250/\text{min} = 15/\text{hour}$

Arbitrary inter-arrival times are exponential

$\lambda = 0.133/\text{min} = 8/\text{hour}$
$\tau_a = 7.5 \text{ min} = 0.125 \text{ h}$

Arbitrary service times are not exponential

$\tau_s = 5.5 \text{ min} = 0.092 \text{ h}$
$\sigma^2 = 8.0 \text{ min}^2$
$\mu = 0.182/\text{min} = 10.91/\text{hour}$

Statistical measures:

Arbitrary unit	Population 1	Population 2
$\rho = \lambda/\mu = 0.733$		
$P_o = 0.267$		
Ls = 0.733	$Ls_1 = 0.333$	$Ls_2 = 0.400$
Lq = 1.270	$Lq_1 = 0.317$	$Lq_2 = 0.952$
L = 2.003	$L_1 = 0.650$	$L_2 = 1.352$
Ws = 5.5 min	$Ws_1 = 10.00$ min	$Ws_2 = 4.00$ min
Wq = 9.54 min	$Wq_1 = 9.54$ min	$Wq_2 = 9.54$ min
W = 15.04 min	$W_1 = 19.54$ min	$W_2 = 13.54$ min
$Wq' = 13.01$ min	$Wq_1' = 13.01$ min	$Wq_2' = 13.01$ min
SL = 0.267	$SL_1 = 0.267$	$SL_2 = 0.267$

Note, $Ls = Ls_1 + Ls_2$, $Lq = Lq_1 + Lq_2$ and $L = L_1 + L_2$. Also, $Ws = \tau_s$, $Ws_1 = \tau_{s1}$ and $Ws_2 = \tau_{s2}$

Chapter 11
M Machines, One Repairman (M/M/1/M)

Abstract This chapter explores a system with a limited number of units in the input population, like M machines in a shop that occasionally require service from one repairman. The inter-arrival time per unit and the service times are exponential. This could be a firm with five copy machines and one repairman. The probability on the number of units in the system is derived, and the performance measures of the system are developed. Examples are presented.

11.1 Introduction

Consider a system with M machines and one repairman and where the run time per machine and the service times have exponential probability densities. The average run time for a machine before it needs repair is $1/\lambda$ and the average service time is $1/\mu$. An example is a cargo ship with six diesel engines and one operator is available to handle all the maintenance and repair needs. The following notation applies here:

M = population size (machines)
$\tau_a = 1/\lambda$ = average run time per unit
$\tau_s = 1/\mu$ = average time to service a unit
λ = average arrival rate per machine in a unit of time
μ = average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu$
n = number of units in the system $n = (0, M)$

Below is a list of the difference equations. Following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_11, © Springer Science+Business Media New York 2012

11.2 Difference Equations

$n = 0$	$P_0(t + h) = (1 - M\lambda h)P_0(t) + \mu h P_1(t) + o(h)$
$n = (1,M - 1)$	$P_n(t + h) = (1 - (M–n)\lambda h - \mu h)P_n(t) + (M–n + 1)\lambda h P_{n-1}(t) + \mu h P_{n+1}(t) + o(h)$
$n = M$	$P_M(t + h) = (1 - \mu h)P_M(t) + \lambda h P_{M-1}(t) + o(h)$

11.3 Equilibrium Equations

$n = 0$	$0 = -M\lambda P_0 + \mu P_1$
$n = (1,M - 1)$	$0 = -([M–n]\lambda + \mu)P_n + (M–n + 1)\lambda P_{n-1} + \mu P_{n+1}$
$n = M$	$0 = -\mu P_M + \lambda P_{M-1}$

11.4 Reduced Equations

$0 = -(M–n)\lambda P_n + \mu P_{n+1}$ $\quad n = (0,M - 1)$

11.5 Probability on n Units in the System

Using the reduced equations and the notation $\rho = \lambda/\mu$, the probability of n = 1, 2 and 3 units in the system becomes:

$P_1 = M\rho P_0 = M\rho^1 P_0$
$P_2 = (M - 1)\rho P_1 = M(M - 1)\rho^2 P_0$
$P_3 = (M - 2)\rho P_2 = M(M - 1)(M - 2)\rho^3 P_0$

and so forth, whereby,

$P_n = M!/(M–n)!\rho^n P_0$ $n = (1,M)$

Because $P_0 = P_0$ and all the probabilities sum to unity,

$$\sum_{n=0}^{M} P_n = P_0 \sum_{n=0}^{M} \rho^n M!/(M - n)! = 1$$

thereby,

$$P_0 = 1/\sum_{n=0}^{M} \rho^n M!/(M - n)!$$

Finally, the probability of n units in the system becomes

$$P_n = \rho^n M!/(M - n)!/\sum_{n=0}^{M} \rho^n M!/(M - n)! \qquad n = (0, M)$$

11.6 Probability the System is Empty

The probability the system is empty is merely the probability that n = 0, i.e.,

$$P_o = 1/\sum_{n=0}^{M} \rho^n M!/(M-n)!$$

11.7 Expected Units in the Service Facility (Ls)

Note the expected number of units in the service facility is listed below.

$$Ls = \sum_{n=1}^{M} P_n = 1 - P_0$$

11.8 Expected Units in the System (L)

The reduced equations are needed to find the expected number of units in the system. Recall, the reduced equations give the following:

$$0 = -(M–n)\lambda P_n + \mu P_{n+1} \qquad n = (0, M-1)$$

Now summing, yields

$$\begin{aligned}\sum_{n=0}^{M-1}[-(M-n)\lambda P_n + \mu P_{n+1}] &= \sum_{n=0}^{M-1}[-M\lambda P_n + n\lambda P_n + \mu P_{n+1}] \\ &= -M\lambda(1-P_M) + \lambda(L - MP_M) + \mu(1-P_0) \\ &= -M\lambda + \lambda L + \mu(1-P_0)\end{aligned}$$

Solving for L, gives,

$$L = M - (1 - P_0)/\rho$$

11.9 Expected Units in the Queue (Lq)

So now, the expected number of units in the queue becomes,

$$Lq = L - Ls$$

11.10 Expected Time in Service (Ws)

The expected time in the service facility is the following,

$$Ws = 1/\mu$$

Running machines	Queue	Repairman
X	X	X
X		
X		

Fig. 11.1 Depiction of a shop with M = five machines and R = one repairman. X identifies each machine. The example shows an instant when three machines are running, one is in the queue and another is in repair. The run time per machine before it needs repair is $\tau_a = 20$ h

Because the arrival rate to the service facility is not constant, Little's Law does not apply, and so it is not possible to compute the expected time for a unit in the queue (Wq) and in the system (W) using Little's Law.

11.11 Service Level

The service level (SL) is the probability a new arrival does not wait for service. This is merely P_0, the probability the system is empty. Hence,

$SL = P_0$.

Example 11.1
Suppose a one operator (repairman) system with five machines, and with exponential arrival and service times. The average run time per machine before needing repair is 20 h and the average service time per machine is two hours. See Fig. 11.1 Some of the key probabilities and statistics associated with this system are listed below.

Input:
M = 5-machines
R = 1-repairman
Run time per machine and service times are exponential
τ_a = average run time per machine = 20 h
τ_s = average service time = 2 h

Computations:
$\lambda = 1/\tau_a = 0.05$ per hour
$\mu = 1/\tau_s = 0.50$ per hour
$\rho = \lambda/\mu = 0.10$
$P_n = 0.564[0.10^n\ 5!/(5 - n)!]$
$P_o = 0.564$
$P_1 = 0.282$
$P_2 = 0.113$
$P_3 = 0.034$
$P_4 = 0.007$

$P_5 = 0.001$
$Ls = P_1 + P_2 + P_3 + P_4 + P_5 = 0.436$
$L = 1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 = 0.640$
$Lq = 1P_2 + 2P_3 + 3P_4 + 4P_5 = 0.204$
$Ws = \tau_s = 2$ h
$SL = P_0 = 0.564$

Example 11.2
Consider the data from Example 11.1 and assume the shop is open 8 h a day and 5 days a week. Each running machine yields $1,000 an hour and the operator cost is $800 per day. The financial statistics are listed below.

Input:
Shop is open 8 h per day and 5 days per week
Yield is $1000 per hour for each running machine
Cost is $800 per day

Computations:
Machine-hours running per week = (5 – L) × 40 174.4
Machine-hours in repair-shop per week = (L) × 40 25.6
Yield per week ($) = (5 – L) × 40 × 1000 174,400
Lost yield per week ($) = (L) × 40 × 1000 25,600
Labor cost per week ($) = 5 × 800 4,000

Example 11.3
The table below gives comparative results for a queuing system with limited population sizes of M = 5, 10 and 15 machines, and when one repairman, R = 1. The utilization ratios per machine are $\rho = 0.01$, 0.05 and 0.10. Further, the arrival times and service times are exponential. The measures listed are P_0, Lq, Ls, L, and SL. Recall, $\rho = \tau_s/\tau_a$ and thereby, $\rho = 0.01$ indicates the average service time for a machine is one percent of the average run time. At $\rho = 0.05$, the average service time is five percent of the average run time for a machine. At $\rho = 0.10$, the average service time is ten percent of the average machine run time.

The values listed in the table below are obtained from computer calculations that use the equations listed in the chapter.

M	R	ρ	P_0	Lq	Ls	L	SL
5	1	0.01	0.96	0.00	0.04	0.04	0.96
5	1	0.05	0.81	0.03	0.09	0.22	0.81
5	1	0.10	0.65	0.11	0.35	0.47	0.65
10	1	0.01	0.90	0.01	0.10	0.11	0.90
10	1	0.05	0.54	0.30	0.46	0.76	0.54

10	1	0.10	0.21	1.36	0.79	2.15	0.21
15	1	0.01	0.85	0.02	0.15	0.17	0.85
15	1	0.05	0.33	0.93	0.67	1.60	0.33
15	1	0.10	0.04	4.40	0.96	5.36	0.04

Chapter 12
M Machines, R Repairmen (M/M/R/M)

Abstract This chapter pertains for a system with a limited number of units in the input population, like M machines in a shop that occasionally require service from R servers (repairmen). The inter-arrival time per unit and the service times are exponential. An example is a taxi fleet of 100 cabs with four service mechanics on duty to maintain and repair the cabs as needed. The probability on number of units in the repair system is generated. The basic performance measures of the system are developed and examples are presented.

12.1 Introduction

Consider a system with M machines and R repairmen and where the run time per machine and the service times have exponential probability densities. The average run time for a machine before it needs repair is $1/\lambda$ and the average service time is $1/\mu$. This could be a firm with five printing presses and two repairmen. The following notation applies here:

M = population size (number machines)
R = number of service facilities (repairmen)
$\tau_a = 1/\lambda$ = average run time per unit
$\tau_s = 1/\mu$ = average time to service a unit
λ = average arrival rate per machine in a unit of time
μ = average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu$
n = number of units in the system n = (0, M)

Below is a list of the difference equations. Following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_12, © Springer Science+Business Media New York 2012

12.2 Difference Equations

$n = 0$	$P_0(t + h) = (1 - M\lambda h)P_0(t) + \mu h P_1(t) + o(h)$
$n = (1,R - 1)$	$P_n(t + h) = (1 - (M–n)\lambda h - n\mu h)P_n(t) + (M–n + 1)$ $\lambda h P_{n\_1}(t) + (n + 1)\mu h P_{n+1}(t) + o(h)$
$n = (R,M - 1)$	$P_n(t + h) = (1 - (M–n)\lambda h - R\mu h)P_n(t) + (M–n + 1)$ $\lambda h P_{n\_1}(t) + R\mu h P_{n+1}(t) + o(h)$
$n = M$	$P_M(t + h) = (1 - R\mu h)P_M(t) + \lambda h P_{M-1}(t) + o(h)$

12.3 Equilibrium Equations

$n = 0$	$0 = -M\lambda P_0 + \mu P_1$
$n = (1,R - 1)$	$0 = [- (M - n)\lambda - n\mu]P_n + (M–n + 1)\lambda P_{n-1}$ $+ (n + 1)\mu P_{n+1}$
$n = (R,M - 1)$	$0 = [- (M - n)\lambda - R\mu]P_n + (M–n + 1)\lambda P_{n-1} + R\mu P_{n+1}$
$n = M$	$0 = [- R\mu]P_M + \lambda P_{M-1}$

12.4 Reduced Equations

$0 = -(M - n)\lambda P_n + (n + 1)\mu P_{n+1}$	$n = (0,R - 1)$
$0 = -(M - n)\lambda P_n + R\mu P_{n+1}$	$n = (R,M - 1)$

12.5 Probability on n Units in the System

When $n < R$, the reduced equations give the probabilities listed below for $n = 1, 2$ and 3:

$P_1 = M\rho P_0$	$= M\rho^1 P_0$
$P_2 = [(M - 1)\rho/2]P_1$	$= M(M - 1)\rho^2/2!P_0$
$P_3 = [(M - 2)\rho/3]P_2$	$= M(M - 1)(M - 2)\rho^3/3!P_0$

and so forth, whereby,

$P_n = \rho^n\ M!/[(M–n)!n!]\ P_0 \quad n = (0,R)$

When $n = (R,M - 1)$, the probabilities are:

$$\begin{array}{lll} n = R & P_{R+1} = [(M - R)/R]\rho P_R & = \rho^{R+1}\, M!/[(M - R)!R!R]P_0 \\ n = R + 1 & P_{R+2} = [(M - R - 1)/R]\rho P_{R+1} & = \rho^{R+2}\, M!/[(M - R - 1)!R!R^2]\, P_0 \end{array}$$

and in general,

$$P_n = \rho^n\, M!/[(M–n)!R!R^{n-R}]\, P_0 \quad n = (R + 1,M)$$

Because all the probabilities sum to unity,

$$\sum_{n=0}^{M} P_n = P_0\{\sum_{n=0}^{R} \rho^n\, M!/[(M–n)!n!] + \sum_{n=R+1}^{M} \rho^n\, M!/[(M–n)!R!R^{n-R}]\}$$

thereby,

$$P_0 = 1/\{\sum_{n=0}^{R} \rho^n\, M!/[(M–n)!n!] + \sum_{n=R+1}^{M} \rho^n\, M!/[(M–n)!R!R^{n-R}]\}$$

Finally, the probability of n units in the system becomes

$$\begin{array}{lll} P_n & = P_0\rho^n\, M!/[(M–n)!n!] & n = (0,R) \\ & = P_0\rho^n\, M!/[(M–n)!R!R^{n-R}] & n = (R + 1,M) \end{array}$$

12.6 Expected Units in the Service Facility (Ls)

The expected number of units in the service facility is computed as below.

$$Ls = \sum_{n=0}^{R-1} nP_n + R\sum_{n=R}^{M} P_n$$

12.7 Expected Units in the Queue (Lq)

The expected number of units in the queue is computed as below,

$$Lq = \sum_{n=R}^{M} (n - R)P_n$$

Running machines	Queue	Repairman
X	X	X
X		X

Fig. 12.1 Depiction of a shop with M = five machines and R = two-repairmen. X identifies each machine. The example shows an instant when two machines are running, one is in the queue and two are in repair. The run-time per machine before it needs repair is $\tau_a = 20$ h

12.8 Expected Units in the System (L)

So now, the expected number of units in the system becomes,

$L = Ls + Lq$

12.9 Expected Time in Service (Ws)

The expected time in the service facility is the following,

$Ws = 1/\mu$

Because the arrival rate to the service facility is not constant Little's Law does not apply, and so it is not possible to compute the expected time for a unit in the queue (Wq) and in the system (W).

12.10 Service Level

The service level (SL) is the probability a new arrival does not wait for service. This is merely $P_{n<R}$. Hence,

$$SL = \sum_{n=0}^{R-1} P_n$$

Example 12.1

Suppose a two-operator (repairmen) system with five machines, and with exponential arrival and service times. The average run time per machine before needing repair is 10 h and the average service time is 2 h (see Fig. 12.1). Some of the key probabilities and statistics associated with this system are listed below.

Input:

M = number of machines = 5

R = number of repairmen = 2

Run times and service times are exponential

τ_a = average run time per machine = 10 h

τ_s = average service time = 2 h

Computations:
$\lambda = 1/\tau_a = 0.10$ per hour
$\mu = 1/\tau_s = 0.50$ per hour
$\rho = \lambda/\mu = 0.20$
$P_n = 0.391[.2^n]5!/[(5-n)!n!] \quad n = (0,2)$
$= 0.391[.2^n]5!/[(5-n)!2!2^{n-2}] \quad n = (3,5)$
$P_o = 0.391$
$P_1 = 0.391$
$P_2 = 0.156$
$P_3 = 0.047$
$P_4 = 0.009$
$P_5 = 0.001$
$Ls = 1P_1 + 2P_2 + 2P_3 + 2P_4 + 2P_5 = 0.82$
$Lq = 1P_3 + 2P_4 + 3P_5 = 0.07$
$L = 1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 = 0.89$
$Ws = \tau_s = 2$ h
$SL = P_0 + P_1 = 0.782$

Example 12.2
Consider the same data from Examples 11.1 and 11.2 where the shop is open 8 h a day and 5 days a week. But now the shop has two operators. Recall, the average run time per machine is 20 h and the average service time is two hours. Also, each running machine yields \$1,000 an hour and each operator's cost is \$800 per day. The financial statistics are listed below.

R = Number of operators	1	2
M = Number of machines	5	5
Shop open 8 h per day and 5 days per week		
Yield = \$1000 per machine running hour		
Cost = \$800 per day per repairman		

Computations:

ρ	0.10	0.10
P_o	0.564	0.622
P_1	0.282	0.311
P_2	0.113	0.062
P_3	0.034	0.005
P_4	0.007	0.001
P_5	0.001	0.000
$L = (1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5)$	0.64	0.45
M-L	4.36	4.55

Machine-hours running per week = (M-L) × 40	174.4	182.0
Machine-hours in-repair shop per week = 40L	25.6	18.0
Yield per week ($) = (M-L) × 40 × 1000	174,400	182,000
Lost yield per week ($) = 40L × 1000	25,600	18,000
Labor cost per week ($) = R×800	4,000	8,000

Example 12.3
The table below gives comparative results for a queuing system with R = 2 and 3 repairmen, and limited population sizes of M machines. When 2 repairmen, M = 10, 20 and 30; when 3 repairmen, M = 20, 30 and 40. The utilization ratios per machine are $\rho = 0.01$ and 0.05, and the arrival times and service times are exponential. The measures listed are P_0, Lq, Ls, L, and SL. Recall, $\rho = \tau_s/\tau_a$ and thereby, $\rho = 0.01$ indicates the average service time for a machine is one percent of the average machine run time. At $\rho = 0.05$, the average service time is five percent of the average machine run time. The run time is the machine operating time before it needs repair.

The values in the table below are derived from computer calculations that use the equations listed in the chapter.

M	R	ρ	P_0	Lq	Ls	L	SL
10	2	0.01	0.91	0.00	0.10	0.10	1.00
10	2	0.05	0.61	0.02	0.48	0.50	0.91
20	2	0.01	0.82	0.00	0.20	0.20	0.98
20	2	0.05	0.35	0.21	0.94	1.15	0.71
30	2	0.01	0.74	0.01	0.30	0.30	0.96
30	2	0.05	0.18	0.94	1.38	2.32	0.44
20	3	0.01	0.82	0.00	0.20	0.20	1.00
20	3	0.05	0.37	0.03	0.95	0.98	0.93
30	3	0.01	0.74	0.00	0.30	0.30	1.00
30	3	0.05	0.22	0.15	1.42	1.57	0.80
40	3	0.01	0.67	0.00	0.40	0.40	0.99
40	3	0.05	0.13	0.51	1.88	2.39	0.62

Example 12.4
The table below lists the values of L (number of units in the queue plus in repair) when: $\rho = 0.01$, the number of repairmen is R = 1, 2, 3, 4, and the number of machines, M, range from 6 to 150. Note $\rho = \tau_s/\tau_a = 0.01$, where τ_a is the average

run time for a machine before it needs repair, and τ_s is the average repair time for a machine. So in this example, the repair time for a machine is small compared to the run time, i.e., one percent. L is the number of machines out of M that are non-productive and are in need of repair.

