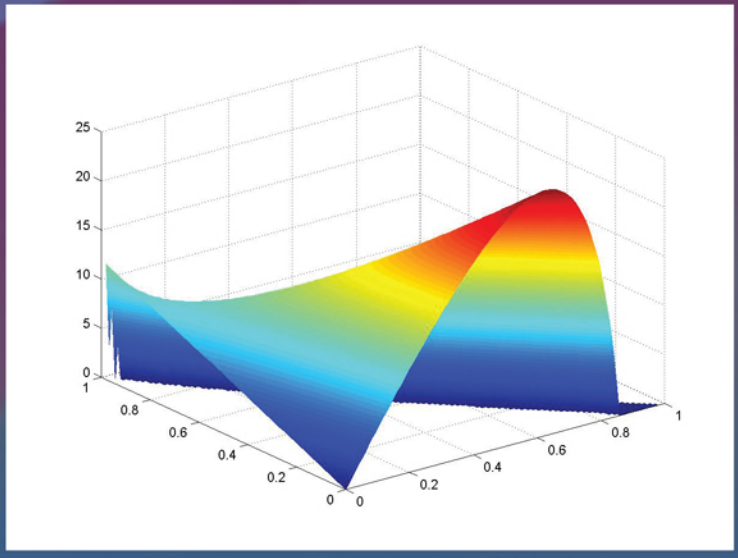
Optimal Design of Queueing Systems



Shaler Stidham, Jr.

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Optimal Design of Queueing Systems

Shaler Stidham, Jr.
University of North Carolina
Chapel Hill, North Carolina, U. S. A.

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20131120

International Standard Book Number-13: 978-1-4200-1000-8 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

List of Figures

Preface

What began a long time ago as a comprehensive book on optimization of queueing systems has evolved into two books: this one on optimal design and a subsequent book (still in the works) on optimal control of queueing systems.

In this setting, "design" refers to setting the parameters of a queueing system (such as arrival rates and service rates) before putting it into operation. By contrast, in "control" problems the parameters are control variables in the sense that they can be varied dynamically in response to changes in the state of the system.

The distinction between design and control, admittedly, can be somewhat artificial. But the available material had outgrown the confines of a single book and I decided that this was as good a way as any of making a division.

Why look at design models? In principle, of course, one can always do better by allowing the values of the decision variables to depend on the state of the system, but in practice this is frequently an unattainable goal. For example, in modern communication networks, real-time information about the buffer contents at the various nodes (routers/switches) of the network would, in principle, help us to make good real-time decisions about the routing of messages or packets. But such information is rarely available to a centralized controller in time to make decisions that are useful for the network as a whole. Even if it were available, the combinatorial complexity of the decision problem makes it impossible to solve even approximately in the time available. (The essential difficulty with such systems is that the time scale on which the system state is evolving is comparable to, or shorter than, the time scale on which information can be obtained and calculations of optimal policies can be made.) For these and other reasons, those in the business of analyzing, designing, and operating communication networks have turned their attention more and more to *flow control*, in which quantities such as arrival (e.g., packet-generation) rates and service (e.g., transmission) rates are computed as time averages over periods during which they may be reasonably expected to be constant (e.g., peak and off-peak hours) and models are used to suggest how these rates can be controlled to achieve certain objectives. Since this sort of decision process involves making decisions about rates (time averages) and not the behavior of individual messages/packets, it falls under the category of what I call a design problem. Indeed, many of the models, techniques, and results discussed in this book were inspired by research on flow and routing control that has been reported in the literature on communication networks.

Of course, flow control is still control in the sense that decision variables can

change their values in response to changes in the state of the system, but the states in question are typically at a higher level, involving congestion averages taken over time scales that are much longer than the time scale on which such congestion measures as queue lengths and waiting times are evolving at individual service facilities. For this reason, I believe that flow control belongs under the broad heading of design of queueing systems.

I have chosen to frame the issues in the general setting of a queueing system, rather than specific applications such as communication networks, vehicular traffic flow, supply chains, etc. I believe strongly that this is the most appropriate and effective way to produce applicable research. It is a belief that is consistent with the philosophy of the founders of operations research, who had the foresight to see that it is the underlying structure of a system, not the physical manifestation of that structure, that is important when it comes to building and applying mathematical models.

Unfortunately, recent trends have run counter to this philosophy, as more and more research is done within a particular application discipline and is published in the journals of that discipline, using the jargon of that discipline. The result has been compartmentalization of useful research. Important results are sometimes rediscovered in, say, the communication and computer science communities, which have been well known for decades in, say, the traffic-flow community.

I blame the research funding agencies, in part, for this trend. With all the best intentions of directing funding toward "applications" rather than "theory," they have conditioned researchers to write grant proposals and papers which purport to deal with specific applications. These proposals and papers may begin with a detailed description of a particular application in which congestion occurs, in order to establish the credibility of the authors within the appropriate research community. When the mathematical model is introduced, however, it often turns out to be the *M/M/1* queue or some other old, familiar queueing model, disguised by the use of a notation and terminology specific to the discipline in which the application occurs.

Another of my basic philosophies has been to present the various models in a unified notation and terminology and, as much as possible, in a unified analytical framework. In keeping with my belief (expressed above) that queueing theory, rather than any one or several of its applications, provides the appropriate modeling basis for this field, it is natural that I should have adopted the notation and terminology of queueing theory. Providing a unified analytical framework was a more difficult task. In the literature optimal design problems for queueing systems have been solved by a wide variety of analytical techniques, including classical calculus, nonlinear programming, discrete optimization, and sample-path analysis. My desire for unity, together with space constraints, led me to restrict my attention to problems that can be solved for the most part by classical calculus, with some ventures into elementary nonlinear programming to deal with constraints on the design variables. A side benefit of this self-imposed limitation has been that, although the book

is mathematically rigorous (I have not shied away from stating results as theorems and giving complete proofs), it should be accessible to anyone with a good undergraduate education in mathematics who is also familiar with elementary queueing theory. The downside is that I have had to omit several interesting areas of queueing design, such as those involving discrete decision variables (e.g., the number of servers) and several interesting and powerful analytical techniques, such as sample-path analysis. (I plan to include many of these topics in my queueing control book, however, since they are relevant also in that context.)

The emphasis in the book is primarily on qualitative rather than quantitative insights. A recurring theme is the comparison between optimal designs resulting from different objectives. An example is the (by-now-classical) result that the individually optimal arrival rate is typically larger than the socially optimal arrival rate.* This is a result of the fact that individual customers, acting in self-interest, neglect to consider the *external effect* of their decision to enter a service facility: the cost of increased congestion which their decision imposes on other users (see, e.g., Section 1.2.4 of Chapter 1). As a general principle, this concept is well known in welfare economics. Indeed, a major theme of the research on queueing design has been to bring into the language of queueing theory some of the important issues and qualitative results from economics and game theory (the Nash equilibrium being another example). As a consequence this book may seem to many readers more like an economics treatise than an operations research text. This is intentional. I have always felt that students and practitioners would benefit from an infusion of basic economic theory in their education in operations research, especially in queueing theory.

Much of the research reported in this book originated in vehicular traffic-flow theory and some of it pre-dates the introduction of optimization into queueing theory in the 1960s. Modeling of traffic flow in road networks has been done mainly in the context of what someone in operations research might call a "minimum-cost multi-commodity flow problem on a network with nonlinear costs". As such, it may be construed as a subtopic in nonlinear programming. An emphasis in this branch of traffic-flow theory has been on computational techniques and results. Chapters 7 and 8 of this book, which deal with networks of queues, draw heavily on the research on traffic-flow networks (using the language and specific models from queueing theory for the behavior of individual links/facilities) but with an emphasis on qualitative properties of optimal solutions, rather than quantitative computational methods.

Although models for optimal design of queueing systems (using my broad definition) have proliferated in the four decades since the field began, I was surprised at how often I found myself developing new results because I could not find what I wanted in the literature. Perhaps I did not look hard enough. If I missed and/or unintentionally duplicated any relevant research, I ask for-

* But see Section 7.4.4 of Chapter 7 for a counterexample.

bearance on the part of those who created it. The proliferation of research on queueing design, together with the explosion of different application areas each with its own research community, professional societies, meetings, and journals, have made it very difficult to keep abreast of all the important research. I have tried but I may not have completely succeeded.