The values of L listed in the table below are derived from computer calculations that use the equations in the chapter.

Values of L at $\rho = 0.01$

M/R	1	2	3	4
6	0.06	0.06	0.06	0.06
8	0.08	0.08	0.08	0.08
10	0.11	0.10	0.10	0.10
15	0.17	0.15	0.15	0.15
20	0.24	0.20	0.20	0.20
30	0.41	0.30	0.30	0.30
40	0.63	0.41	0.40	0.40
50	0.93	0.52	0.50	0.50
60	1.35	0.65	0.60	0.59
70	1.97	0.78	0.70	0.69
80	2.95	0.93	0.81	0.79
90	4.61	1.10	0.92	0.89
100	7.57	1.29	1.03	1.00
150	50.00	3.04	1.70	1.53

Note when M = 100 and R = 1, the number of machines in queue plus in repair is L = 7.57. When R = 2 repairmen, L drops to 1.29

Chapter 13
One Server, Repeat Service (M/M/1/θ)

Abstract This chapter concerns a system with one-server, infinite capacity, exponential inter-arrival and service times, and where the service may need to be repeated. An example is a one-operator machine shop fabrication of a fixture that is tested at the end to see if it passes a strength test. If not, another fixture must be fabricated. The probability on n units in the system is developed. Examples are presented.

13.1 Introduction

Consider a system with one server and an infinite queue where the inter-arrival and the service times have exponential probability densities. The average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. The probability is θ of a fault in the service, whereby the service needs to be repeated. This could be a print shop where the quality of a job could be faulty and the job must be repeated. The following notation applies here:

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility
$\theta =$ probability the service must be repeated
$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$\rho/(1 - \theta) < 1$ is needed to assure the system is in equilibrium
$n =$ number of units in the system ($n \geq 0$)

Below is a list of the difference equations for the system. Following are the corresponding equilibrium equations and then the reduced equations.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_13,

13.2 Difference Equations

$n = 0 \qquad P_0(t + h) = (1 - \lambda h)P_0(t) + (1 - \theta)\mu h P_1(t) + o(h)$

$n \geq 1 \qquad P_n(t + h) = (1 - \lambda h - (1 - \theta)\mu h)P_n(t) + \lambda h P_{n-1}(t) + (1 - \theta)\mu h P_{n+1}(t) + \theta\mu h P_n(t) + o(h)$

13.3 Equilibrium Equations

$n = 0 \qquad 0 = -\lambda P_0 + (1 - \theta)\mu P_1$

$n \geq 1 \qquad 0 = -(\lambda + \mu)P_n + \lambda P_{n-1} + (1 - \theta)\mu P_{n+1}$

13.4 Reduced Equations θ

$0 = -\lambda P_{n-1} + (1 - \theta)\mu P_n \qquad n \geq 1$

13.5 Probability on n Units in the System

Using the reduced equations and the notation $\rho' = \lambda/[(1 - \theta)\mu]$, the probability of n units in the system becomes.

$P_n = \lambda/[(1 - \theta)\mu]P_{n-1} = \rho' P_{n-1} \; n \geq 1$

It is observed that

$P_0 = \qquad \rho'^0 P_0$

$P_1 = \rho' P_0 = \qquad \rho'^1 P_0$

$P_2 = \rho' P_1 = \qquad \rho'^2 P_0$

$P_n = \rho'^n P_0 \qquad n \geq 0$

and so forth, whereby,

Because all the probabilities sum to unity,

$\Sigma_{n\geq 0} P_n = P_0 \Sigma_{n\geq 0}\rho'^n = 1$

To maintain equilibrium, it is necessary for $\rho' < 1$, and because $\rho' = \rho/(1 - \theta)$, we need $\rho < (1 - \theta)$. This allows applying (2.1) to the above relation to yield,

$P_0 \Sigma_{n\geq 0}\rho'^n = P_0 1/(1 - \rho')$

thereby,
$P_0 = (1 - \rho')$.

Finally, the probability of n units in the system becomes

$P_n = \rho'^n(1 - \rho') \quad n \geq 0$

13.6 Expected Runs

The expected number of runs to get a good unit is obtained from the geometric distribution. Let x represent the number of runs until a good unit results when θ is the probability of a defective unit. The expected value of x becomes,

$E(x) = 1/(1 - \theta)$.

Example 13.1
Suppose a one-service facility system with infinite capacity, and with exponential arrival and service times. The average time between arrivals is 10 min, and the average time per service is 8 min. The probability the service must repeat is 0.10. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One server
Infinite capacity
Arrival and service times are exponential
τ_a = expected time between arrivals = 10 min
τ_s = expected time for service = 8 min
$\theta = 0.10$ = probability service must repeat

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$\rho' = \rho/(1 - \theta) = 0.889$
$P_n = (.111).889^n \quad n \geq 0$
$P_o = 0.111$
$P_1 = 0.099$
$P_2 = 0.088$
$P_3 = 0.078$
…

Example 13.2

Consider the data from Example 13.1, and assume the average fee per job is $100 and the material cost is $40 per job. Recall, the average time between arrivals is 10 min, and the average time per service is 8 min. Also, the shop is open 40 h per week. The data below compares the financial costs when the repeat probability goes from 0.10 to 0.05.

Input:

Shop open 40 h per week		
One-server		
Infinite capacity		
Average fee per job is $100		
Material cost per job is $40		
λ = expected arrivals per hour = 6		
Repeat probability = θ	0.10	0.05

Computations:

Jobs per week = (λ40)	240	240
Expected (runs/good unit) = $1/(1-\theta)$	1.111	1.053
Expected repeat runs per week = $\theta/(1-\theta)\lambda 40$	26.7	12.6
Expected runs per week = $1/(1-\theta)\lambda 40$	252.6	266.7
Expected fees per week ($) = $\lambda 40 \times 100$	24,000	24,000
Expected material cost per week ($) = $\lambda 40 \times 40 \times 1/(1-\theta)$	10,666	10,105

Chapter 14
Multi Servers, Repeat Service (M/M/k/θ)

Abstract This chapter explores a system with multi-servers, infinite capacity, exponential inter-arrival and service times, and where the service may need to be repeated. An example is a warehouse with several order pickers that receive customer orders. When an order is picked incorrectly it must be repeated. The probability on n units in the system is developed. Examples are presented.

14.1 Introduction

Consider a system with k servers and an infinite queue where the inter-arrival and the service times have exponential probability densities. The average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. The probability is θ that the serviced unit has a fault and thereby the service must be repeated. An example might be a busy shoe store with two salesmen and where the foot of each customer is measured prior to bringing out a pair of shoes for the customer to test. Should the customer not like the comfort of the shoe, the salesman is obliged to fetch another size or style shoe and repeat the process. The following notation applies here:

θ = probability the service must repeat
$\tau_a = 1/\lambda$ = average time between arrivals
$\tau_s = 1/\mu$ = average time to service a unit
λ = average number of arrivals per unit of time
μ = average number of units processed in a unit of time for a continuously busy service facility
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
$\rho' = \rho/[1 - \theta]$
$\rho'/k < 1$ is needed to ensure the system is in equilibrium
k = number of service facilities
n = number of units in the system $(n \geq 0)$

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_14, © Springer Science+Business Media New York 2012

Below is a list of the difference equations. Following are the corresponding equilibrium equations and then the reduced equations.

14.2 Difference Equations

$$
\begin{array}{ll}
n = 0 & P_0(t+h) = (1-\lambda h)P_0(t) + (1-\theta)\mu h P_1(t) + o(h) \\
n = (1,k-1) & P_n(t+h) = (1-\lambda h-(1-\theta)n\mu h)P_n(t) + \lambda h P_{n-1}(t) \\
 & \quad + (1-\theta)(n+1)\mu h P_{n+1}(t) + \theta n\mu h P_n(t)\ o(h) \\
n \geq k & P_n(t+h) = (1-\lambda h-(1-\theta)k\mu h)P_n(t) + \lambda h P_{n-1}(t) \\
 & \quad + (1-\theta)k\mu h P_{n+1}(t) + \theta k\mu h P_n(t) + o(h)
\end{array}
$$

14.3 Equilibrium Equations

$$
\begin{array}{ll}
n = 0 & 0 = -\lambda P_0 + (1-\theta)\mu P_1 \\
n = (1,k-1) & 0 = -(\lambda + n\mu)P_n + \lambda P_{n-1} + (1-\theta)(n+1)\mu P_{n+1} \\
n \geq k & 0 = -(\lambda + k\mu)P_n + \lambda P_{n-1} + (1-\theta)k\mu P_{n+1}
\end{array}
$$

14.4 Reduced Equations

$$
\begin{array}{ll}
0 = -\lambda P_{n-1} + (1-\theta)n\mu P_n & n = (1,k) \\
0 = -\lambda P_{n-1} + (1-\theta)k\mu P_n & n > k
\end{array}
$$

14.5 Probability on n Units in the System

For convenience, the following analysis uses the notations $\rho = \lambda/\mu$, and $\rho' = \rho/(1-\theta)$. At n = 0 to k, the reduced equations becomes the following;

$$
\begin{array}{llll}
P_0 & & = \rho'^0 P_0 & \\
P_1 & = \rho' P_0 & = \rho'^1 P_0 & \\
P_2 & = \rho'/2 P_1 & = \rho'^2/2! P_0 & \\
P_3 & = \rho'/3 P_2 & = \rho'^3/3! P_0 & \\
\ldots & & & \\
P_n & = \rho'/n P_{n-1} & = \rho'^n/n! P_0 & n = (0,k)
\end{array}
$$

When n is k + 1 and larger, the reduced equations yield the relations listed below.

$$\begin{aligned} P_{k+1} &= \rho'/kP_k &&= \rho'^{k+1}/[k!k]P_0 \\ P_{k+2} &= \rho'/kP_{k+1} &&= \rho'^{k+2}/[k!k^2]P_0 \\ &\cdots \\ P_n &= \rho'/kP_{n-1} &&= \rho'^{n}/[k!k^{n-k}]P_0 \qquad n > k \end{aligned}$$

Summarizing,

$$\begin{aligned} P_n &= \rho'^{n}/n!P_0 && n = (0,k) \\ P_n &= \rho'^{n}/[k!k^{n-k}]P_0 && n > k \end{aligned}$$

At n = k, both of the above equations are the same; and because probabilities across all values of n sum to unity, the relation below applies.

$$\begin{aligned} \Sigma_{n\geq 0}\, P_n &= P_0\{\sum_{n=0}^{k-1} \rho'^{n}/n! + \sum_{n\geq k} \rho'^{n}/[k!k^{n-k}]\} \\ &= P_0\{\sum_{n=0}^{k-1} \rho'^{n}/n! + \rho'^{k}/k! \sum_{n\geq k} \rho'^{n-k}/k^{n-k}]\} \end{aligned}$$

For equilibrium, $\rho'/k < 1$. Applying (2.2) on the above right-hand term yields,

$$\Sigma_{n\geq 0}\, P_n = P_0\{\sum_{n=0}^{k-1} \rho'^{n}/n! + \rho'^{k}/[(k\text{-}1)!(k\text{-}\rho')]\}$$

So now, the probability of n = 0 becomes:

$$P_0 = 1/\{\sum_{n=0}^{k-1} \rho'^{n}/n! + \rho'^{k}/[(k-1)!(k-\rho')]\}$$

Finally, the probability of n units in the system becomes

$$P_n = \begin{cases} \rho'^{n}/n!P_0 & n = (0, k-1) \\ \rho'^{n}/\left[k!k^{n-k}\right]P_0 & n \geq k \end{cases}$$

14.6 Expected Runs

The expected number of runs to get a good unit is obtained from the geometric distribution. Let x represent the number of runs until a good unit results when θ is the probability of a defective unit. The expected value of x becomes,

$$E(x) = 1/(1 - \theta).$$

Example 14.1
Suppose a two-service facility system with infinite capacity, and with exponential arrival and service times. The average time between arrivals is 10 min, and the average time per service is 8 min. The probability the service must repeat is $\theta = 0.20$. Some of the key probabilities and statistics associated with this system are listed below.

Input:
Two-servers
Infinite capacity
Arrival and service times are exponential
τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min
$\theta = 0.20$ = probability the service must repeat

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$\rho' = \lambda/[(1 - \theta)\mu] = 1.00$
$\rho'/2 = 0.50$
$P_n = 0.333/n! \quad n = (0,2)$
$P_n = 0.167/2^{n-2} \quad n \geq 3$
$P_o = 0.333$
$P_1 = 0.333$
$P_2 = 0.167$
$P_3 = 0.083$
…

Example 14.2
Consider the data from Example 14.1, and assume the average fee per job is \$100 and the material cost is \$40 per job. The shop is open 40 h per week. The financial data below compares the costs when the repeat probability goes from 0.20 to 0.10.

Input:
Shop open 40 h per week
Two-servers
Infinite capacity
Average fee per job is \$100
Material cost per job is \$40
λ = expected arrivals per hour = 6
Repeat probability = θ 0.20 0.10

Computations:

Expected jobs per week $= \lambda 40$	240	240
Expected runs per good unit $= 1/(1 - \theta)$	1.25	1.11
Expected repeat runs per week $= \theta/(1 - \theta)\lambda 40$	60	26.7
Expected runs per week $= 1/(1 - \theta)\lambda 40$	300	266.7
Expected fees per week (\$) $= \lambda 40 \times 100$	24,000	24,000
Expected material cost per week (\$) $= \lambda 40 \times 40 \times 1/(1 - \theta)$	12,000	10,667

Chapter 15
Tandem Queues (M/M/1: M/M/1)

Abstract This chapter considers a series of two systems where the arriving units are treated in one system and then in another system in a tandem way. This could be patients arriving to a medical center where the first system is filling out the paper and insurance forms, and the second system is receiving the medical attention. The probability on the number of units in the system, and the basic performance measures of the system are developed. Examples are presented.

15.1 Introduction

Tandem queues occur when the departing units from one system become the new arrivals to a downstream system. Tandem queues are also called queues-in-series. These are two systems, each with one server and an infinite queue capacity where the inter-arrival and the service times have exponential probability densities. The average inter-arrival time to system 1 is $1/\lambda$. The average service times are $1/\mu_1$ for system 1, and $1/\mu_2$ for system 2. An example may be the vast of citizens entering a state auto license bureau to receive a new driver's license. After passing the preliminary quiz and eye site exam, the citizen enters a queue to fill out forms and pay a fee to the state. From there, the citizen proceeds to another queue to take a road test on driving capability. It is noted that the application for two systems in series could readily be extended to three or more systems in series. The following notation applies here:

$\tau_a = 1/\lambda$ = average time between arrivals
$\tau_{s1} = 1/\mu_1$ = average time to service a unit in system 1
$\tau_{s2} = 1/\mu_2$ = average time to service a unit in system 2
λ = average number of arrivals per unit of time
μ_1, μ_2 = service rates for systems 1 and 2, respectively
$\rho_1 = \tau_{s1}/\tau_a = \lambda/\mu_1$ = utilization ratio for system 1

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_15, © Springer Science+Business Media New York 2012

$\rho_2 = \tau_{s2}/\tau_a = \lambda/\mu_2 =$ utilization ratio for system 2
$\rho_1 < 1$ and $\rho_2 < 1$ are needed to assure the systems are in equilibrium
n_1, $n_2 =$ number of units in systems 1 and 2, respectively
$n = n_1 + n_2 =$ number of units in both systems $(n \geq 0)$

15.2 Statistics for System 1

Recall the probability and statistics developed for an (M/M/1) system in Chap. 3. Since system 1 is also an (M/M/1) system and is completely independent from system 2, the system 1 probability and statistics are the same as listed in Chap. 3. These are summarized below:

$P_{n1} = \rho_1^{n1} (1 - \rho_1)\ n_1 \geq 0$
$P_{n1=0} = (1 - \rho_1)$
$Ls_1 = \rho_1$
$Lq_1 = \rho_1^2/(1 - \rho_1)$
$L_1 = Ls_1 + Lq_1 = \rho_1/(1 - \rho_1)$
$Ws_1 = 1/\mu_1$
$Wq_1 = \rho_1/[\mu_1(1 - \rho_1)]$
$W_1 = Ws_1 + Wq_1$
$Wq_1' = Wq_1/\rho_1$
$SL_1 = (1 - \rho_1)$

15.3 Output from System 1

Because the service times from system 1 are exponentially distributed with an average of $1/\mu_1$, the output rate while the facility is continuously busy is Poisson distributed with a rate of μ_1. But also, because the system is in equilibrium, the input and output distributions and the corresponding rates are the same. That is, since the inter-arrival times to system 1 are exponential with an average time of $1/\lambda$, the inter-departure times from the system are also exponential with a rate of $1/\lambda$. At the same time the number of outputs from the system is Poisson distributed with a rate of λ.

15.4 Statistics for System 2

Because the output units from system 1 are Poisson with a rate of λ, and these are the input units to system 2 downstream, system 2 also has exponential arrivals with an average time of $1/\lambda$. Also since system 2 has one server and an infinite queue

with exponential service times and with an average of $1/\mu_2$, the system is classified as (M/M/1). As long as the utilization rate ρ_2 for system 2 is less than one, the system also is in equilibrium. The probability and statistics for this system are listed below:

$P_{n2} = \rho_2^{n2}(1 - \rho_2) \quad n_2 \geq 0$
$P_{n2=0} = (1 - \rho_2)$
$Ls_2 = \rho_2$
$Lq_2 = \rho_2^2/(1 - \rho_2)$
$L_2 = Ls_2 + Lq_2 = \rho_2/(1 - \rho_2)$
$Ws_2 = 1/\mu_2$
$Wq_2 = \rho_2/[\mu_2(1 - \rho_2)]$
$W_2 = Ws_2 + Wq_2$
$Wq_2' = Wq_2/\rho_2$
$SL_2 = (1 - \rho_2)$

15.5 Number of Units in Both Systems

The number of units in both systems is $n = n_1 + n_2$. The probability of n is developed from the convolution below.

$$P_n = \sum_{n1=0}^{n} P_{n1} P_{n-n1}$$
$$= (1 - \rho_1)(1 - \rho_2) \sum_{n1=0}^{n} \rho_1^{n1} \rho_2^{n-n1}$$
$$= (1 - \rho_1)(1 - \rho_2)\rho_2^n \sum_{n1=0}^{n} (\rho_1/\rho_2)^{n1}$$

Now applying (2.7),

$$P_n = (1 - \rho_1)(1 - \rho_2)\rho_2^n[1 - (\rho_1/\rho_2)^{n+1}]/[1 - (\rho_1/\rho_2)]$$
$$= (1 - \rho_1)(1 - \rho_2)[\rho_2^{n+1} - \rho_1^{n+1}]/[\rho_{2-}\rho_1]$$

15.6 Statistics for the Total System

The statistics for the total system can now be computed. These are listed below:

$Ls = Ls_1 + Ls_2$
$Lq = Lq_1 + Lq_2$

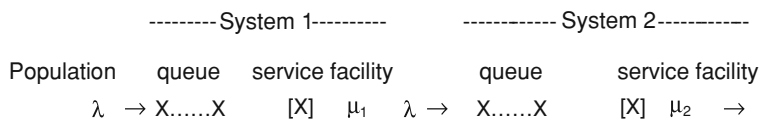

Fig. 15.1 Two queuing systems in series where the arrival units are first serviced in system 1 and then in system 2

$L = L_1 + L_2$
$Ws = Ws_1 + Ws_2$
$Wq = Wq_1 + Wq_2$
$W = W_1 + W_2$

Example 15.1
Suppose two systems in series where the departures from system 1 go directly to system 2 for service. Each system has one service facility and an infinite queue. Also, all of the arrival and service times are exponential. The average time between arrivals to system 1 is 10 min, and the average time per service is 8 min for system 1, and 5 min for system 2. See Fig. 15.1. Some of the key probabilities and statistics associated with this system are listed below.

Input:
2 systems in series (1,2)
One-server per system
Infinite capacity per system
Inter-arrival times and service times are exponential
τ_a = expected time between arrivals = 10 min
τ_{s1} = expected service time for system 1 = 8 min
τ_{s2} = expected service time for system 2 = 5 min

Computations:
System 1:

$\lambda = 1/\tau_a = 0.10$ per minute
$\mu_1 = 1/\tau_{s1} = 0.125$ per minute
$\rho_1 = \lambda/\mu_1 = 0.80$
n_1 = number of units in system 1
$P_{n1} = (.20).80^{n1} \qquad n_1 \geq 0$

System 2:

$\lambda = 1/\tau_a = 0.10$ per minute
$\mu_2 = 1/\tau_{s2} = 0.200$ per minute
$\rho_2 = \lambda/\mu_2 = 0.50$
n_2 = number of units in system 2
$P_{n2} = (.50).50^{n2} \qquad n_2 \geq 0$

Total:

$n = n_1 + n_2 =$ number of units in systems 1 and 2
$P_n = (1 - .8)(1 - .5)[.5^{n+1} - .8^{n+1}]/[.5 - .8]\ n \geq 0$

	System 1	System 2	Total
P_o	0.200	0.500	0.100
P_1	0.160	0.250	0.130
P_2	0.128	0.125	0.129
P_3	0.102	0.067	0.116
…			
Ls	0.80	0.50	1.30
Lq	3.20	0.50	3.70
L	4.00	1.00	5.00
Ws	8 min	5 min	13 min
Wq	32 min	5 min	37 min
W	40 min	10 min	50 min
Wq′	40 min	10 min	50 min
SL	0.20	0.50	

Chapter 16
Priority System, One Server, Infinite Queue (M/M/1//P)

Abstract This chapter pertains to a systems where the service discipline behaves in a preemptive priority way. The system has one server, infinite capacity, exponential inter-arrival and service times. An example is a military unit using a one-frequency radio system where the top commander can interrupt any ongoing call whenever needed. The probability on n units, and the basic performance measures of the system are developed. Examples are presented.

16.1 Introduction

This chapter concerns two populations of customers, one with high priority and the other with low priority. Should a high priority unit enter the system while a low priority unit is being serviced, the high priority unit bumps the low priority unit and takes over the service facility. Afterwards, when the bumped unit continues service, it does so from where it left off. This system is called a preemptive priority system. The inter-arrival and service times from each population follows the exponential probability density. The chapter shows how to measure the statistics for the total system and for each of the individual populations. The average arrival rates are λ_1 and λ_2 for populations 1 and 2, respectively. The average service times are $1/\mu$ for both populations. This two-priority system could be patients coming to an emergency clinic where some need immediate emergency treatment and others do not. The reader should be aware that the methods of this chapter could be extended for three or more priority populations. The method could also be extended to allow separate service times for each population.

The following notation applies here:

Populations 1 and 2 have high and low priority, respectively
τ_{a1}, τ_{a2} = average inter-arrival time for units from populations 1 and 2
λ_1, λ_2 = average number of arrivals per unit of time from populations 1 and 2

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_16, © Springer Science+Business Media New York 2012

τ_{s1}, τ_{s2} = average service time for units from populations 1 and 2
$\mu_1 = \mu_2 = \mu$ = average number of units of populations 1 and 2 that are processed in a unit of time from a continuously busy service facility
$\rho_1 = \lambda_1/\mu$ = utilization ratio for population 1
$\rho_2 = \lambda_2/\mu$ = utilization ratio for population 2
$\lambda = \lambda_1 + \lambda_2$ = arrival rate for the total system
$\rho = \rho_1 + \rho_2 = \lambda/\mu$ = utilization ratio for the total system
$\rho < 1$ is needed to assure the system is in equilibrium
n_1, n_2 = number of units in the system from populations 1 and 2
$n = n_1 + n_2$ = total number of units in the system $(n \geq 0)$

16.2 Statistics for the Total System

Since the arrival rates from populations 1 and 2 are λ_1 and λ_2, the expected arrival rate for an arbitrary unit to the total system is

$$\lambda = \lambda_1 + \lambda_2$$

So now, the time between arrivals (t_a) to the system is exponential with an average time of $E(t_a) = \tau_a = 1/\lambda$

For both the high and low priority units, the service times are also exponential and the service rates are μ. Hence μ is the service rate for the total system and thereby the associated utilization ratio is

$$\rho = \lambda/\mu$$
$$= \rho_1 + \rho_2$$

Note, the total system is classified as an (M/M/1) system, the same as described in Chap. 3. In the following analysis, n is the number of units in the total system. So using the Chap. 3 results, the probability and statistics for the total system are as below:

$P_n = \rho^n (1 - \rho) \quad n \geq 0$
$P_{n=0} = (1 - \rho)$
$Ls = \rho$
$Lq = \rho^2/(1 - \rho)$
$L = Ls + Lq = \rho/(1 - \rho)$
$Ws = 1/\mu$
$Wq = \rho/[\mu(1 - \rho)]$
$W = Ws + Wq = 1/[\mu(1 - \rho)]$
$Wq' = Wq/\rho$
$SL = (1 - \rho)$

16.3 Statistics for the Top Priority Units

In this preemptive priority system, the top priority units are using the system as though the low priority units are not involved. Should a top priority unit enter when a low priority is in service, without hesitation, the top priority unit takes over the service facility. So in this way, the top priority units are operating the system with a utilization ratio of ρ_1 and the system is classified as (M/M/1). In the following analysis, n_1 is the number of high priority units in the system. The probability and statistics for the high priority units are also taken from those developed in Chap. 3. They are the following:

$P_{n1} = \rho_1^{n1}\ (1 - \rho_1) \quad n_1 \geq 0$
$P_{n1=0} = (1 - \rho_1)$
$Ls_1 = \rho_1$
$Lq_1 = \rho_1^2/(1 - \rho_1)$
$L_1 = Ls_1 + Lq_1 = \rho_1/(1 - \rho_1)$
$Ws_1 = 1/\mu_1$
$Wq_1 = \rho_1/[\mu_1(1 - \rho_1)]$
$W_1 = Ws_1 + Wq_1 = 1/[\mu_1(1 - \rho_1)]$
$Wq_1' = Wq_1/\rho_1$
$SL_1 = (1 - \rho_1)$

16.4 Statistics for the Low Priority Units

In this preemptive system, the low priority units are allowed to use the service facilities as long as there are no high priority units in the system. In essence, they are delegated to remain in the queue while the high priority units are being serviced. Below shows how to measure the statistics for the low priority units.

16.5 Expected Units in Service (Ls), Queue (Lq) and System (L)

Since,

$$Ls = Ls_1 + Ls_2,$$

and

$$Ls = \rho$$
$$= \rho_1 + \rho_2$$

then

$$Ls_2 = Ls - Ls_1 = \rho_2$$

In the same way,

$$Lq_2 = Lq - Lq_1$$

and thereby,

$$L_2 = Ls_2 + Lq_2$$

16.6 Expected Time in Service (Ws), Queue (Wq) and System (W)

The expected time is the service facility for a low priority unit is merely,

$$Ws_2 = 1/\mu$$

To obtain the expected time in the queue for a low priority item, the following relation for the expected queue time for the total system is needed. This is

$$Wq = Wq_1(\lambda_1/\lambda) + Wq_2(\lambda_2/\lambda)$$

Thereby,

$$Wq_2 = [Wq - Wq_1(\lambda_1/\lambda)]/(\lambda_2/\lambda)$$

Using Ws_2 and Wq_2, the total time in the system becomes,

$$W_2 = Ws_2 + Wq_2$$

16.7 Expected Time in Queue (Wq′) for a Delayed Item

To compute the wait time in the queue, given the unit is delayed, is obtained from the relation below,

$$Wq' = Wq'_1(\lambda_1/\lambda) + Wq'_2(\lambda_2/\lambda)$$

Hence,

$$Wq'_2 = [Wq' - Wq'_1(\lambda_1/\lambda)]/(\lambda_2/\lambda)$$

Example 16.1
Suppose a one-service facility system with infinite capacity, and with two input populations with exponential arrival and service times. One population has high preemptive priority over the other low priority population. The average time

between arrivals is 25 min for population 1, and 25 min for population 2. The average time per service is 10 min for populations 1 and 2. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One server
Infinite capacity
2 input populations (1,2)
The population inter-arrival and service times are exponential
Population 1 has high priority

$$\tau_{a1} = \text{expected time between arrivals for population 1} = 25\,\text{min}$$
$$\tau_{s1} = \text{expected service time for population 1} = 10\,\text{min}$$

Population 2 has low priority

$$\tau_{a2} = \text{expected time between arrivals for population 2} = 25\,\text{min}$$
$$\tau_{s2} = \text{expected service time for population 2} = 10\,\text{min}$$

Computations:

	Total system	high priority	low priority
τ_a	12.5 min	25 min	25 min
τ_s	10 min	10 min	10 min
λ $(1/\tau_a)$	0.08/min	0.04/min	0.04/min
μ $(1/\tau_s)$	0.10/min	0.10/min	0.10/min
ρ (λ/μ)	0.80	0.40	0.40
P_n	$0.20(0.80^n)$	$0.60(0.40^{n1})$	
P_o	0.2000	0.6000	
P_1	0.1600	0.2400	
P_2	0.1280	0.0960	
P_3	0.1024	0.0384	
…			
Ls	0.800	0.400	0.400
Lq	3.200	0.267	2.933
L	4.000	0.667	3.333
Ws	10 min	10 min	10 min
Wq	40 min	6.67 min	73.33 min
W	50 min	16.67 min	83.33 min
Wq′	50 min	16.67 min	83.33 min
SL	0.20	0.60	

Chapter 17
Priority, One Server, Arbitrary Service (M/G/1//P)

Abstract This chapter explores a system where the service discipline is preemptive priority. The system has one-server, infinite capacity, exponential inter-arrival times and arbitrary service times. An example is a clinic where some patients need immediate emergency treatment and others do not. The emergency patients override the non-emergency patients. The probability on n units, and the basic performance measures of the system are developed. Examples are presented.

17.1 Introduction

This chapter concerns two populations of customers, one with high priority and the other with low priority. Should a high priority unit enter the system while a low priority unit is being serviced, the high priority unit bumps the low priority unit and takes over the service facility. Subsequently, when the bumped unit continues service, it does so from where it left off. This system is called a preemptive priority system. The inter-arrival times follows the exponential probability density. The service times have arbitrary distributions that could be discrete or continuous. The chapter shows how to measure the statistics for the total system and for each of the individual populations. The average arrival rates are λ_1 and λ_2 for populations 1 and 2, respectively. The average service times are $1/\mu_1$ and $1/\mu_2$, and the associated variances are σ_1^2 and σ_2^2 for populations 1 and 2, respectively. This two priority system could be calls for help in a police beat where some calls need immediate emergency treatment and others do not. The methods of this chapter could be extended for three or more priority populations.

The following notation applies here:

Populations 1 and 2 have high and low priority, respectively

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_17,

τ_{a1}, τ_{a2} = average inter-arrival time for units from populations 1 and 2
λ_1, λ_2 = average number of arrivals per unit of time from populations 1 and 2
τ_{s1}, τ_{s2} = average service time for units from populations 1 and 2
σ_1^2, σ_2^2 = variance of service times from populations 1 and 2
$\mu_1 = 1/\tau_{a1}$, $\mu_2 = 1/\tau_{a2}$
$\rho_1 = \lambda_1/\mu_1$ = utilization ratio for population 1
$\rho_2 = \lambda_2/\mu_2$ = utilization ratio for population 2
$\lambda = \lambda_1 + \lambda_2$ = arrival rate for the total system
τ_a = average inter-arrival time for an arbitrary unit
τ_s = average service time for an arbitrary unit
$\mu = 1/\tau_s$
$\rho = \rho_1 + \rho_2 = \lambda/\mu$ = utilization ratio for the total system
$\rho < 1$ is needed to assure the system is in equilibrium
n_1, n_2 = number of units in the system from populations 1 and 2
$n = n_1 + n_2$ = total number of units in the system ($n \geq 0$)

17.2 Statistics for the Total System

Since the arrivals from populations 1 and 2 follow a Poisson distribution with rates λ_1 and λ_2, the total arrivals are Poisson, and the expected arrival rate for an arbitrary unit to the total system is

$$\lambda = \lambda_1 + \lambda_2$$

So now, the time between arrivals (t_a) to the system is exponential with an average time of $E(t_a) = \tau_a = 1/\lambda$

17.3 Expected Time and Variance of Time in Service ($1/\mu$) and σ^2

For an arbitrary unit, the expected service time is the following:

$$\begin{aligned} E(t_s) &= 1/\mu \\ &= (1/\mu_1)(\lambda_1/\lambda) + (1/\mu_2)(\lambda_2/\lambda) \\ &= (\rho_1 + \rho_2)/\lambda \end{aligned}$$

and thereby the corresponding service rate is

$$\mu = 1/E(t_s)$$

To compute the associated variance, $E(t_s^2)$ is needed. This requires using the variances for populations 1 and 2 in the following way. For population 1, the variance of the service time (t_{s1}) is computed as below,

$$V(t_{s1}) = E\left(t_{s1}^2\right) - E(t_{s1})^2$$

Hence,

$$E\left(t_{s1}^2\right) = V(t_{s1}) + E(t_{s1})^2$$

In the same way, the corresponding relation for a unit from population 2 becomes,

$$E\left(t_{s2}^2\right) = V(t_{s2}) + E(t_{s2})^2$$

Thereby, for an arbitrary unit in the system, the expression below applies,

$$E\left(t_s^2\right) = (\sigma_1^2 + 1/\mu_1^2)(\lambda_1/\lambda) + (\sigma_2^2 + 1/\mu_2^2)(\lambda_2/\lambda)$$

So now the variance for an arbitrary unit in the system is computed by,

$$V(t_s) = \sigma^2 = E\left(t_s^2\right) - E(t_s)^2$$

17.4 Statistics for an Arbitrary Unit in the System

Now with the parameters, λ, μ, ρ and σ^2 known for the total system, the results from Chap. 9 can be used to measure the statistics. These are the following;

$P_o = 1 - \lambda/\mu = (1 - \rho)$
$Ls = \rho$
$Lq = [\lambda^2\sigma^2 + \rho^2]/[2(1 - \rho)]$
$L = [\lambda^2\sigma^2 + 2\rho - \rho^2]/[2(1 - \rho)]$
$Ws = Ls/\lambda = 1/\mu$
$Wq = Lq/\lambda$
$W = Ws + Wq$
$Wq' = Wq/\rho$
$SL = P_0 = 1 - \rho$

17.5 Statistics for the Top Priority Units

In this preemptive priority system, the top priority units are using the system as though the low priority units are not involved. Should a top priority unit enter when a low priority is in service, without hesitation, the top priority unit takes over

the service facility. So in this way, the top priority units are operating the system with a utilization ratio of ρ_1 and the system is classified as (M/G/1). The parameters for the top priority units are λ_1, μ_1, ρ_1 and σ_1^2. In the following analysis, n_1 is the number of high priority units in the system. The probability and statistics for the high priority units are also taken from those developed in Chap. 9. They are the following:

$P_{n1=0} = 1 - \lambda_1/\mu_1 = (1 - \rho_1)$
$Ls_1 = \rho_1$
$Lq_1 = [\lambda_1^2\sigma_1^2 + \rho_1^2]/[2(1 - \rho_1)]$
$L_1 = [\lambda_1^2\sigma_1^2 + 2\rho_1 - \rho_1^2]/[2(1 - \rho_1)]$
$Ws_1 = Ls_1/\lambda_1 = 1/\mu_1$
$Wq_1 = Lq_1/\lambda_1$
$W_1 = Ws_1 + Wq_1$
$Wq_1' = Wq_1/\rho_1$
$SL_1 = P_{n1=0} = 1 - \rho_1$

17.6 Statistics for the Low Priority Units

In this preemptive system, the low priority units are allowed to use the service facilities as long as there are no high priority units in the system. In essence, they are delegated to remain in the queue while the high priority units are being serviced. Below shows how to measure the statistics for the low priority units.

17.7 Expected Units in Service (Ls), Queue (Lq) and System (L)

Since,

$$Ls = Ls_1 + Ls_2,$$

and

$$Ls = \rho$$
$$= \rho_1 + \rho_2$$

then

$$Ls_2 = Ls - Ls_1 = \rho_2$$

In the same way,

$$Lq_2 = Lq - Lq_1$$

and thereby,

$$L_2 = Ls_2 + Lq_2$$

17.8 Expected Time in Service (Ws), Queue (Wq) and System (W)

The expected time is the service facility for a low priority unit is merely,

$$Ws_2 = 1/\mu_2$$

To obtain the expected time in the queue for a low priority item, the following relation for the expected queue time for the total system is needed. This is

$$Wq = Wq_1(\lambda_1/\lambda) + Wq_2(\lambda_2/\lambda)$$

Thereby,

$$Wq_2 = [Wq - Wq_1(\lambda_1/\lambda)]/(\lambda_2/\lambda)$$

Using Ws_2 and Wq_2, the total time in the system becomes,

$$W_2 = Ws_2 + Wq_2$$

17.9 Expected Time in Queue (Wq′) for a Delayed Item

To compute the wait time in the queue, given the unit is delayed, is obtained from the relation below,

$$Wq' = Wq'_1(\lambda_1/\lambda) + Wq'_2(\lambda_2/\lambda)$$

Hence,

$$Wq'_2 = [Wq' - Wq'_1(\lambda_1/\lambda)]/(\lambda_2/\lambda)$$

Example 17.1

Suppose a one service facility system with infinite capacity, and with two input populations with exponential arrivals. One population has high preemptive priority over the other low priority population. The average time between arrivals is 25 min for population 1, and 25 min for population 2. The average time per service is 10 min for populations 1 and 2, and the associated variances are $\sigma_1^2 = 1$

and $\sigma_2^2 = 1$, respectively. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One-server
Infinite capacity
2 arrival populations (1,2)
Inter-arrival times are exponential
Population 1

High preemptive priority
τ_{a1} = expected time between arrivals = 25 min
τ_{s1} = expected service time = 10 min
σ_1^2 = variance of service times = 1 min^2

Population 2

Low priority
τ_{a2} = expected time between arrivals = 25 min
τ_{s2} = expected service time = 10 min
σ_2^2 = variance of service times = 1 min^2

Computations:
Population 1:

$\lambda_1 = 1/\tau_{a1} = 0.04$/min
$\mu_1 = 1/\tau_{s1} = 0.10$/min
$E(ts_1^2) = 101$

Population 2:

$\lambda_2 = 1/\tau_{a2} = 0.04$/min
$\mu_2 = 1/\tau_{s2} = 0.10$/min
$E(ts_2^2) = 101$

Total system:

$\lambda = \lambda_1 + \lambda_2 = 0.08$/min
$\tau_a = 1/\lambda = 12.5$ min
$E(ts) = \tau_s = \tau_{s1}(\lambda_1/\lambda) + \tau_{s2}(\lambda_2/\lambda) = 10$ min
$\mu = 1/\tau_s = 0.10$/min
$E(ts^2) = 101$
$\sigma^2 = E(ts^2) - E(ts)^2 = 1.0$ min^2

	Total system	Population 1 (high priority)	Population 2 (low priority)
τ_a	12.5 min	25 min	25 min
τ_s	10 min	10 min	10 min
λ	0.08/min	0.04/min	0.04/min
μ	0.10/min	0.10/min	0.10/min
σ^2	1.00 min^2	1 min^2	1 min^2
ρ	0.80	0.40	0.40
P_o	0.200	0.600	
Ls	0.800	0.400	0.400
Lq	1.616	0.135	1.481
L	2.416	0.535	1.881
Ws	10 min	10 min	10 min
Wq	20.20 min	3.375 min	37.025 min
W	30.20 min	13.375 min	47.025 min
Wq′	25.25 min	8.437 min	42.063 min
SL	0.20	0.60	

Chapter 18
One Server, Constant Service (M/D/1)

Abstract This chapter concerns a one-server system with infinite capacity where the inter-arrival times are exponential and the service times are constant. An example is the cars arriving to a carwash where the service time is always the same. The probability on n units in the system, and the related performance measures are developed. Examples are presented.