A word about the organization of the book: I have tried to minimize the use of references in the text, with the exception of references for "classical" results in queueing theory and optimization. References for the models and results on optimal design of queues are usually given in an endnote (the final section of the chapter), along with pointers to material not covered in the book.

Acknowledgements

I would like to thank my editors at Chapman Hall and CRC Press in London for their support and patience over the years that it took me to write this book. I particularly want to thank Fred Hillier for introducing me to the field of optimization of queueing systems a little over forty years ago. I am grateful to my colleagues at the following institutions where I taught courses or gave seminars covering the material in this book: Cornell University (especially Uma Prabhu), Aarhus University (especially Niels Knudsen and Søren Glud Johansen), N.C. State University (especially Salah Elmaghraby), Technical University of Denmark, University of Cambridge (especially Peter Whittle, Frank Kelly, and Richard Weber), and INRIA Sophia Antipolis (especially François Baccelli and Eitan Altman). My colleagues in the Department of Statistics and Operations Research at UNC-CH (especially Vidyadhar Kulkarni and George Fishman) have provided helpful input, for which I am grateful. I owe a particular debt of gratitude to the graduate students with whom I have collaborated on optimal design of queueing systems (especially Tuell Green and Christopher Rump) and to Yoram Gilboa, who helped teach me how to use MATLAB® to create the figures in the book. Finally, my wife Carolyn deserves special thanks for finding just the right combination of encouragement, patience, and (at appropriate moments) prodding to help me bring this project to a conclusion.

CHAPTER 1

Introduction to Design Models

Like the descriptive models in "classical" queueing theory, optimal design models may be classified according to such parameters as the arrival rate(s), the service rate(s), the interarrival-time and service-time distributions, and the queue discipline(s). In addition, the queueing system under study may be a network with several facilities and/or classes of customers, in which case the nature of the flows of the classes among the various facilities must also be specified.

What distinguishes an optimal design model from a traditional descriptive model is the fact that some of the parameters are subject to decision and that this decision is made with explicit attention to economic considerations, with the preferences of the decision maker(s) as a guiding principle. The basic distinctive components of a design model are thus:

1. the decision variables,
2. benefits and costs, and
3. the objective.

Decision variables may include, for example, the arrival rates, the service rates, and the queue disciplines at the various service facilities. Typical benefits and costs include rewards to the customers from being served, waiting costs incurred by the customers while waiting for service, and costs to the facilities for providing the service. These benefits and costs may be brought together in an objective function, which quantifies the implicit trade-offs. For example, increasing the service rate will result in less time spent by the customers waiting (and thus a lower waiting cost), but a higher service cost. The nature of the objective function also depends on the horizon (finite or infinite), the presence or absence of discounting, and the identity of the decision maker (e.g., the facility operator, the individual customer, or the collective of all customers).

Our goal in this chapter is to provide a quick introduction to these basic components of a design model. We shall illustrate the effects of different reward and cost structures, the trade-offs captured by different objective functions, and the effects of combining different decision variables in one model. To keep the focus squarely on these issues, we use only the simplest of descriptive queueing models – primarily the classical *M/M/1* model. By further restricting attention to infinite-horizon problems with no discounting, we shall be able to use the well-known steady-state results for these models to derive closed-form

expressions (in most cases) for the objective function in terms of the decision variables. This will allow us to do the optimization with the simple and familiar tools of differential calculus. Later chapters will elaborate on each of the models introduced in this chapter, relaxing distributional assumptions and considering more general cost and reward structures and objective functions. These more general models will require more sophisticated analytical tools, including linear and nonlinear programming and game theory.

We begin this chapter (Sections 1.1 and 1.2) with two simple examples of optimal design of queueing systems. Both examples are in the context of an isolated *M/M/1* queue with a linear cost/reward structure, in which the objective is to minimize the expected total cost or maximize the expected net benefit per unit time in steady state. In the first example the decision variable is the service rate and in the second, the arrival rate. The simple probabilistic and cost structure makes it possible to use classical calculus to derive analytical expressions for the optimal values of the design variables.