18.1 Introduction

This is a system with one server and infinite queue where the time between arriving customers is exponentially distributed with an average of $1/\lambda$. The service time is constant and is always $1/\mu$. An example could be a clothes washing machine in an apartment building where the total cycle time is always the same. The following notation applies here:

$\tau_a = 1/\lambda$ = average time between arrivals
$\tau_s = 1/\mu$ = fixed time to service a unit
$\sigma^2 = 0$ = variance of the time to service a unit
λ = average number of arrivals per unit of time
$\mu = 1/\tau_a$
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
$\rho < 1$ is needed to assure the system is in equilibrium
n = number of units in the system
P_n = probability of n units in the system $\quad (n \geq 0)$

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_18, © Springer Science+Business Media New York 2012

18.2 Summary of the Statistical Measures

This system is one with exponential input times and constant service times. There is one service facility and an infinite queue. This system conforms with the Pollaczek–Khintchin formula presented in Chap. 9 where the input times are exponential and the output times are arbitrary. In the system of this chapter, the service time is constant and as such, the variance is zero, i.e. $\sigma^2 = 0$. With the parameters: λ, μ, ρ and $\sigma^2 = 0$, the statistical measures developed in Chap. 9 apply here. These are listed below using $\sigma^2 = 0$.

$P_o = 1 - \lambda/\mu = (1 - \rho)$
$Ls = \rho$
$Lq = \rho^2/[2(1 - \rho)]$
$L = [2\rho - \rho^2]/[2(1 - \rho)]$
$Ws = Ls/\lambda = 1/\mu$
$Wq = Lq/\lambda$
$W = Ws + Wq$
$Wq' = Wq/\rho$
$SL = P_0 = 1 - \rho$

Example 18.1
Suppose a one service facility system with infinite capacity, and with exponential arrivals. The average time between arrivals is 10 min. The time per service is always 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One-server
Infinite capacity
Inter-arrival times are exponential
Service times are constant
τ_a = expected time between arrivals = 10 min
τ_s = service time = 8 min
σ^2 = variance of service time = 0

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\rho = \lambda/\mu = 0.80$
$P_o = 0.2000$
$Ls = 0.80$
$Lq = 1.60$
$L = 2.40$

Ws = 8 min
Wq = 16 min
W = 24 min
Wq′ = 20 min
SL = 0.20

18.3 The Probability Distribution of n

For this constant service time system, it is also possible to generate the probability of n units in the system. To begin, consider two moments in time, t and $t' = t + 1/\mu$. Note, if an unit is in the service facility at time t, the unit will complete its service by t′, since the service time is $1/\mu$ and is constant. So, the following probability statements are defined:

$P_n(t)$ = probability of n units in the system at time t
$P_n(t')$ = probability of n units in the system at time t′
Note, because of equilibrium, $P_n(t') = P_n(t) = P_n$

$P[j(t, t')]$ = probability of j arrivals from t to t′

$$= \rho^j \exp(-\rho)/j!\ j = 0, 1, 2, \ldots$$

where j Is Poisson with $E(j) = \lambda/\mu = \rho$.

Note j is the number of arrivals in the length of time $1/\mu$, between t and t′.

To begin in the pursuit to finding P_n, recall:

$P_0 = 1 - \rho$.

n = 0

The difference equation for n = 0 is the following:

$P_0(t') = P_0(t)\ P[0(t,t')] + P_1(t)\ P[0(t,t')]$

Since $P_0(t') = P_0(t) = P_0$, the relation below is formed:

$P_0 = P_0 \exp(-\rho) + P_1 \exp(-\rho)$

With some algebra,

$P_1 = (1 - \rho)[\exp(\rho) - 1]$

n = 1

The difference equation for n = 1 is the following:

$$P_1(t') = P_0(t)\ P[1(t,t')] + P_1(t)\ P[1(t,t')] + P_2(t)\ P[0(t,t')]$$

Therefore,

$$P_1 = P_0\ \rho\ \exp(-\rho) + P_1\ \rho\ \exp(-\rho) + P_2\exp(-\rho)$$

Again, with algebra,

$$P_2 = (1 - \rho)\ [\exp(2\rho) - \rho\ \exp(\rho) - \exp(\rho)]$$

n ≥ 2

For n of two or larger, the following expression for the probability is formulated:

$$P_n = (1 - \rho) \sum_{j=1}^{n} (-1)^{n-j} \exp(j\rho)\ [(j\rho)^{n-j}/(n - j)! + (j\rho)^{n-j-1}/(n - j - 1)!]\ n \geq 2$$

When j = n, the right-hand term in the above equation is ignored.

Example 18.2
Continuing with the above example, $\lambda = 0.10$ per minute, $\mu = 1/\tau_s = 0.125$ per minute and $\rho = \lambda/\mu = 0.80$. The probabilities for n = 0, 1 and 2 are listed below:

$P_0 = (1 - \rho) = 0.200$
$P_1 = (1 - \rho)[\exp(\rho) - 1] = 0.245$
$P_2 = (1 - \rho)\ [\exp(2\rho) - \rho\ \exp(\rho) - \exp(\rho)] = 0.189$
…

Example 18.3
The table below gives comparative results for $\rho = 0.1$ to 0.9 when one service facility, k = 1, exponential arrival times, constant service times and infinite queue capacity. The measures listed are P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL. For simplicity, the average service time is $\tau_s = 1.00$, and thereby Ws = 1.00 for all situations.

K	ρ	cov	P_0	Lq	Ls	L	Wq	Ws	W	Wq′	SL
1	0.1	0.0	0.90	0.01	0.10	0.11	0.06	1.00	1.06	0.56	0.90
1	0.2	0.0	0.80	0.02	0.20	0.23	0.12	1.00	1.13	0.63	0.80
1	0.3	0.0	0.70	0.06	0.30	0.36	0.21	1.00	1.21	0.71	0.70
1	0.4	0.0	0.60	0.13	0.40	0.53	0.33	1.00	1.33	0.83	0.60
1	0.5	0.0	0.50	0.25	0.50	0.75	0.50	1.00	1.50	1.00	0.50
1	0.6	0.0	0.40	0.45	0.60	1.05	0.75	1.00	1.75	1.25	0.40
1	0.7	0.0	0.30	0.82	0.70	1.52	1.17	1.00	2.17	1.67	0.30
1	0.8	0.0	0.20	1.60	0.80	2.40	2.00	1.00	3.00	2.50	0.20
1	0.9	0.0	0.10	4.05	0.90	4.95	4.50	1.00	5.50	5.00	0.10

The table above can be used for any queuing system with one server and constant service time. For example, if the average service time is $\tau_s = 8$ min, and the utilization ratio was $\rho = 0.80$, as in Example 18.1, all the measures listed above are the same, with a minor adjustment to the wait time measures. For this situation, $Wq = 2.00 \times \tau_s = 16.00$ min, $Ws = 1.00 \times \tau_s = 8.00$ min, $W = 3.00 \times \tau_s = 24.00$ min, and $Wq' = 2.50 \times \tau_s = 20.00$ min.

Chapter 19
Exponential Arrivals, Erlang Service (M/E2/1)

Abstract This chapter explores a one-server system with infinite capacity, exponential inter-arrival times and Erlang 2-stage service times. Could be a jogging shoe manufacturer that uses a mold (called a last) to produce a shoe of a certain size and width. The arrival time between demands for the mold is exponential, and the time to use the mold on the shoe is Erlang. For an infinite capacity system, the performance measures are generated. For a finite capacity system, matrix methods are introduced and the chapter shows how to compute the probability of n units in the system, and also the performance measures. The chapter also shows how to extend the matrix method to compute the probabilities for an infinite capacity system. Examples are presented.

19.1 Introduction

Suppose a system with one server and where the inter-arrival times have exponential probability densities, and the service times have a 2-stage Erlang probability density. Further, the average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This could be the rental of a jack-hammer at a hardware rental shop. For brevity, this chapter presents the 2-stage Erlang service system. But the reader should recognize that the results given in this chapter could readily be extended for a k-stage Erlang service system.

19.2 Connection Between the Exponential and Erlang Distributions

Recall from Chap. 2, where the connection between the exponential and Erlang distributions is given. If y is exponentially distributed with mean $1/\theta$, and if $x = (y_1 + \ldots + y_k)$, then x is Erlang with mean and variance, $E(x) = k/\theta$ and

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_19, © Springer Science+Business Media New York 2012

$V(x) = k/\theta^2$, respectively. In this chapter, however, the variables are t and t_s, where t is exponential with a mean of $1/(2\mu)$ and $t_s = (t_1 + t_2)$. Thereby, $k = 2$, t_s is Erlang where the mean is $E(ts) = 1/\mu$ and the variance is $\sigma^2 = 1/(2\mu^2)$.

In summary, the system of this chapter has one service facility, the arrival times are exponential, the service times are Erlang with a known mean and variance, and the queue capacity is infinite. Hence, the system conforms with the Pollaczek–Khintchin formula that is presented in Chap. 9.

The following notation applies here:

$\tau_a = 1/\lambda$ = average time between arrivals
λ = average number of arrivals per unit of time
$k = 2$ = Erlang parameter
$\tau_s = \tau_1 + \tau_2 = 1/(2\mu) + 1/(2\mu) = 1/\mu$ = average time to service a unit
$\sigma^2 = 1/(2\mu^2)$ = variance of the service time
$\mu = 1/\tau_s$
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
$\rho < 1$ is needed to assure equilibrium
n = number of units in the system ($n \geq 0$)

19.3 Measuring the Summary Statistics

So now, the summary statistics for this system can be computed using the results developed in Chap. 9. These are the following:

$P_o = 1 - \lambda/\mu = (1 - \rho)$
$Ls = \rho$
$Lq = [\lambda^2\sigma^2 + \rho^2]/[2(1 - \rho)]$
$L = [\lambda^2\sigma^2 + 2\rho - \rho^2]/[2(1 - \rho)]$
$Ws = Ls/\lambda = 1/\mu$
$Wq = Lq/\lambda$
$W = Ws + Wq$
$Wq' = Wq/\rho$
$SL = P_0 = 1 - \rho$

Example 19.1
Suppose a one service facility system with infinite capacity, and with exponential arrivals. The average time between arrivals is 10 min. The service times are Erlang with $k = 2$ and the average time per service is 8 min. Some of the key probabilities and statistics associated with this system are listed below.

Input:
One-server
Infinite capacity
Inter-arrival times are exponential
Service times are Erlang with $k = 2$

τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\lambda = 60/\tau_a = 6$ per hour
$\mu = 60/\tau_s = 7.5$ per hour
$\sigma^2 = 1/(2\mu^2) = 32$ minute^2
$\rho = \lambda/\mu = 0.80$
$P_o = 0.20$
Ls = 0.80
Lq = 2.40
L = 3.20
Ws = 8 min
Wq = 24 min
W = 32 min
Wq′ = 30 min

19.4 Finding the Probability of n Units in the System

With a bit more effort, it is also possible to compute the probabilities for this system. To begin, it is necessary to artificially set the capacity of the system to a finite size of N. The method presented here will later show how to find the value of N that emulates an infinite capacity system.

The following notation is now added to those listed earlier for this system.

N = an artificial limit on the maximum units allowed in the system
(n,j) = states of the system
n = number of units in the system n = (0,N)
j = 0 at n = 0; j = 1, 2 at n ≥ 1 to identify the stage of service

Note for this system, the states have two arguments, (n,j), where n = the number of units in the system, and j identifies the stage of service from 1 to 2. Recall from Chap. 2 where each stage is exponential and the two exponential stages added together give the Erlang shape. The reader should be aware that when the service times are Erlang with three or more stages, the method of solution is merely an extension of the procedure shown here.

Below is a list of the difference equations and the corresponding equilibrium equations. The equilibrium equations are needed to develop the probability and statistical measures for the system.

19.5 Difference Equations

(n,j)		
(0,0)		$P_{00}(t + h) = (1 - \lambda h)P_{00}(t) + 2\mu h P_{12}(t) + o(h)$
(1,1)		$P_{11}(t + h) = (1 - \lambda h - 2\mu h)P_{11}(t) + \lambda h P_{00}(t) + 2\mu h$ $P_{22}(t) + o(h)$
(1,2)		$P_{12}(t + h) = (1 - \lambda h - 2\mu h)P_{12}(t) + 2\mu h P_{11}(t) + o(h$
(n,1)	$n = (2,N - 1)$	$P_{n1}(t + h) = (1 - \lambda h - 2\mu h)P_{n1}(t) + \lambda h P_{n-1,1}(t) + 2\mu h$ $P_{n+1,2}(t) + o(h)$
(n,2)	$n = (2,N - 1)$	$P_{n2}(t + h) = (1 - \lambda h - 2\mu h)P_{n2}(t) + \lambda h P_{n-1,2}(t) + 2\mu h$ $P_{n1}(t) + o(h$
(N,1)		$P_{N1}(t + h) = (1 - 2\mu h)P_{N1}(t) + \lambda h P_{N-1,1}(t)$
(N,2)		$P_{N2}(t + h) = (1 - 2\mu h)P_{N2}(t) + \lambda h P_{N-1,2}(t) + 2\mu h$ $P_{N1}(t) + o(h$

19.6 Equilibrium Equations

(n,j)		
(0,0)		$0 = -\lambda P_{00} + 2\mu P_{12}$
(1,1)		$0 = -(\lambda + 2\mu)P_{11} + \lambda P_{00} + 2\mu P_{22}$
(1,2)		$0 = -(\lambda + -2\mu)P_{12} + 2\mu P_{11}$
(n,1)	$n = (2,N - 1)$	$0 = -(\lambda + 2\mu)P_{n1} + \lambda P_{n-1,1} + 2\mu P_{n+1,2}$
(n,2)	$n = (2,N - 1)$	$0 = -(\lambda + -2\mu)P_{n2} + \lambda P_{n-1,2} + 2\mu P_{n1}$
(N,1)		$0 = -(2\mu)P_{N1} + \lambda P_{N-1,1}$
(N,2)		$0 = -(2\mu)P_{N2} + \lambda P_{N-1,2} + 2\mu P_{N1}$

19.7 Matrix Solution

To solve for the probabilities, P_{nj}, requires matrix methods as will be shown below. To illustrate, a small example where $N = 2$ is used for simplicity. When $N = 2$, the unknown probabilities are: $P_{00}, P_{11}, P_{12}, P_{21}, P_{22}$, and the equilibrium equations become:

(0,0)	$0 = -\lambda P_{00} + 2\mu P_{12}$
(1,1)	$0 = -(\lambda + 2\mu)P_{11} + \lambda P_{00} + 2\mu P_{22}$
(1,2)	$0 = -(\lambda + 2\mu)P_{12} + 2\mu P_{11}$
(2,1)	$0 = -(2\mu)P_{21} + \lambda P_{11}$
(2,2)	$0 = -(2\mu)P_{22} + \lambda P_{12} + 2\mu P_{21}$

19.8 Zero = Zero

When dealing with the more complex set of equilibrium equations, like above, it is good practice to sum all the equations to ensure that $0 = 0$. Since the sum of the left-hand-side of the five equations sum to zero, the sum for the right-hand-side should also be zero. This is needed to assure that all elements are correctly

installed in the equations. This check is helpful to locate any elements that should not be included and/or identify any that are missing. The analyst needs to do this by checking that all the elements with a negative value have a corresponding element with a positive value.

19.9 AP = BP_{00}

For an N = 2 system, that has five equations, only the first four equations are needed in the matrix form that are shown below.

$AP = BP_{00}$

A is a 4 × 4 matrix, P and B are 4 × 1 vectors. P_{00} is a probability and is a single 1 × 1 term. For this example, the elements of the matrices are listed below:

19.10 A, P and B

The matrix A is: $\begin{bmatrix} 0 & 2\mu & 0 & 0 \\ -(\lambda+2\mu) & 0 & 0 & 2\mu \\ 2\mu & -(\lambda+2\mu) & 0 & 0 \\ \lambda & 0 & -2\mu & 0 \end{bmatrix}$

The vector P is: $\begin{bmatrix} P_{11} \\ P_{12} \\ P_{21} \\ P_{22} \end{bmatrix}$

and the vector B is: $\begin{bmatrix} \lambda \\ -\lambda \\ 0 \\ 0 \end{bmatrix}$

19.11 Solving for the Probabilities

Solving for the vector P requires the inverse of A, as below:

$P = A^{-1} B P_{00}$

$= Q P_{00}$

The vector Q has elements: q_{11}, q_{12}, q_{21} and q_{22}. Further, another q element (q_{00}) can be assigned to the probability, P_{00} since, $P_{00} = q_{00}P_{00}$. Hence, $q_{00} = 1$. Since the sum of all the probabilities equal unity, the sum of the q elements are used as below to compute P_{00}:

$$P_{00} = 1/(q_{00} + q_{11} + q_{12} + q_{21} + q_{22})$$

So now, all of the probabilities can be computed as follows:

$$P_{11} = q_{11}P_{00}$$
$$P_{12} = q_{12}P_{00}$$
$$P_{21} = q_{21}P_{00}$$
$$P_{22} = q_{22}P_{00}$$

The probabilities P_n are now obtained as below:

$$P_0 = P_{00}$$
$$P_1 = P_{11} + P_{12}$$
$$P_2 = P_{21} + P_{22}$$

Note where the sum, $P_0 + P_1 + P_2 = 1$.

19.12 When n = (0,N)

In the general case, when N is any positive integer, the method to solve for the probabilities follows the same pattern. First, the equilibrium equations (less the final one) are placed in the matrix form $AP = BP_{00}$, and the inverse of A yields the vector Q. The sum of the Q elements (plus q_{00}) yields the value of P_{00}, as shown earlier, from where all the probabilities P_{nj} are derived. Finally, the probabilities of P_n become:

$$P_0 = P_{00}$$
$$P_n = P_{n1} + P_{n2} \quad n = (1,N)$$

19.13 Lambda and Rho Effective

Recall from Chap. 4 where lambda effective is computed as below:

$$\lambda_e = \lambda[1 - P_N] = \text{“lambda effective”}$$
$$\rho_e = \lambda_e/\mu = \text{“rho effective”}$$

In this context,

λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ = utilization ratio,
ρ_e = effective utilization ratio, and is less than one.

19.14 Probability and Statistics for an Infinite Capacity System

As λ_e/λ approaches one, the system statistics and probabilities that are calculated for this finite capacity system are very close to a system with infinite capacity. At the outset, the analyst can set $N = 20$, say, and then measure the above ratio. If λ_e/λ is too low, then N is increased by increments of 5, say, until the accuracy desired is met. In this way, this system that artificially sets a limit on the capacity can be used to find the results needed for a system with infinite capacity. As ρ approaches one, N increases.

19.15 Expected Number of Units in the Service Facility (Ls)

The expected number of units in the service facility becomes:

$$\text{Ls} = \sum_{n=1}^{N} P_n$$

19.16 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained as below,

$$\text{Lq} = \sum_{n=1}^{N} (n-1)P_n$$

19.17 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$$\text{L} = \text{Ls} + \text{Lq}$$

19.18 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$$\text{Ws} = \text{Ls}/\lambda_e$$
$$\text{Wq} = \text{Lq}/\lambda_e$$
$$\text{W} = \text{L}/\lambda_e = \text{Ws} + \text{Wq}$$

19.19 Expected Time in the Queue Given a Delay (Wq')

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. The probabilities for these events are:

$P(D') = P_o$
$P(D) = (1 - P_o)$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given a delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$$

Since $W_{q|D'} = 0$,

$$Wq' = W_{q|D} = Wq/P(D) = Wq/(1 - P_0)$$

19.20 Service Level and Loss Probability

The service level (SL) is the probability an arrival to the system is not delayed in the queue, and this is simply P_0. The loss probability (Ploss) is the probability a new arrival is lost because the system capacity is too small. This is merely P_N, the probability the system is full, where any new arrival is blocked from entering. Hence,

$SL = P_0$
$Ploss = P_N$

Note, Ploss is another measure on how high to set N so that the probability and statistics results are close enough to an infinite capacity system. When Ploss is less than 0.005, say, then N is adequate.

Example 19.2
Assume a system where arrivals are exponential with an average time of 1.0, and the service is Erlang with $k = 2$ stages and has an average time of 0.5. Hence, $\lambda = 1.0$ and $\mu = 2.0$. To illustrate, using $N = 2$ as the system capacity, the matrices of interest are shown below.

Input:
One-server
Finite capacity with $N = 2$
Inter-arrival times are exponential
Service times are Erlang with $k = 2$
Expected arrival rate is $\lambda = 1.0$
Expected service rate is $\mu = 2.0$

Computations:
The matrices A and B are formulated.
Then the vector $Q = A^{-1}B$ is generated.
Finally, all the probabilities P_{nj} are computed.

Below shows all the computations.

$$A = \begin{bmatrix} 0 & 4 & 0 & 0 \\ -5 & 0 & 0 & 4 \\ 4 & -5 & 0 & 0 \\ 1 & 0 & -4 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

Now, taking the inverse of A and solving for Q yields,

$$Q = \begin{bmatrix} 0.3125 \\ 0.2500 \\ 0.0781 \\ 0.1406 \end{bmatrix}$$

From here, the probability P_{00} is computed as below:

$$P_{00} = 1/[1.000 + 0.3125 + 0.2500 + 0.0781 + 0.1406] = 0.5614$$

Hence,

$P_{11} = 0.1754$
$P_{12} = 0.1404$
$P_{21} = 0.0438$
$P_{22} = 0.0789$

Finally, the state probabilities are the following:

$P_0 = 0.5614$
$P_1 = 0.3158$
$P_2 = 0.1227$

Note, in this system with $N = 2$, the probability of a lost arrival is $Ploss = P_2 = 0.1227$. Since this is high, an approximation to an infinite capacity system would require a much larger parameter N. Perhaps, $N = 10$ to start with.

Example 19.3

The coefficient of variation for an Erlang variable with k stages is $cov = 1/\sqrt{k}$. The table below lists the cov values for k ranging from 1 to 9. Note, when $k = 1$, $cov = 1.00$ and the Erlang is the same as the exponential, and when $k = 9$, $cov = 0.33$, whereby the Erlang distribution looks much like a normal distribution.

Erlang k	cov
1	1.00
2	0.71
3	0.58
4	0.50
5	0.45
6	0.41
7	0.38
8	0.35
9	0.33

Example 19.4

The table below gives comparative results for a queuing system with one service facility, where the utilization ratio is $\rho = 0.50$, the arrival times are exponential, the service times are Erlang with parameters $k = 1$ to 9, and the queue capacity is infinite. The measures listed are P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL. For simplicity, the average service time is $\tau_s = 1.00$, and thereby $Ws = 1.00$ for all situations. The table entries are computed using the Pollaczek–Khintchin equations developed in Chap. 9.

Erlang k	cov	P_0	Lq	Ls	L	Wq	Ws	W	Wq′	SL
1	1.00	0.50	0.50	0.50	1.00	1.00	1.00	2.00	2.00	0.50
2	0.71	0.50	0.38	0.50	0.88	0.75	1.00	1.75	1.50	0.50
3	0.58	0.50	0.33	0.50	0.83	0.67	1.00	1.67	1.33	0.50
4	0.50	0.50	0.31	0.50	0.81	0.63	1.00	1.63	1.25	0.50
5	0.45	0.50	0.30	0.50	0.80	0.60	1.00	1.60	1.20	0.50
6	0.41	0.50	0.29	0.50	0.79	0.58	1.00	1.58	1.17	0.50
7	0.38	0.50	0.29	0.50	0.79	0.57	1.00	1.57	1.14	0.50
8	0.35	0.50	0.28	0.50	0.78	0.56	1.00	1.56	1.13	0.50
9	0.33	0.50	0.28	0.50	0.78	0.56	1.00	1.56	1.11	0.50

Note, the statistical measures: P_0, Ls, Ws and SL are the same for all examples. Further, Lq, L, Wq, W and Wq′ become smaller as the cov decreases from 1.00 to 0.33.

Chapter 20
Erlang Arrivals, Exponential Service (E2/M/1)

Abstract This chapter considers a one-server sytem with finite capacity, Erlang 2-stage inter-arrival times and exponential service times. An example is the trucks that arrive to a receiving dock with one unloading crew. As the trucks come in, the crew (the service facility) unloads each truck in the order of arrival. Matrix methods are used to compute the probability of n units in the system. The probabilities are used to derive the performance measures. The chapter shows how the matrix method can extend to an infinite capacity system. Examples are presented.

20.1 Introduction

Consider a system with one server and where the inter-arrival times are from a 2-stage Erlang probability distribution, and the service times are exponential. Further, the average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This could be a call for a windshield mold in a glass company that builds windshields by model and year for the wide array of automobiles. The windshield mold has the role of the service facility and the customer orders for a particular windshield are the arrivals. New windshields are needed to replace those that are damaged on automobiles.

Recall from Chap. 2, if $x = (y_1 + \ldots + y_k)$ and y is exponential with mean $1/\theta$, then x is Erlang with mean and the variance $E(x) = k/\theta$ and $V(x) = k/\theta^2$, respectively. But in this situation, the inter-arrival time is $t_a = (t_1 + t_2)$ where t is exponential with a mean of $1/(2\lambda)$, and thereby t_a is Erlang with $k = 2$ and has a mean of $E(t_a) = 1/\lambda$.

To obtain the probabilities and summary statistics for this system, it is necessary to artificially set the capacity to a finite size of N. The method of this chapter will show how N can be set in a way where the results will be almost the same as when the capacity is infinite.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_20,

The following notation applies here:

$\tau_a = 1/\lambda$ = average time between arrivals
λ = average number of arrivals per unit of time
k = 2 = Erlang parameter
$\tau_a = \tau_1 + \tau_2 = 1/(2\lambda) + 1/(2\lambda) = 1/\lambda$ = average inter-arrival time to the system
$\mu = 1/\tau_s$
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
$\rho < 1$ is needed to maintain equilibrium
N = an artificial limit on the maximum units allowed in the system
(n,i) = states of the system
n = number of units in the system n = (0,N)
i = 1, 2 at $n \geq 0$ to identify the stage of the next arrival

For this system, the states have two arguments, (n,i) where n = the number of units in the system, and i identifies the stage of the next arrival from 1 to 2. Recall from Chap. 2 where each stage is exponential and the two exponential stages added together give the Erlang shape. The reader should be aware that when the inter-arrival times are Erlang with three or more stages, the method of solution is merely an extension of the procedure shown here.

Below is a list of the difference equations and the corresponding equilibrium equations. The equilibrium equations are needed to develop the probability and statistical measures for the system.

20.2 Difference Equations

(n,i)		
(0,1)		$P_{01}(t + h) = (1 - 2\lambda h)P_{01}(t) + \mu h P_{11}(t) + o(h)$
(0,2)		$P_{02}(t + h) = (1 - 2\lambda h)P_{02}(t) + 2\lambda h P_{01}(t) + \mu h P_{12}(t) + o(h)$
(n,1)	n = (1,N − 1)	$P_{n1}(t + h) = (1 - 2\lambda h - \mu h)P_{n1}(t) + 2\lambda h P_{n-1,2}(t) + \mu h P_{n+1,1}(t) + o(h)$
(n,2)	n = (1,N − 1)	$P_{n2}(t + h) = (1 - 2\lambda h - \mu h)P_{n2}(t) + 2\lambda h P_{n1}(t) + \mu h P_{n+1,2}(t) + o(h)$
(N,1)		$P_{N1}(t + h) = (1 - 2\lambda h - \mu h)P_{N1}(t) + 2\lambda h P_{N-1,2}(t) + 2\lambda h P_{N2}(t) + o(h)$
(N,2)		$P_{N2}(t + h) = (1 - 2\lambda h - \mu h)P_{N2}(t) + 2\lambda h P_{N1}(t) + o(h)$

20.3 Equilibrium Equations

(n,i)		
(0,1)		$0 = -2\lambda P_{01} + \mu P_{11}$
(0,2)		$0 = -2\lambda P_{02} + 2\lambda P_{01} + \mu P_{12}$
(n,1)	n = (1,N − 1)	$0 = -(2\lambda + \mu)P_{n1} + 2\lambda P_{n-1,2} + \mu P_{n+1,1}$
(n,2)	n = (1,N − 1)	$0 = -(2\lambda + \mu)P_{n2} + 2\lambda P_{n1} + \mu P_{n+1,2}$
(N,1)		$0 = -(2\lambda + \mu)P_{N1} + 2\lambda P_{N-1,2} + 2\lambda P_{N2}$
(N,2)		$0 = -(2\lambda + \mu)P_{N2} + 2\lambda P_{N1}$

20.4 Matrix Solution

To solve for the probabilities, P_{ni}, requires matrix methods as will be shown below. To illustrate, a small example where $N = 2$ is used for simplicity. When $N = 2$, the unknown probabilities are: P_{01}, P_{02}, P_{11}, P_{12}, P_{21}, P_{22}, and the equilibrium equations become:

(0,1)	$0 = -2\lambda P_{01} + \mu P_{11}$
(0,2)	$0 = -2\lambda P_{02} + 2\lambda P_{01} + \mu P_{12}$
(1,1)	$0 = -(2\lambda + \mu)P_{11} + 2\lambda_{02} + \mu P_{21}$
(1,2)	$0 = -(2\lambda + \mu)P_{12} + 2\lambda P_{11} + \mu P_{22}$
(2,1)	$0 = -(2\lambda + \mu)P_{21} + 2\lambda P_{12} + 2\lambda P_{22}$
(2,2)	$0 = -(2\lambda + \mu)P_{22} + 2\lambda P_{21}$

20.5 Zero = Zero

When dealing with the more complex set of equilibrium equations, like above, it is good practice to sum all the equations to ensure that $0 = 0$. Since the sum of the left-hand-side of the five equations sum to zero, the sum for the right-hand-side should also be zero. This is needed to assure that all elements are correctly installed in the equations. This check is helpful to locate any elements that should not be included and/or identify any that are missing. The analyst needs to do this by checking that all the elements with a negative value have a corresponding element with a positive value.

20.6 AP = BP_{01}

For an $N = 2$ system, that has six equations, only the first five equations are needed in the matrix formulation of $AP = BP_{01}$.

A is a 5×5 matrix, P and B are 5×1 vectors. P_{01} is a single 1×1 term. For this example, the elements of the matrices are listed below:

20.7 A, P and B

The matrix A is:

$$\begin{bmatrix} 0 & \mu & 0 & 0 & 0 \\ -2\lambda & 0 & \mu & 0 & 0 \\ 2\lambda & -(2\lambda+\mu) & 0 & \mu & 0 \\ 0 & 2\lambda & -(2\lambda+\mu) & 0 & \mu \\ 0 & 0 & 2\lambda & -(2\lambda+\mu) & 2\lambda \end{bmatrix}$$

The vector P is: $\begin{bmatrix} P_{02} \\ P_{11} \\ P_{12} \\ P_{21} \\ P_{22} \end{bmatrix}$

and the vector B is: $\begin{bmatrix} 2\lambda \\ -2\lambda \\ 0 \\ 0 \\ 0 \end{bmatrix}$

20.8 Solving for the Probabilities

Solving for the vector P requires the inverse of A, as below:

$$P = A^{-1}B\ P_{01}$$
$$= Q\ P_{01}$$

The vector Q has elements: q_{02}, q_{11}, q_{12}, q_{21} and q_{22}. Further, another q element (q_{01}) can be assigned to the probability, P_{01} since, $P_{01} = q_{01}P_{01}$. Hence, $q_{01} = 1$. Since the sum of all the probabilities equal unity, the sum of the q elements are used as below to compute P_{01}.

$$P_{01} = 1/(q_{01} + q_{02} + q_{11} + q_{12} + q_{21} + q_{22})$$

So now, all of the probabilities can be computed as follows:

$$P_{02} = q_{02}P_{01}$$
$$P_{11} = q_{11}P_{01}$$
$$P_{12} = q_{12}P_{01}$$
$$P_{21} = q_{21}P_{01}$$
$$P_{22} = q_{22}P_{01}$$

The probabilities P_n are now obtained as below:

$$P_0 = P_{01} + P_{02}$$
$$P_1 = P_{11} + P_{12}$$
$$P_2 = P_{21} + P_{22}$$

Note where the sum, $P_0 + P_1 + P_2 = 1$.

20.9 When n = (0,N)

In the general case, when N is any positive integer, the method to solve for the probabilities follows the same pattern. First, the equilibrium equations (less the final one) are placed in the matrix form $AP = BP_{01}$, and the inverse of A yields the

vector Q. The sum of the Q elements (plus q_{01}) yields the value of P_{01}, as shown earlier, from where all the probabilities P_{ni} are derived. Finally, the probabilities of P_n become:

$P_n = P_{n1} + P_{n2}$ n = (0,N)

20.10 Lambda and Rho Effective

Recall from Chap. 4 where lambda effective and rho effective are computed. In this situation, they are as below:

$\lambda_e = \lambda[1 - P_{N2}]$ = "lambda effective"
$\rho_e = \lambda_e/\mu$ = "rho effective"

In this context,

λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ = utilization ratio,
ρ_e = effective utilization ratio, and is less than one.

20.11 Probability and Statistics for an Infinite Capacity System

As λ_e/λ approaches one, the system statistics and probabilities that are calculated for this finite capacity system are very close to a system with infinite capacity. At the outset, the analyst can set N = 20, say, and then measure the above ratio. If λ_e/λ is too low, then N is increased by increments of 5, say, until the accuracy desired is met. In this way, this system, that artificially sets a limit on the capacity, can be used to find the results needed for a system with infinite capacity. The reader should recognize that as ρ increases, the larger N becomes.

20.12 Expected Number of Units in the Service Facility (Ls)

The expected number of units in the service facility becomes:

$$Ls = \sum_{n=1}^{N} P_n$$

20.13 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained as below,

$$\mathrm{Lq} = \sum_{n=1}^{N} (n-1)P_n$$

20.14 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$$\mathrm{L} = \mathrm{Ls} + \mathrm{Lq}$$

20.15 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$$\mathrm{Ws} = \mathrm{Ls}/\lambda_e$$
$$\mathrm{Wq} = \mathrm{Lq}/\lambda_e$$
$$\mathrm{W} = \mathrm{L}/\lambda_e = \mathrm{Ws} + \mathrm{Wq}$$

20.16 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. The probabilities for these events are:

$$\mathrm{P(D')} = \mathrm{P_o}$$
$$\mathrm{P(D)} = (1 - \mathrm{P_o})$$

The corresponding conditional waiting times in the queue are:

$\mathrm{W_{q|D'}}$ = wait time in queue given no delay
$\mathrm{W_{q|D}}$ = wait time in queue given a delay

The relation between the waiting time (Wq) and the conditional waiting times ($\mathrm{W_{q|D'}}$, $\mathrm{W_{q|D}}$) is below:

$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$

Since $W_{q|D'} = 0$,

$Wq' = W_{q|D} = Wq/P(D) = Wq/(1 - P_0)$

20.17 Service Level and Loss Probability

The service level (SL) is the probability an arrival to the system is not delayed in the queue, and this is simply P_0. The loss probability (Ploss) is the probability a new arrival is lost because the system capacity is too small. This is merely P_{N2}, the probability the system is full, where any new arrival is blocked from entering. Hence,

$SL = P_0$
$Ploss = P_{N2}$

Note, Ploss is another measure on how high to set N so that the probability and statistics results are close enough to an infinite capacity system. When Ploss is less than 0.005, say, then N is deemed adequate.

Example 20.1
Assume a system where arrivals are Erlang with $k = 2$ stages and with an average arrival time of 1.0, and the service is exponential with an average time of 0.5. Hence, $\lambda = 1.0$ and $\mu = 2.0$. To illustrate, using $N = 2$ as the system capacity, the matrices of interest are shown below.

Input:
One-server
Finite capacity with $N = 2$
Inter-arrival times are Erlang with $k = 2$
Service times are exponential
The arrival rate is $\lambda = 1.0$
The service rate is $\mu = 2.0$

Computations:
The matrices A and B are formulated.
Then, the vector $Q = A^{-1}B$ is generated.
Finally, all the probabilities P_{ni} are computed.

The computations for the example are showed below.

$$A = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 \\ -2 & 0 & 2 & 0 & 0 \\ 2 & -4 & 0 & 2 & 0 \\ 0 & 2 & -4 & 0 & 2 \\ 0 & 0 & 2 & -4 & 2 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 \\ -2 \\ 0 \\ 0 \\ \\ 0 \end{bmatrix}$$

Now, taking the inverse of A and solving for Q yields,

$$Q = \begin{bmatrix} 1.6 \\ 1.0 \\ 0.6 \\ 0.4 \\ \\ 0.2 \end{bmatrix}$$

From here, the probability P_{00} is computed as below:

$$\begin{aligned} P_{01} &= 1/[1.0 + 1.6 + 1.0 + 0.6 + 0.4 + 0.2] \\ &= 0.2083 \end{aligned}$$

Hence,

$$\begin{aligned} P_{02} &= 0.3333 \\ P_{11} &= 0.2083 \\ P_{12} &= 0.1250 \\ P_{21} &= 0.0833 \\ P_{22} &= 0.0416 \end{aligned}$$

Finally, the state probabilities are the following:

$$\begin{aligned} P_0 &= 0.5416 \\ P_1 &= 0.3333 \\ P_2 &= 0.1249 \end{aligned}$$

Note, in this system with $N = 2$, the probability of a lost arrival is $Ploss = P_{N2} = 0.0416$. Since this is high, an approximation to an infinite capacity system would require a larger parameter N. Perhaps, $N = 10$ to start with.

Chapter 21
Erlang Arrivals, Erlang Service (E2/E2/1)

Abstract This chapter pertains to a one-server system with finite capacity, and with Erlang 2-stage inter-arrival and service times. Could be a furniture store where, on each sale, the store has a stockman who fetches the item in the back storage area of the store, brings it to the customer's vehicle and helps to load the item in the vehicle. In this situation, the stockman is the service facility. Matrix methods are used to compute the probability of n units in the system. The probabilities are used to calculate the performance measures. The chapter shows how to extend the matrix method for an infinite capacity system.