The next three sections consider problems in which more than one design parameter is a decision variable. In Section 1.3, we consider the case where both the arrival rate and service rate are decision variables. Here a simple analysis based on calculus breaks down, since the objective function is not jointly concave and therefore the first-order optimality conditions do not identify the optimal solution. (This will be a recurring theme in our study of optimal design models, and we shall explore it at length in later chapters.) Section 1.4 revisits the problem of Section 1.2 – finding optimal arrival rates – but now in the context of a system with two classes of customers, each with its own reward and waiting cost and arrival rate (decision variable). Again the objective function is not jointly concave and the first-order optimality conditions do not identify the optimal arrival rates. Indeed, the only interior solution to the first-order conditions is a saddle-point of the objective function and is strictly dominated by *both* boundary solutions, in which only one class has a positive arrival rate. Finally, in Section 1.5, we consider the simplest of networks – a system of parallel queues in which each arriving customer must be routed to one of several independent facilities, each with its own queue.

A final word before we start. In a design problem, the values of the decision variables, once chosen, cannot vary with time nor in response to changes in the *state* of the system (e.g., the number of customers present). Design problems have also been called *static control* problems, in contrast to *dynamic control* problems in which the decision variables can assume different values at different times, depending on the observed state of the system. In the literature a static control problem is sometimes called an *open-loop control* problem, whereas a dynamic control problem is called a *closed-loop control* problem. We shall simply use the term *design* for the former and *control* for the latter type of problem.

1.1 Optimal Service Rate

Consider an *M/M/1* queue with arrival rate λ and service rate μ. That is, customers arrive according to a Poisson process with parameter λ. There is a single server, who serves customers one at a time according to a *FIFO* (*F*irst-*I*n-*F*irst-*O*ut) queue discipline. Service times are independent of the arrival process and i.i.d. with an exponential distribution with mean μ^{-1}. Suppose that λ is fixed, but μ is a decision variable.

Examples

1. A machine center in a factory: how fast a machine should we install?
2. A communication system: what should the transmission rate in a communication channel be (e.g., in bits/sec.)?

Performance Measures and Trade-offs.

Typical performance measures are the number of customers in the system (or in the queue) and the waiting time of a customer in the system (or in the queue). If the system operates for a long time, then we might be interested in the long-run average or the expected steady-state number in the system, waiting time, and so forth. All these are measures of the level of *congestion*. As μ increases, the congestion (as measured by any of these quantities) decreases. (Of course this property is not unique to *M/M/1* systems.) Therefore, to minimize congestion, we should choose as large a value of μ as possible (e.g., $\mu = \infty$, if there is no finite upper bound on μ). But, in all real systems, increasing the service rate costs something. Thus there is a trade-off between decreasing the congestion and increasing the cost of providing service, as μ increases. One way to capture this trade-off is to consider a simple model with linear costs.

1.1.1 A Simple Model with Linear Service and Waiting Costs

Suppose there are two types of cost:

(i) a service-cost rate, c (cost per unit time per unit of service rate); and

(ii) a waiting-cost rate h (cost per unit time per customer in system).

In other words, (i) if we choose service rate μ, then we pay a service cost $c \cdot \mu$ per unit time; (ii) a customer who spends t time units in the system accounts for $h \cdot t$ monetary units of waiting cost, or equivalently, the system incurs $h \cdot i$ monetary units of waiting cost per unit time while i customers are present. Suppose our objective is to minimize the long-run average cost per unit time. Now it follows from standard results in descriptive queueing theory (or the general theory of continuous-time Markov chains) that the long-run average cost equals the expected steady-state cost, if steady state exists (which is true if and only if $\mu > \lambda$). Otherwise the long-run average cost equals ∞. Therefore, without loss of generality let us assume $\mu > \lambda$.