21.1 Introduction

Consider a system with one server and where the inter-arrival times and the service times are from 2-stage Erlang probability distributions. Also, the average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. An example could be calls arriving to a stock exchange to buy and/or sell stock and the exchange has but one service facility to receive the calls.

Recall from Chap. 2, if $x = (y_1 + \ldots + y_k)$ and y is exponential with mean $1/\theta$, then x is Erlang with mean and the variance $E(x) = k/\theta$ and $V(x) = k/\theta^2$, respectively. In this situation, the inter-arrival time is $t_a = (t_1 + t_2)$ where the right-hand-variables are exponential with a mean of $1/(2\lambda)$, and so, t_a is a 2-stage Erlang variable with mean $1/\lambda$. The service times are also Erlang where $t_s = t_1 + t_2$. and the t variables on the right-hand-side are exponential with expected times of $E(t) = 1/(2\mu)$. Thereby the service times have a 2-stage Erlang distribution with expected time of $E(t_s) = 1/\mu$.

To obtain the probabilities and summary statistics for this system, it is necessary to artificially set the capacity to a finite size of N. The method of this chapter will show how N can be set in a way where the results will be almost the same as when the capacity is infinite.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_21, © Springer Science+Business Media New York 2012

The following notation applies here:

$k = 2$ = Erlang parameter for the arrival times
$k = 2$ = Erlang parameter for the service times
$\tau_a = \tau_1 + \tau_2 = 1/(2\lambda) + 1/(2\lambda) = 1/\lambda$ = average inter-arrival time to the system
$\tau_s = \tau_1 + \tau_2 = 1/(2\mu) + 1/(2\mu) = 1/\mu$ = average service time
$\lambda = 1/\tau_a$ = average number of arrivals per unit of time
$\mu = 1/\tau_s$
$\rho = \tau_s/\tau_a = \lambda/\mu$ = utilization ratio
$\rho < 1$ to ensure equilibrium
N = an artificial limit on the maximum units allowed in the system
(n,i,j) = states of the system
n = number of units in the system n = (0,N)
i = 1, 2 at $n \geq 0$ to identify the stage of next arrival
j = 0, 1, 2. At n = 0, j = 0, and at $n \geq 1$, j = 1, 2 to identify the stage of the current unit in service

For this system, the states have three arguments, (n,i,j) where n = the number of units in the system; i identifies the stage (1 or 2) of the next arrival; and j gives the stage (1 or 2) of the service time for the current unit in the service facility, and j = 0 is when no unit is in the service facility. Recall from Chap. 2, two exponential stages added together give the Erlang shape with k = 2. The reader should be aware that when the inter-arrival and the service times are Erlang with three or more stages, the method of solution is merely an extension of the procedure described here.

Below is a list of the difference equations and the corresponding equilibrium equations. The equilibrium equations are needed to develop the probability and statistical measures for the system.

21.2 Difference Equations

(n,I,j)	
(0,1,0)	$P_{010}(t + h) = (1 - 2\lambda h)P_{010}(t) + 2\mu h P_{112}(t) + o(h)$
(0,2,0)	$P_{020}(t + h) = (1 - 2\lambda h)P_{020}(t) + 2\lambda h P_{010}(t) + 2\mu h P_{122}(t) + o(h)$
(1,1,1)	$P_{111}(t + h) = (1 - 2\lambda h - 2\mu h)P_{111}(t) + 2\lambda h P_{020}(t) + 2\mu h P_{212}(t) + o(h)$
(1,1,2)	$P_{n12}(t + h) = (1 - 2\lambda h - 2\mu h)P_{112}(t) + 2\mu h P_{111}(t) + o(h)$
(1,2,1)	$P_{121}(t + h) = (1 - 2\lambda h - 2\mu h)P_{121}(t) + 2\lambda h P_{111}(t) + 2\mu h P_{222}(t) + o(h)$
(1,2,2)	$P_{122}(t + h) = (1 - 2\lambda h - 2\mu h)P_{122}(t) + 2\lambda h P_{112}(t) + 2\mu h P_{121}(t) + o(h)$

$$
\begin{array}{lll}
(n,1,1) & n = (2,N-1) & P_{n11}(t+h) = (1-2\lambda h-2\mu h)P_{n11}(t) + 2\lambda h P_{n-1,21}(t) \\
 & & \quad + 2\mu h P_{n+1,12}(t) + o(h) \\
(n,1,2) & n = (2,N-1) & P_{n12}(t+h) = (1-2\lambda h-2\mu h)P_{n12}(t) + 2\lambda h P_{n-1,22}(t) \\
 & & \quad + 2\mu h P_{n11}\,(t) + o(h) \\
(n,2,1) & n = (2,N-1) & P_{n21}(t+h) = (1-2\lambda h-2\mu h)P_{n21}(t) + 2\lambda h P_{n11}(t) \\
 & & \quad + 2\mu h P_{n+1,22}(t) + o(h) \\
(n,2,2) & n = (2,N-1) & P_{n22}(t+h) = (1-2\lambda h-2\mu h)P_{n22}(t) + 2\lambda h P_{n12}(t) \\
 & & \quad + 2\mu h P_{n21}(t) + o(h) \\
(N,1,1) & & P_{N11}(t+h) = (1-2\lambda h-2\mu h)P_{N11}(t) \\
 & & \quad + 2\lambda h P_{N-1,21}(t) + 2\lambda h P_{N21}(t) + o(h) \\
(N,1,2) & & P_{N12}(t+h) = (1-2\lambda h-2\mu h)P_{N12}(t) \\
 & & \quad + 2\lambda h P_{N-1,22}(t) + 2\lambda h P_{N22}(t) \\
 & & \quad + 2\mu h P_{N11}(t) + o(h) \\
(N,2,1) & & P_{N21}(t+h) = (1-2\lambda h-2\mu h)P_{N21}(t) + 2\lambda h P_{N11}(t) \\
 & & \quad + o(h) \\
(N,2,2) & & P_{N22}(t+h) = (1-2\lambda h-2\mu h)P_{N22}(t) + 2\lambda h P_{N12}(t) \\
 & & \quad + 2\mu h P_{N21}(t) + o(h)
\end{array}
$$

21.3 Equilibrium Equations

$$
\begin{array}{lll}
(n,I,j) & & \\
(0,1,0) & & 0 = -2\lambda P_{010} + 2\mu P_{112} \\
(0,2,0) & & 0 = -2\lambda P_{020} + 2\lambda P_{010} + 2\mu P_{122} \\
(1,1,1) & & 0 = -(2\lambda+2\mu)P_{111} + 2\lambda P_{020} + 2\mu P_{212} \\
(1,1,2) & & 0 = -(2\lambda+2\mu)P_{112} + 2\mu P_{111} \\
(1,2,1) & & 0 = -(2\lambda+2\mu)P_{121} + 2\lambda P_{111} + 2\mu P_{222} \\
(1,2,2) & & 0 = -(2\lambda+2\mu)P_{122} + 2\lambda P_{112} + 2\mu P_{121} \\
(n,1,1) & n = (2,N-1) & 0 = -(2\lambda+2\mu)P_{n11} + 2\lambda P_{n-1,21} + 2\mu P_{n+1,12} \\
(n,1,2) & n = (2,N-1) & 0 = -(2\lambda+2\mu)P_{n12} + 2\lambda P_{n-1,22} + 2\mu P_{n11} \\
(n,2,1) & n = (2,N-1) & 0 = -(2\lambda+2\mu)P_{n21} + 2\lambda P_{n11} + 2\mu P_{n+1,22} \\
(n,2,2) & n = (2,N-1) & 0 = -(2\lambda+2\mu)P_{n22} + 2\lambda P_{n12} + 2\mu P_{n21} \\
(N,1,1) & & 0 = -(2\lambda+2\mu)P_{N11} + 2\lambda P_{N-1,21} + 2\lambda P_{N21} \\
(N,1,2) & & 0 = -(2\lambda+2\mu)P_{N12} + 2\lambda P_{N-1,22} + 2\lambda P_{N22} + 2\mu P_{N11} \\
(N,2,1) & & 0 = -(2\lambda+2\mu)P_{N21} + 2\lambda P_{N11} \\
(N,2,2) & & 0 = -(2\lambda+2\mu)P_{N22} + 2\lambda P_{N12} + 2\mu P_{N21}
\end{array}
$$

21.4 Matrix Solution

To solve for the probabilities, P_{nij} requires matrix methods as is shown below. To illustrate, a small example where $N = 2$ is used for simplicity. When $N = 2$, the unknown probabilities are: P_{010}, P_{020}, P_{111}, P_{112}, P_{121}, P_{122}, P_{211}, P_{212}, P_{221}, P_{222}. The equilibrium equations are listed below:

$$
\begin{aligned}
&(0,1,0) && 0 = -2\lambda P_{010} + 2\mu P_{112}\\
&(0,2,0) && 0 = -2\lambda P_{020} + 2\lambda P_{010} + 2\mu P_{122}\\
&(1,1,1) && 0 = -(2\lambda + 2\mu)P_{111} + 2\lambda P_{020} + 2\mu P_{212}\\
&(1,1,2) && 0 = -(2\lambda + 2\mu)P_{112} + 2\mu P_{111}\\
&(1,2,1) && 0 = -(2\lambda + 2\mu)P_{121} + 2\lambda P_{111} + 2\mu P_{222}\\
&(1,2,2) && 0 = -(2\lambda + 2\mu)P_{122} + 2\lambda P_{112} + 2\mu P_{121}\\
&(2,1,1) && 0 = -(2\lambda + 2\mu)P_{211} + 2\lambda P_{121} + 2\lambda P_{221}\\
&(2,1,2) && 0 = -(2\lambda + 2\mu)P_{212} + 2\lambda P_{122} + 2\lambda P_{222} + 2\mu P_{211}\\
&(2,2,1) && 0 = -(2\lambda + 2\mu)P_{221} + 2\lambda P_{211}\\
&(2,2,2) && 0 = -(2\lambda + 2\mu)P_{222} + 2\lambda P_{212} + 2\mu P_{221}
\end{aligned}
$$

21.5 Zero = Zero

When dealing with the more complex set of equilibrium equations, like above, it is good practice to sum all the equations to ensure that $0 = 0$. Since the sum of the left-hand-side of the ten equations sum to zero, the sum for the right-hand-side should also be zero. This is needed to assure that all elements are correctly installed in the equations. This check is helpful to locate any elements that should not be included and/or identify any that are missing. The analyst needs to do this by checking that all the elements with a negative value have a corresponding element with a positive value.

21.6 AP = BP_{010}

For an $N = 2$ system, that has ten equations, the first nine equations are needed in the matrix form that are shown below.

$AP = BP_{010}$
A is a 9×9 matrix, P and B are 9×1 vectors. P_{010} is a single 1×1 term.

21.7 Solving for the Probabilities

Solving for the vector P requires the inverse of A, as below:

$$P = A^{-1}B\ P_{010}$$
$$= Q\ P_{010}$$

The vector Q has elements: q_{020}, q_{111}, q_{112}, q_{121}, q_{122}, q_{211}, q_{212}, q_{221}, q_{222}. Further, another q element (q_{010}) can be assigned to the probability, P_{010} since, $P_{010} = q_{010}P_{010}$. Hence,

$q_{010} = 1$. Since the sum of all the probabilities equal unity, the sum of the q elements are used as below to compute P_{010}:

$P_{010} = 1/(q_{010} + q_{020} + q_{111} + q_{112} + q_{121} + q_{122} + q_{211} + q_{212} + q_{221} + q_{222})$

So now, all of the probabilities can be computed as follows:

$P_{020} = q_{020}P_{010}$
$P_{111} = q_{111}P_{010}$
$P_{112} = q_{112}P_{010}$
$P_{121} = q_{121}P_{010}$
$P_{122} = q_{122}P_{010}$
$P_{211} = q_{211}P_{010}$
$P_{212} = q_{212}P_{010}$
$P_{221} = q_{221}P_{010}$
$P_{222} = q_{222}P_{010}$

The probabilities P_n are obtained as below:

$P_0 = P_{010} + P_{020}$
$P_1 = P_{111} + P_{112} + P_{121} + P_{122}$
$P_2 = P_{211} + P_{212} + P_{221} + P_{222}$

Note where the sum, $P_0 + P_1 + P_2 = 1$.

21.8 When n = (0,N)

In the general case, when N is any positive integer, the method to solve for the probabilities follows the same pattern. First, the equilibrium equations (less the final one) are placed in the matrix form $AP = BP_{010}$, and the inverse of A yields the vector Q. The sum of the Q elements (plus q_{010}) yields the value of P_{010}, as shown earlier, from where all the probabilities P_{nij} are derived. Finally, the probabilities of P_n become:

$P_0 = P_{010} + P_{020}$ $\quad n = 0$
$P_n = P_{n11} + P_{n12} + P_{n21}\ P_{n22}$ $\quad n = (1,N)$

21.9 Lambda and Rho Effective

In this system, lambda effective and rho effective are computed as below:

$\lambda_e = \lambda[1 - P_{N21} - P_{N22}]$ = "lambda effective"
$\rho_e = \lambda_e/\mu$ = "rho effective"

In this context,

λ = expected number of arrivals in a unit of time,
λ_e = expected number of units that enter the system in a unit of time,
$\lambda - \lambda_e$ = expected number of units that are lost per unit of time,
ρ_e = effective utilization ratio.

21.10 Probability and Statistics for an Infinite Capacity System

As λ_e/λ approaches one, the system statistics and probabilities that are calculated for this finite capacity system are very close to a system with infinite capacity. At the outset, the analyst can set N = 20, say, and then measure the above ratio. If λ_e/λ is too low, then N is increased by increments of 5, say, until the accuracy desired is met. In this way, this system that artificially sets a limit on the capacity can be used to find the results needed for a system with infinite capacity.

21.11 Expected Number of Units in the Service Facility (Ls)

The expected number of units in the service facility becomes:

$$\text{Ls} = \sum_{n=1}^{N} P_n$$

21.12 Expected Units in the Queue (Lq)

The expected number of units in the queue is obtained as below,

$$\text{Lq} = \sum_{n=1}^{N} (n-1)P_n$$

21.13 Expected Units in the System (L)

The expected number of units in the system (service facility plus queue) is

$\text{L} = \text{Ls} + \text{Lq}$

21.14 Expected Time in Service (Ws), Queue (Wq) and System (W)

Using Little's Law,

$Ws = Ls/\lambda_e$
$Wq = Lq/\lambda_e$
$W = L/\lambda_e = Ws + Wq$

21.15 Expected Time in the Queue Given a Delay (Wq′)

Another useful system statistic is the expected time in the queue for an arrival that is delayed in the queue. Note that an arrival that is not delayed will not have to wait in the queue. Wq is the average of both of these events. So it is helpful to introduce the events D and D′, where D = the event a new arrival is delayed, and D′ = the event of not delayed. The probabilities for these events are:

$P(D') = P_o$
$P(D) = (1 - P_o)$

The corresponding conditional waiting times in the queue are:

$W_{q|D'}$ = wait time in queue given no delay
$W_{q|D}$ = wait time in queue given a delay

The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:

$$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$$

Since $W_{q|D'} = 0$,

$$Wq' = W_{q|D} = Wq/P(D) = Wq/(1 - P_0)$$

21.16 Service Level and Loss Probability

The service level (SL) is the probability an arrival to the system is not delayed in the queue, and this is simply P_0. The loss probability (Ploss) is the probability a new arrival is lost because the system capacity is too small. This is merely ($P_{N21} + P_{N22}$), the probability the system is full, where any new arrival is blocked from entering. Hence,

$SL = P_0$
$Ploss = P_{N21} + P_{N22}$

Note, Ploss is another measure on how high to set N so that the probability and statistics results are close enough to an infinite capacity system. When Ploss is less than 0.005, say, then N would be deemed adequate.

Chapter 22
Waiting Time Density, One Server (M/M/1)

Abstract This chapter shows how to calculate the waiting time probabilities for a one-server system, with infinite capacity, exponential inter-arrival times and exponential service times, where the customers are serviced in a first-in-first-out discipline. An example is when a city designs a beat for a squad car and wants to determine the probability that at least 90 percent of the calls received for the beat can begin service within 10 min. The squad car is the service facility and the calls within the beat are the arrivals. Examples are presented.

22.1 Introduction

Consider a system with one server and an infinite queue where the inter-arrival and the service times have exponential probability densities. The average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This is the (M/M/1) system described in Chap. 3. The reader should know that the expected wait time values developed in Chap. 3 are averages for any service discipline. This chapter shows how to measure the waiting time distribution when the first-in-first-out (FIFO) service discipline is in use. The average times (Wq, Wq′) given earlier are valid for any service discipline, but the probability density could vary by service discipline. An example on the use of the probability density is when a small hospital has one ambulance and wants to determine the probability that an emergency call will have to wait more than 20 min to receive service. Some of the notation and results for this system are listed below.

$\tau_a = 1/\lambda =$ average time between arrivals
$\tau_s = 1/\mu =$ average time to service a unit
$\lambda =$ average number of arrivals per unit of time
$\mu =$ average number of units processed in a unit of time for a continuously busy service facility

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_22, © Springer Science+Business Media New York 2012

$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio.
$\rho < 1$ is needed to assure the system is in equilibrium
$\lambda < \mu$
n = number of units in the system $(n \geq 0)$

$P_n = (1 - \rho)\rho^n \qquad n \geq 0$
$P_0 = (1 - \rho) \qquad n = 0$

22.2 Conditional Probability of Wait Time in Queue

Suppose n units are in the system when a new arrival enters the system. Of interest is to find the probability that the wait time in the queue, t, for the new arrival lies between (t′ and t′ + dt). This is a conditional probability that depends on the size of n, and is defined as below:

$$\begin{aligned} P(t' < t < t' + dt|\ n) &= P[(n-1 \text{ departures in } t') \,\&\, (1 \text{ departure in dt})|n] \\ &= P[(n-1 \text{ departures in } t')] \times P[(1 \text{ departure in dt})] \\ &= [e^{-\mu t'}(\mu t')^{n-1}]/(n-1)! \times \mu dt \end{aligned}$$

22.3 Probability of Wait Time in Queue

Thereby,

$$\begin{aligned} P(t' < t < t' + dt) &= \sum_{n=1}^{\infty} P(t' < t < t' + dt \mid n) P_n \\ &= \sum_{n=1}^{\infty} [e^{-\mu t'}(\mu t')^{n-1}]/(n-1)!\,(\mu dt)\rho^n(1-\rho) \\ &= (\mu dt) e^{-\mu t'} \rho(1-\rho) \sum_{n=1}^{\infty} (\mu t' \rho)^{n-1}/(n-1)! \end{aligned}$$

Applying Equation (2.9) to the summation portion of the above, the probability becomes:

$$\begin{aligned} P(t' < t < t' + dt) &= \lambda(1-\rho)e^{(\lambda-\mu)t'}dt \\ &= f(t')dt \end{aligned}$$

Thus,
$f(t) = \lambda(1 - \rho)\, e^{(\lambda-\mu)t}$ at $n \geq 1$

Finally, the probability of the time in the queue becomes a mixed discrete and continuous distribution as listed below:

$P(t = 0) = P_0 = (1 - \rho) \quad (t = 0)$
$f(t) = \lambda(1 - \rho)\, e^{(\lambda-\mu)t} \quad (t > 0)$

To find the probability that the time in the queue is greater than t' becomes, $P(t > t') = \int_{t>t}^{t'} f(t)dt = \rho\, e^{(\lambda-\mu)t'}$

Example 22.1
Suppose a one service facility system with infinite capacity, and with exponential arrival and service times. The average time between arrivals is 10 min, and the average time per service is 8 min. Customers are serviced on a first-come-first served basis. Some of the wait time statistics for this system are listed below.

Input:
One-server
Infinite capacity
Inter-arrival times are exponential
Service times are exponential
τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min
Service discipline is first-come-first-serve

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\rho = \lambda/\mu = 0.80$
$P_n = (.20).80^n \; n \geq 0$
$P_o = 0.2000$
Ws = 8 min = 0.1333 h
Wq = 32 min = 0.5333 h
W = 40 min = 0.6666 h
Wq′ = 40 min = 0.6666 h
t = wait time in queue
$P(t > t') = 0.80e^{(-0.025)t'}$
P(t = 0 min) = 0.200
P(t > 0 min) = 0.800
P(t > 10 min) = 0.623
P(t > 20 min) = 0.485
P(t > 30 min) = 0.378
P(t > 40 min) = 0.294

Example 22.2
The table below pertains to an infinite capacity queuing system with one service facility, exponential arrivals and service times, where the average service time is $1/\mu = 1.00$, and the customers are serviced on a first-in-first-out basis. In the table, t is the waiting time in the queue, and the entries give the probability of t larger than t', $P(t > t')$, for utilization ratios, ρ, ranging from 0.10 to 0.90. Note that $t' = 1$ is the same as the average service time.

			$P(t > t')$					
ρ	$P(t = 0)$	t'	0	1	2	3	4	5
0.10	0.90		0.10	0.04	0.02	0.01	0.00	0.00
0.20	0.80		0.20	0.09	0.04	0.02	0.01	0.00
0.30	0.70		0.30	0.15	0.07	0.04	0.02	0.01
0.40	0.60		0.40	0.22	0.12	0.07	0.04	0.02
0.50	0.50		0.50	0.30	0.18	0.11	0.07	0.04
0.60	0.40		0.60	0.40	0.27	0.18	0.12	0.08
0.70	0.30		0.70	0.52	0.38	0.28	0.21	0.16
0.80	0.20		0.80	0.65	0.54	0.44	0.36	0.29
0.90	0.10		0.90	0.81	0.74	0.67	0.60	0.55

The above table can be used for any average service time. For example, if $\rho = 0.8$ and the average service time is 8 min, as in Example 22.1, the probability $P(t > 8 \text{ min}) = 0.65$, since in the above table, $P(t > 1) = 0.65$. Further, $P(t > 16 \text{ min}) = 0.54$, and so forth.

Chapter 23
Waiting Time Density, Multi Servers (M/M/k)

Abstract This chapter shows how to calculate the waiting time probabilities for a multi-server system, with infinite capacity, exponential inter-arrival times and exponential service times, where the customers are serviced in a first-in-first-out discipline. Could be a package delivery service that wants to determine the number of delivery vehicles to have in its fleet so 90 percent of deliveries begin within 20 min of the call. Examples are presented.

23.1 Introduction

Consider a system with k servers and an infinite queue where the inter-arrival and the service times have exponential probability densities. The average time between arriving customers is $1/\lambda$ and the average service time is $1/\mu$. This is the (M/M/k) system described in Chapter 6. The reader should know that the expected wait time values developed in Chapter 6 are averages for any service discipline. This chapter shows how to measure the waiting time distribution when the first-in-first-out (FIFO) service discipline is in use. An example is when a popular pizza store wants to determine the number of delivery drivers to have available to assure the customers that 95 percent of the deliveries will take less than 40 min. The average times (Wq, Wq′) given earlier are valid for any service discipline, but the probability density could vary by service discipline. Some of the notation and results for this system are listed below.

k = number of service facilities
$\tau_a = 1/\lambda$ = average time between arrivals
$\tau_s = 1/\mu$ = average time to service a unit
λ = average number of arrivals per unit of time
μ = average number of units processed in a unit of time for a continuously busy service facility

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_23,

$\rho = \tau_s/\tau_a = \lambda/\mu =$ utilization ratio
$\rho < k$ is needed to assure the system is in equilibrium
$\lambda < k\mu$
n = number of units in the system

$$P_n \quad \begin{matrix} = \rho^n/n!P_0 & n = (0,k-1) \\ \rho^n/[k!k^{n-k}]P_0 & n \geq k \end{matrix}$$

$$P_0 = 1/\{\sum_{n=0}^{k-1} \rho^n/n! + \rho^k/[(k-1)!(k-\rho)]\}$$

23.2 Conditional Probability of Wait Time in Queue

Suppose n units, where $(n > k)$, are in the system when a new arrival enters the system. Of interest is to find the probability that the wait time in the queue, t, for the new arrival lies between (t' and $t' + dt$). This is a conditional probability that depends on the size of n, and is defined as below:

$$P(t' < t < t' + dt \mid n) = P[(n - k \text{ departures in } t') \ \& \ (1 \text{ departure in } dt) \mid n]$$

$$= P[(n - k \text{ departures in } t')] \times P[(1 \text{ departure in } dt)]$$

$$= [e^{-k\mu t'} (k\mu t')^{n-k}]/(n-k)! \times k\mu dt$$

23.3 Probability of Wait Time in Queue

Thereby,

$$P(t' < t < t' + dt) = \sum_{n \geq k} P(t' < t < t' + dt \mid n)P_n$$

$$= \sum_{n \geq k} [e^{-k\mu t'}(k\mu t')^{n-k}/(n-k)!](k\mu dt)\ P_n$$

$$= P_0(k\mu dt)e^{-k\mu t'}\rho^k/k! \sum_{n \geq k} (k\mu t'\rho)^{n-k}/[(n-k)!k^{n-k}]$$

Applying Eq. (2.9) to the summation portion on the right-hand-side of the above, the probability becomes:

$$P(t' < t < t' + dt) = P_0(k\mu dt)\ [e^{-k\mu t'}\rho^k/k!]e^{\lambda t'}$$

$$= f(t')dt$$

Thus,

$$f(t) = P_0\mu\rho^k e^{(\lambda - k\mu)t}/(k-1)!$$

Finally, the probability of the time in the queue becomes a mixed discrete and continuous distribution as listed below:

$P(t = 0) = P_{n<k}\ (t = 0)$

$f(t) = P_0 \mu \rho^k e^{(\lambda - k\mu)t}/(k - 1)!\ (t > 0)$

To find the probability that the time in the queue t is greater than t′ becomes,

$P(t > t') = \int_{t > t'}^{\infty} f(t)dt = P_0 \rho^k e^{(\lambda - k\mu)t'}/[(k - 1)!(k - \rho)]$

Example 23.1

Suppose a two-service facility system with infinite capacity, and with exponential arrival and service times. The average time between arrivals is 10 min, and the average time per service is 8 min. Customers are serviced on a first-come-first serve basis. Some of the key probabilities and statistics associated with the waiting time for this system are listed below.

Input:
Two-servers
Infinite capacity
Inter-arrival times are exponential
Service times are exponential
τ_a = expected time between arrivals = 10 min
τ_s = expected service time = 8 min
Service discipline is first-come-first-served

Computations:
$\lambda = 1/\tau_a = 0.10$ per minute
$\mu = 1/\tau_s = 0.125$ per minute
$\rho = \lambda/\mu = 0.80$
$P_0 = 0.4286$
$P_n = (0.4286).80^n/n! \quad n = (0,2)$
$P_n = (0.2143).80^n/2^{n-2} \quad n \geq 3$
Ws = 8 min = 0.133 h
Wq = 1.52 min = 0.025 h
W = 9.52 min = 0.159 h
Wq′ = 6.67 min = 0.111 h
$P(t > t') = 0.2286e^{(-0.15)t'}$
P(t = 0 min) = P(n = 0) + P(n = 1) = 0.771
P(t > 0 min) = 0.229
P(t > 10 min) = 0.051
P(t > 20 min) = 0.011

Example 23.2

The table below pertains to queuing systems with k = 2 and 3 service facilities, infinite capacity, exponential arrivals and service times, where the average service time is $1/\mu = 1.00$, and the customers are serviced on a first-in-first-out basis. In

the table, t is the waiting time in the queue, and the entries give the probability of t larger than t', $P(t > t')$, for utilization ratios, ρ, ranging from 0.50 to 1.50 at $k = 2$; and 0.50 to 2.50 at $k = 3$. Note that $t' = 1$ is the same as the average service time.

				$P(t > t')$				
K	ρ	$P(t = 0)$	t'	0	1	2	3	4
2	0.50	0.90		0.10	0.02	0.01	0.00	0.00
2	1.00	0.67		0.33	0.12	0.05	0.02	0.00
2	1.50	0.36		0.64	0.39	0.24	0.14	0.09
3	0.50	0.98		0.02	0.00	0.00	0.00	0.00
3	1.00	0.91		0.09	0.01	0.00	0.00	0.00
3	1.50	0.76		0.24	0.05	0.01	0.00	0.00
3	2.00	0.56		0.44	0.16	0.06	0.02	0.01
3	2.50	0.30		0.70	0.43	0.26	0.16	0.10

The above table can be used for any average service time. For example, if $k = 2$ servers and $\rho = 1.0$ and the average service time is 8 min, $P(t > 8 \text{ min}) = 0.12$ and $P(t > 16 \text{ min}) = 0.05$. If $k = 3$ servers and $\rho = 1.00$, $P(t > 8 \text{ min}) = 0.01$ and $P(t > 16 \text{ min}) = 0.00$, and so forth.

Bibliography

Asmussen, S. (2003). *Applied probability and queues* (2nd ed). New York: Springer.

Baccelli, F., & Bremaud, P. (2003). *Elements of queueing theory: Palm martingale calculus and stochastic recurrences* (2nd ed). New York: Springer.

Bhat, U. N., & Basawa, I. V. (1992). *Queuing and related models*. Oxford: Oxford University Press.

Borovkov, A. A. (1976). *Stochastic processes in queueing theory*. Heidelberg: Springer

Bose, S. (2001). *An introduction to queueing systems*. New York: Kluwer Academic

Bolch, G., Greiner, S. de Meer, H., & Shridharbhai, K. S. (1998). Queueing networks and markov chains. New York: Wiley.

Cooper, R. B. (1981). *Introduction to queueing theory* (2nd ed.) North Holland: Elsevier.

Dshalalow, J. H. (1997). *Frontiers in queueing*. New York: CRC Press.

Giambene, G. (2005). *Queuing theory and telecommunications : Networks and applications*. New York: Springer.

Gross, D., Shortle, J., Thompson, J., & Harris, C. (2008). *Fundamentals of queueing theory* (4th ed.). New York: Wiley (Series in Probability and Statistics).

Kleinrock, L. (1975). *Queueing systems. Volume 2. theory*. New York: Wiley.

Kleinrock, L. (1976). *Queueing systems. Volume 2. applications*. New York: Wiley.

Morse, P. M. (1958). *Queues, inventories and maintenance*.New York: Wiley.

Ng, C. H. (1997). *Queueing modelling fundamentals*. New York: Wiley.

Saaty, T. L. (1961). *Elements of queueing theory with applications*. New York: McGraw-Hill.

Tanner, M. (1995). *Practical queueing analysis/Book and disk*. New York: IBM McGraw-Hill.

White, J. A., Schmidt J. W., & Bennett G. K. (1975). *Analysis of queueing systems*. New York: Academic Press.

DOI: 10.1007/978-1-4614-3713-0,

Problems

Chapter 3

3.1 A queuing system has one service facility, an infinite queue capacity, and has exponential arrival and service times. The average arrival time is 5 h and the average service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

3.2 Assume the same system as Problem 3.1.
Find P_0, P_1, P_2, P_3 and $P_{n \geq 4}$.

3.3 Assume the same system as Problem 3.1. Suppose this is an auto repair shop that is open 8 h a day and 5 days a week. Also, the average fee per car is $500.
Find, the following:
a. The expected fees per week.
b. The expected idle hours per week.

Chapter 4

4.1 A queuing system has one service facility, with a maximum capacity of N = 3 customers. The arrival and service times are exponential. The average arrival time is 5 h and the average service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W and Wq′.

4.2 Assume the same system as Problem 4.1.
Find P_0, P_1, P_2,P_3 SL, Ploss and λ_e.

4.3 Assume the same system as Problem 4.1. Suppose this is an auto repair shop that is open 8 h a day and 5 days a week. Also, the average fee per car is $500.
Find, the following:
a. The expected fees per week.
b. The expected idle hours per week.
c. The expected customers lost in a week.

4.4 Assume a one service facility system with capacity of N = 2 where all arrival and service times are exponential. The arrival rates per hour are $\lambda_0 = 4$, $\lambda_1 = 3$, and $\lambda_2 = 0$, the service rates per hour are $\mu_1 = 1$ and $\mu_2 = 2$.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0, © Springer Science+Business Media New York 2012

Find, P_0, P_1, P_2, Lq, Ls, L and SL.
Note the equilibrium equations listed below:
$0 = -\lambda_0 P_0 + \mu_1 P_1$
$0 = -(\lambda_1 + \mu_1)P_1 + \lambda_0 P_0 + \mu_2 P_2$
$0 = -\mu_2 P_2 + \lambda_1 P_1$

4.5 Consider the system of Problem 4.4 and suppose the facility is open 40 h a week and the fee per customer is $50. Also the cost of labor per hour is $20.
Note, $E(\lambda) = \lambda_0 P_0 + \lambda_1 P_1 + \lambda_2 P_2$.
Find the E(fees/week), E(labor cost/week) and E(idle hours/week)

Chapter 5

5.1 A queuing system has one service facility, with no queue space. The arrival and service times are exponential. The average arrival time is 5 h and the average service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W, Wq′, Sl, Ploss and λ

5.2 Assume the same system as Problem 5.1. Find P_0 and P_1

5.3 Assume the same system as Problem 5.1. Suppose this is an auto repair shop that is open 8 h a day and 5 days a week. Also, the average fee per car is $500.
Find, the following:
a. The expected fees per week.
b. The expected idle hours per week.
c. The expected customers lost in a week.

Chapter 6

6.1 A queuing system has two service facilities, an infinite queue capacity, and has exponential arrival and service times. The average arrival time is 5 h and the average service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

6.2 Assume the same system as Problem 6.1.
Find P_0, P_1, P_2, P_3 and $P_{n \geq 4}$.

6.3 Assume the same system as Problem 6.1. Suppose this is an auto repair shop that is open 8 h a day and 5 days a week. Also, the average fee per car is $500.
Find, the following:
a. The expected fees per week.
b. The expected idle hours per week.

Chapter 7

7.1 A queuing system has two service facilities and has a maximum capacity of N = 3 customers. The arrival and service times are exponential. The average arrival time is 5 h and the average service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W and Wq′.

7.2 Assume the same system as Problem 7.1.
Find P_0, P_1, P_2 , P_3, SL, Ploss and λ_e.

7.3 Assume the same system as Problem 7.1. Suppose this is an auto repair shop that is open 8 h a day and 5 days a week. Also, the average fee per car is $500. Find, the following:
a. The expected fees per week.
b. The expected operator idle hours per week.
c. The expected customers lost in a week.

7.4 Assume a two service facility system with capacity of N = 3 where all arrival and service times are exponential. The arrival rates per hour are $\lambda_0 = 5$, $\lambda_1 = 4$, $\lambda_2 = 3$, and $\lambda_3 = 0$, the service rates per hour are $\mu_1 = 2$, $\mu_2 = 3$ and $\mu_3 = 4$. Find, P_0, P_1, P_2, P_3, Lq, Ls, L and SL.
Note the equilibrium equations listed below:
$0 = -\lambda_0 P_0 + \mu_1 P_1$
$0 = -(\lambda_1 + \mu_1)P_1 + \lambda_0 P_0 + 2\mu_2 P_2$
$0 = -(\lambda_2 + 2\mu_2)P_2 + \lambda_1 P_1 + 2\mu_3 P_3$
$0 = -2\mu_3 P_3 + \lambda_2 P_2$

7.5 A parking lot has 2 spaces where the customers are cars (need one space each) and trucks (need two spaces each) and all arrival and service times are exponential. The arrival rates for cars are 4 per hour, and for trucks, it is 2 per hour. The average parking rate is 6 per hour for cars and trucks. Find P_{n1n2} where n_1 is the number of cars in the system, and n_2 is the number of trucks in the system.
Note the equilibrium equations listed below:
$0 = -(\lambda_1 + \lambda_2)P_{00} + \mu_1 P_{10} + \mu_2 P_{01}$
$0 = -(\lambda_1 + \mu_1)P_{10} + \lambda_1 P_{00} + 2\,\mu_1 P_{20}$
$0 = -(2\mu_1)P_{20} + \lambda_1 P_{10}$
$0 = -\mu_2 P_{01} + \lambda_2 P_{00}$

Chapter 8

8.1 A queuing system has two service facilities and no queue space. The arrival and service times are exponential. The average arrival time is 5 h and the average service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W and Wq′.

8.2 Assume the same system as Problem 8.1.
Find P_0, P_1, P_2, SL, Ploss and λ_e.

8.3 Assume the same system as Problem 8.1. Suppose this is an auto repair shop that is open 8 h a day and 5 days a week. Also, the average fee per car is $500. Find, the following:
a. The expected fees per week.
b. The expected operator idle hours per week.
c. The expected customers lost in a week.

8.4 Assume a two service facility system with capacity of N = 2 where all arrival and service times are exponential. The arrival rates per hour are $\lambda_0 = 4$, $\lambda_1 = 2$, and $\lambda_2 = 0$, the service rates per hour are $\mu_1 = 8$ and $\mu_2 = 4$.

Find, P_0, P_1, P_2, Ls and SL.
Note the equilibrium equations listed below:
$0 = -\lambda_0 P_0 + \mu_1 P_1$
$0 = -(\lambda_1 + \mu_1)P_1 + \lambda_0 P_0 + 2\mu_2 P_2$
$0 = -2\mu_2 P_2 + \lambda_1 P_1$

Chapter 9

9.1 A one server system has exponential arrivals and an infinite queue capacity. The average arrival times are 5 min and the service times are normally distributed with an average of 3 min and a standard deviation of 1 min.
Find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

9.2 A one server system has exponential arrivals with an average time of 4 min and the queue capacity is infinite. The service times are uniform ranging from 2 to 4 min.
Find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

9.3 A one server system has exponential arrivals with an average time of 4 min and the queue capacity is infinite. The service times, t_s, are discrete distributed with: $P(t_s = 2) = 0.2$ and $P(t_s = 4) = 0.8$.
Find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

Chapter 10

10.1 A one server system has one service facility with an infinite queue capacity and exponential arrivals from two populations. The average arrival times are 8 min from population 1, and 2 min from population 2. The service times have a mean of two minutes and a standard deviation of zero minutes from population 1, and a mean of one minute and a standard deviation of zero minutes from population 2.
Find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

10.2 For the system in 10.1, find the following statistics for population 1 and 2.
For populations 1, find: Lq, Ls, L, Wq, Ws, W, Wq′ and SL.
For populations 2, find: Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

Chapter 11

11.1 Suppose a one repairman shop with six machines and the average run time per machine is 5 h and the average service time is one h. All is exponential.
Find P_0, P_1, P_2, P_3, P_4, P_5, P_6.

11.2 For the shop in 11.1, find the following: Ls, Lq, L, Ws and SL.

11.3 Suppose the shop in 11.1, is open 8 h a day and the yield per machine is 1,000 units per hour.
Find the expected yield per day; the expected yield lost per day; and the expected repairman idle h per day.

11.4 A one repairman shop has two machines, 1 and 2, where the average run time for machine 1 is $1/\lambda_1$ and for machine 2, it is $1/\lambda_2$. The average service times are $1/\mu$ for both machines. All times are exponential. The probabilities for this system is P_{n1n2j} where n_1 is the number of machine 1 in the service facility, n_2 is the number for machine 2, and j is the machine that is currently in repair. List the equilibrium equations.

Chapter 12

12.1 Suppose a two repairman shop with six machines and the average run time per machine is 5 h and the average service time is one hour. All is exponential.
Find P_0, P_1, P_2, P_3, P_4, P_5, P_6.

12.2 For the shop in 12.1, find the following: Ls, Lq, L, Ws and SL.

12.3 Suppose the shop in 12.1 is open 8 h a day and the yield per machine is 1,000 units per hour. Find the expected yield per day; the expected yield lost per day; and the expected repairman idle hours per day.

12.4 A two repairmen shop has three machines, 1, 2 and 3, where the average run times are $1/\lambda_1$, $1/\lambda_2$ and $1/\lambda_3$, for machines 1, 2 and 3, respectively. The average service times are $1/\mu_1$, $1/\mu_2$ and $1/\mu_3$, for machines 1, 2 and 3, respectively. The probabilities for this system is $P_{n1n2n3j}$ where n_1 is the number of machine 1 in the service facility, n_2 is the number for machine 2, n_3 is the number for machine 3, and j is the machine that is currently in repair. List the probabilities that pertain to this system.

Chapter 13

13.1 A queuing system has one service facility, an infinite queue capacity, and has exponential arrival and service times. The average arrival time is 5 h and the average service time is 3 h. The probability of a faulty unit coming out of the service facility is 0.20; and all faulty units have to repeat the service.
Find P_0, P_1, P_2.

13.2 For the system in 13.1, assume the fee per good unit is $1000, the material cost per unit is $400, and the system is open 40 h a week. Find the following:
a. Expected fee per week.
b. Expected material cost per week.
c. Expected service facility idle hours per week.

Chapter 14

14.1 A queuing system has two service facilities, an infinite queue capacity, and has exponential arrival and service times. The average arrival time is 5 h and the average service time is 3 h. The probability of a faulty unit coming out of the service facility is 0.20 and all faulty units have to repeat the service.
Find P_0, P_1, P_2.

14.2 For the system in 14.1, assume the fee per good unit is $1,000, the material cost per unit is $400, and the system is open 40 h a week. Find the following:
a. Expected fee per week.
b. Expected material cost per week.
c. Expected service facility idle hours per week.

Chapter 15

15.1 Suppose a situation where the units are in tandem for three systems. All is exponential and the queue capacity is infinite in each system. The average arrival time to system 1 is 20 min. The service times to systems 1, 2 and 3 are 10 min, 8 and 4 min, respectively.
For each of the systems in tandem, find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

15.2 For the queuing systems in 15.1, find the average hours in a queue, and the total time in the three systems.

Chapter 16

16.1 A preemptive priority one server system has one service facility with an infinite queue capacity and exponential arrivals from two populations. The average arrival times are 60 min from population 1 (high priority), and 15 min from population 2 (low priority). The average service times are 6 min for low and high priority units. All are exponential and the queue capacity is infinite. For an arbitrary unit in the system, find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

16.2 For the system in 16.1, find the following statistics for high priority units: P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

16.3 For the system in 16.2, find the following statistics for the low priority units: Lq, Ls, L, Wq, Ws, W, Wq′.

Chapter 17

17.1 A preemptive priority one server system has one service facility with an infinite queue capacity and exponential arrivals from two populations. The average arrival times for high priority units are 8 min, and for the low priority units it is 2 min. The service times have a mean of 2 min and a standard deviation of zero for high priority, and a mean of one minute and a standard deviation of zero for low priority.
For an arbitrary unit in the system, find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

17.2 For the system in 17.1, find the following statistics for high priority units: P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.

17.3 For the system in 17.2, find the following statistics for the low priority units: Lq, Ls, L, Wq, Ws, W, Wq′.

Chapter 18

18.1 A queuing system has one service facility, an infinite queue capacity, and has exponential arrival times. The service times are constant. The average arrival time is 5 h and the service time is 3 h.
Find, Lq, Ls, L, Wq, Ws, W, Wq′ and SL.
18.2 For the system in 18.1, find P_0, P_1 and P_2.

Chapter 19

19.1 Consider a one server queuing system with exponential arrivals and Erlang service with three stages. The queue capacity is infinite. The average arrival time is 20 min and the average service time is 18 min.
Find P_0, Lq, Ls, L, Wq, Ws, W, Wq′ and SL
19.2 On Problem 19.1, suppose the capacity is N = 2 for the number of units in the system. List the equilibrium equations.

Chapter 20

20.1 Consider a one server queuing system with Erlang arrivals of 2 stages, and exponential service. Assume the capacity is N = 3. The average arrival time is 20 min and the average service time is 18 min.
List the equilibrium equations for this system.
20.2 For the system in 20.1, list the matrices, A, B, P that would be needed to solve for the state probabilities.

Chapter 21

21.1 Consider a one server queuing system with Erlang arrivals of 2 stages, and Erlang service of 2 stages. Assume the capacity is N = 2. The average arrival time is 20 min and the average service time is 18 min.
List the equilibrium equations for this system.
21.2 For the system in 21.1, list the matrices, A, B, P that would be needed to solve for the state probabilities.

Chapter 22

22.1 A queuing system has one service facility, an infinite queue capacity, and has exponential arrival and service times. The average arrival time is 5 h and the average service time is 3 h. If t is the minutes of wait time in the queue, find, P(t = 0), P(t > 0), P(t > 5) and P(t > 10).
22.2 For the system in 22.1, find the conditional probability: P(t >10 min |t>5 min).

Chapter 23

23.1 A queuing system has two service facilities, an infinite queue capacity, and has exponential arrival and service times. The average arrival time is 5 h and the average service time is 3 h. If t is the wait time in the queue, find, $P(t = 0)$, $P(t > 0)$ and $P(t > 10$ min).

23.2 For the system in 23.1, find the conditional probability: $P(t > 10$ min$|t > 5$ min).

Solutions

3.1 Lq = 0.90, Ls = 0.60, L = 1.50
Wq = 4.50 h, Ws = 3.00 h, W = 7.50 h, Wq′ = 7.50 h
SL = 0.40
3.2 P_n (n = 0,3) = (0.400, 0.240, 0.144, 0.086)
P(n ≥4) = 0.130
3.3 E(fee/week) = $4000, E(idle hours per week) = 16 h.

4.1 Lq = 0.444, Ls = 0.178, L = 0.622
Wq = 2.22 h, Ws = 3.00 h, W = 5.22 h, Wq′ = 4.11 h
4.2 P_n (n = 0,3) = (0.460, 0.276, 0.166, 0.099)
SL = 0.46, Ploss = 0.099, λ_e = 0.18/hour
4.3 E(fee/week) = $3600, E(idle hours per week) = 18.4 h,
E(customers lost/week) = 0.80
4.4 P_0 = 1/11, P_1 = 4/11, P_2 = 6/11, Ls = 10/11, Lq = 6/11, L = 16/11, SL = 1/11.
4.5 E(fee/week) = $1164, E(labor cost/week) = $800, E(idle hours /week) = 3.64

5.1 Lq = 0.00, Ls = 0.375, L = 0.375
Wq = 0 h, Ws = 3.00 h, W = 3.00 h, Wq′ = 0 h
SL = 0.625, Ploss = 0.375, λ_e = 0.125/hour
5.2 P_0 = 0.625, P_1 = 0.375
5.3 E(fee/week) = $2500, E(idle hours per week) = 25 h,
E(customers lost/week) = 3.00

6.1 Lq = 0.06, Ls = 0.60, L = 0.66
Wq = 0.30 h, Ws = 3.00 h, W = 3.30 h, Wq′ = 2.13 h
SL = 0.86
6.2 P_n (n = 0,3) = (0.538, 0.323, 0.097, 0.032)
P(n≥4) = 0.900
6.3 E(fee/week) = $4000, E(idle hours per week) = 21.56 h.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0, © Springer Science+Business Media New York 2012

7.1 Lq = 0.029, Ls = 0.583, L = 0.612
Wq = 0.149 h, Ws = 3.00 h, W = 3.149 h, $Wq' = 1.173$ h
7.2 P_n (n = 0,3) = (0.545, 0.327, 0.098, 0.029)
SL = 0.872, Ploss = 0.029, $\lambda_e = 0.194$/hour
7.3 E(fee/week) = $3880, E(idle hours per week) = 56.68 h,
E(customers lost/week) = 0.232
7.4 $P_0 = 0.173$, $P_1 = 0.431$, $P_2 = 0.289$, $P_3 = 0.108$,
Ls = 1.225, Lq = 0.108, L = 1.333, SL = 0.604.
7.5 $P_{00} = 0.15$, $P_{10} = 0.60$, $P_{20} = 0.20$, $P_{01} = 0.05$

8.1 Lq = 0.00, Ls = 0.539, L = 0.539
Wq = 0 h, Ws = 3.00 h, W = 3.00 h, $Wq' = 0$ h
8.2 $P_0 = 0.562$, $P_1 = 0.337$, $P_2 = 0.101$
SL = 0.899, Ploss = 0.101, $\lambda_e = 0.099$/hour
8.3 E(fee/week) = $1980, E(idle hours per week) = 58.44 h,
E(customers lost/week) = 3.96
8.4 $P_0 = 0.615$, $P_1 = 0.308$, $P_2 = 0.077$, Ls = 0.462, SL = 0.923,

9.1 $P_0 = 0.40$, Lq = 0.50, Ls = 0.60, L = 1.10
Wq = 2.5 min, Ws =3.0 min, W = 5.5 min, $Wq' = 4.17$ min
SL = 0.40
9.2 $P_0 = 0.25$, Lq = 1.15, Ls = 0.75, L = 1.90,
Wq = 5.75 min, Ws =3.00 min , W = 8.75 min, $Wq' = 7.67$ min
SL = 0.25
9.3 $P_0 = 0.278$, Lq = 0.984, Ls = 0.722, L = 1.7060,
Wq = 4.92 min, Ws =3.60 min , W = 8.52 min, $Wq' = 6.81$ min
SL = 0.278

10.1 $P_0 = 0.25$, Lq = 1.125, Ls = 0.750, L = 1.875
Wq = 1.80 min, Ws = 1.20 min, W = 3.00 min, $Wq' = 2.40$ min
SL = 0.25
10.2 Population 1: Lq = 0.225, Ls = 0.150, L = 0.375
Wq = 0.36 min, Ws = 2.00 min, W = 2.36 min, $Wq' = 0.48$ min
SL = 0.25
Populations 2: Lq = 0.900 min, Ls = 0.600 min, L = 1.500 min
Wq = 1.44 min, Ws = 1.00 min, W = 2.44 min, $Wq' = 1.92$ min
SL = 0.25

11.1 P_n (n =0,6) = (0.185, 0.222, 0.222, 0.178, 0.107, 0.074, 0.015)
11.2 Ls = 0.81, Lq = 1.15, L = 1.96, Ws = 1.00, SL = 0.19
11.3 E(yield/day) = 32,320, E(lost yield/day) = 15,680,
E(Repairman lost hours/day) =1.48 h
11.4 $0 = -(\lambda_1 + \lambda_2)P_{000} + \mu_1 P_{101} + \mu_2 P_{012}$
$0 = -(\lambda_2 + \mu_1)P_{101} + \lambda_1 P_{000} + \mu_2 P_{112}$
$0 = -(\lambda_1 + \mu_2)P_{012} + \lambda_2 P_{000} + \mu_1 P_{111}$

$0 = -\mu_1 P_{111} + \lambda_2 P_{101}$
$0 = -\mu_2 P_{112} + \lambda_1 P_{012}$

12.1 P_n (n = (0,6)= (0.320, 0.384, 0.192, 0.077, 0.023, 0.004, 0.000)
12.2 Ls = 0.98, Lq = 0.14, L = 1.12, Ws = 1.00 hour, SL = 0.70
12.3 E(yield/day) = 39040, E(yield lost/day) = 8,960)
E(repairmen idle hours/day) = 8.192 h
12.4 P_{0000}, P_{1001}, P_{0102}, P_{0013}, P_{1101}, P_{1102}, P_{1011}, P_{1013}, P_{0112}, P_{0113}, P_{1111}, P_{1112}, P_{1113}

13.1 $P_0 = 0.2500$, $P_1 = 0.1875$, $P_2 = 0.1405$
13.2 E(fee/week) = \$8,000, E(material cost/week) = \$4000
E(service facility idle hours/week) = 10 h

14.1 $P_0 = 0.211$, $P_1 = 0.158$, $P_2 = 0.059$
14.2 E(fee/week) = \$8,000, E(material cost/week) = \$4000
E(service facility idle hours/week) = 23.2 h

15.1 System 1: $P_0 = 0.50$, Lq = 0.50, Ls = 0.50, L = 1.00
Wq = 10 min, Ws = 10 min, W = 20 min, Wq′ = 20 min
SL = 0.50
System 2: $P_0 = 0.60$, Lq = 0.27, Ls = 0.40, L = 0.67
Wq = 5.328 min, Ws = 8 min, W = 13.328 min, Wq′ = 13.328 min
SL = 0.60
System 3: $P_0 = 0.80$, Lq = 0.05, Ls = 0.20, L = 0.25
Wq = 1.00 min, Ws = 4.00 min, W = 5.00 min, Wq′ = 5.00 min
SL = 0.80

16.1 All: $P_0 = 0.50$, Lq = 0.50, Ls = 0.50, L = 1.00
Wq = 6 min, Ws = 6 min, W = 12 min, Wq′ = 12 min
SL = 0.50
16.2 High priority: $P_0 = 0.90$, Lq = 0.01, Ls = 0.10, L = 0.11
Wq = 0.66 min, Ws = 6 min, W = 6.66 min, Wq′ = 6.66 min
SL = 0.90
16.3 Low priority: Lq = 0.49, Ls = 0.40, L = 0.89
Wq = 7.78 min, Ws = 6 min, W = 13.78 min, Wq′ = 13.78 min

17.1 All: $P_0 = 0.250$, Lq = 1.25, Ls = 0.75, L = 2.00
Wq = 1.04 min, Ws = 1.20 min, W = 2.24 min, Wq′ = 1.39 min
SL = 0.25
17.2 High priority: $P_0 = 0.75$, Lq = 0.042, Ls = 0.250, L = 0.292
Wq = 0.336 min, Ws = 2 min, W = 2.336 min, Wq′ = 1.394 min
SL = 0.75

17.3 Low priority: Lq = 1.208, Ls = 0.50, L = 1.708
Wq = 1.216 min, Ws = 1 min, W = 2.216 min, Wq′ = 1.389 min

18.1 $P_0 = 0.40$, Ls = 0.60, Lq = 0.45, L = 1.05
Ws = 3 h, Wq = 2.25 h, W = 5.25 h, Wq′ = 0.75 h
18.2 $P_0 = 0.40$, $P_1 = 0.329$, $P_2 = 0.162$

19.1 $P_0 = 0.10$, Ls = 0.90, Lq = 5.40, L = 6.30
Ws = 18 min, Wq = 108 min, W = 116 min, Wq′ = 120 min
SL = 0.10
19.2 $0 = -\lambda P_{00} + 3\mu P_{13}$
$0 = -(\lambda+3\mu)P_{11} + \lambda P_{00} + 3\mu P_{23}$
$0 = -(\lambda+3\mu)P_{12} + 3\mu P_{11}$
$0 = -(\lambda+3\mu)P_{13} + 3\mu P_{12}$
$0 = -(3\mu)P_{21} + \lambda P_{11}$
$0 = -(3\mu)P_{22} + \lambda P_{12} + 3\mu P_{21}$
$0 = -(3\mu)P_{23} + \lambda P_{13} + 3\mu P_{22}$

20.1 $0 = -2\lambda P_{01} + \mu P_{11}$
$0 = -2\lambda P_{02} + 2\lambda P_{01} + \mu P_{12}$
$0 = -(2\lambda+\mu)P_{11} + 2\lambda P_{02} + \mu P_{21}$
$0 = -(2\lambda+\mu)P_{12} + 2\lambda P_{11} + \mu P_{22}$
$0 = -(2\lambda+\mu)P_{21} + 2\lambda P_{12} + \mu P_{31}$
$0 = -(2\lambda+\mu)P_{22} + 2\lambda P_{21} + \mu P_{32}$
$0 = -(2\lambda+\mu)P_{31} + 2\lambda P_{22} + 2\lambda P_{32}$
$0 = -(2\lambda+\mu)P_{32} + 2\lambda P_{31}$
20.2

$$\mathrm{A} = \begin{bmatrix} 0 & 0.056 & 0 & 0 & 0 & 0 & 0 \\ -0.10 & 0 & 0.056 & 0 & 0 & 0 & 0 \\ 0.10 & -0.156 & 0 & 0.056 & 0 & 0 & 0 \\ 0 & 0.10 & -0.156 & 0 & 0.056 & 0 & 0 \\ 0 & 0 & 0.10 & -0.156 & 0 & 0.056 & 0 \\ 0 & 0 & 0 & 0.10 & -0.156 & 0 & 0.056 \\ 0 & 0 & 0 & 0 & 0.10 & -0.156 & 0.10 \end{bmatrix}$$

$$\mathrm{B} = \begin{bmatrix} 0.10 \\ -0.10 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathrm{P} = \begin{bmatrix} P_{02} \\ P_{11} \\ P_{12} \\ P_{21} \\ P_{22} \\ P_{31} \\ P_{32} \end{bmatrix}$$

21.1 $0 = -0.10P_{010} + 0.11P_{112}$
$0 = -0.10P_{020} + 0.10P_{010} + 0.11P_{122}$
$0 = -0.21P_{111} + 0.10P_{020} + 0.11P_{212}$
$0 = -0.21P_{112} + 0.11P_{111}$
$0 = -0.21P_{121} + 0.10P_{111} + 0.11P_{222}$
$0 = -0.21P_{122} + 0.10P_{112} + 0.11P_{121}$
$0 = -0.21P_{211} + 0.10P_{121} + 0.10P_{221}$
$0 = -0.21P_{212} + 0.10P_{122} + 0.10P_{222} + 0.11P_{211}$
$0 = -0.21P_{221} + 0.10P_{211}$
$0 = -0.21P_{222} + 0.10P_{212} + 0.11P_{221}$

21.2

$$A = \begin{bmatrix} 0 & 0 & 0.11 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0..10 & 0 & 0 & 0 & 0.11 & 0 & 0 & 0 & 0 \\ 0.10 & -0.21 & 0 & 0 & 0 & 0 & 0.11 & 0 & 0 \\ 0 & 0.11 & -0.21 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.10 & 0 & -0.21 & 0 & 0 & 0 & 0 & 0.11 \\ 0 & 0 & 0.10 & 0.11 & -0.21 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.10 & 0 & -0.21 & 0 & 0.10 & 0 \\ 0 & 0 & 0 & 0 & 0.10 & 0.11 & -0.21 & 0 & 0.10 \\ 0 & 0 & 0 & 0 & 0 & 0.10 & 0 & -0.21 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} -0.10 \\ 0.10 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad P = \begin{bmatrix} P_{020} \\ P_{111} \\ P_{112} \\ P_{121} \\ P_{122} \\ P_{211} \\ P_{212} \\ P_{221} \\ P_{222} \end{bmatrix}$$

22.1 $P(t = 0) = 0.40$, $P(t > 0) = 0.60$, $P(t > 5) = 0.594$, $P(t > 10) = 0.587$
22.2 $P(t > 10|t>5) = 0.988$

23.1 $P(t = 0) = 0.865$, $P(t > 0) = 0.135$, $P(t > 5) = 0.014$, $P(t > 10) = 0.001$
23.2 $P(t > 10|t>5) = 0.071$

Index

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0, © Springer Science+Business Media New York 2012