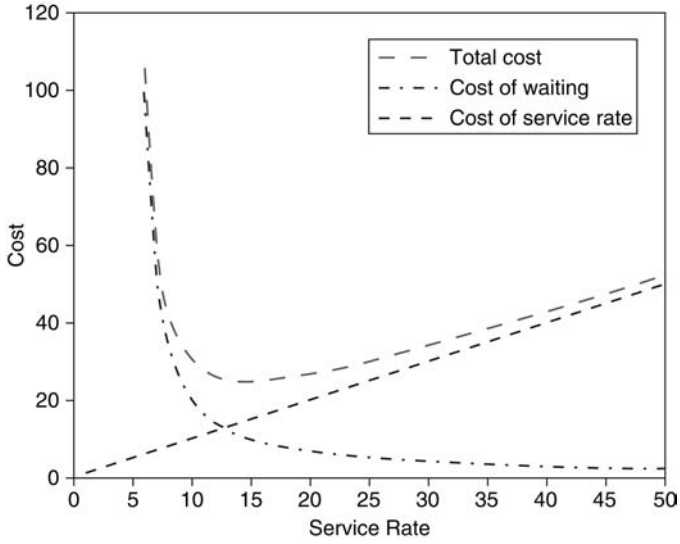


Figure 1.1 *Total Cost as a Function of Service Rate*

Let $C(\mu)$ denote the expected steady-state total cost per unit time, when service rate μ is chosen. Then

$$C(\mu) = c \cdot \mu + h \cdot L(\mu) \ ,$$

where $L(\mu)$ is the expected steady-state number in system. For a *FIFO M/M/1* queue, it is well known (see, e.g., Gross and Harris [79]) that

$$L(\mu) = \lambda W(\mu) = \frac{\lambda}{\mu - \lambda} \ , \tag{1.1}$$

where $W(\mu)$ is the expected steady-state waiting time in system.* Thus our optimization problem takes the form:

$$\min_{\{\mu:\mu>\lambda\}} C(\mu) = c \cdot \mu + h \cdot \left(\frac{\lambda}{\mu - \lambda} \right) \ . \tag{1.2}$$

Note that

$$C''(\mu) = \frac{2h\lambda}{(\mu - \lambda)^3} > 0 \ , \text{ for all } \mu > \lambda \ ,$$

so that $C(\mu)$ is convex in $\mu \in (\lambda, \infty)$. Moreover, $C(\mu) \to \infty$ as $\mu \downarrow \lambda$ and as $\mu \uparrow \infty$. (See Figure 1.1.) Hence we can solve this problem by differentiating $C(\mu)$ and setting the derivative equal to zero:

$$C'(\mu) = c - \frac{h\lambda}{(\mu - \lambda)^2} = 0 \ . \tag{1.3}$$

* The expression (1.1) holds more generally for any work-conserving queue discipline that does not use information about customer service times. See, e.g., El-Taha and Stidham [60].

This yields the following expression for the unique optimal value of the service rate, denoted by μ^*:

$$\mu^* = \lambda + \sqrt{\frac{\lambda h}{c}} \ . \tag{1.4}$$

The optimal value of the objective function is thus given by

$$C(\mu^*) = c\left(\lambda + \sqrt{\lambda h/c}\right) + \lambda h/\sqrt{\lambda h/c} = c\lambda + \sqrt{\lambda hc} + \sqrt{\lambda hc} \ .$$

This expression has the following interpretation. The term $c \cdot \lambda$ represents the *fixed* cost of providing the minimum possible level of service, namely, $\mu = \lambda$. The next two terms – both equal to $\sqrt{\lambda hc}$ – represent, respectively, the service cost and the waiting cost associated with the optimal "surplus" service level, $\mu^* - \lambda$. Note that an optimal solution divides the *variable* cost equally between service cost and waiting cost.

More explicitly, if one reformulates the problem in equivalent form with the *surplus* service rate, $\tilde{\mu} := \mu - \lambda$, as the decision variable and removes the fixed-cost term, $c\lambda$, from the objective function, then the new objective function, denoted by $\tilde{C}(\tilde{\mu})$, takes the form

$$\tilde{C}(\tilde{\mu}) = c\tilde{\mu} + h\lambda/\tilde{\mu} \ . \tag{1.5}$$

The optimal value of $\tilde{\mu}$ is given by

$$\tilde{\mu}^* = \sqrt{\frac{\lambda h}{c}} \ ,$$

and the optimal value of the objective function by

$$\tilde{C}(\tilde{\mu}^*) = c\sqrt{\lambda h/c)} + \lambda h/\sqrt{\lambda h/c} = \sqrt{\lambda hc} + \sqrt{\lambda hc} \ .$$

It is the particular structure of the objective function (1.5) – the sum of a term proportional to the decision variable and a term proportional to its reciprocal – that leads to the property that an optimal solution equates the two terms, a property that of course does not hold in general when one is minimizing the sum of two cost terms. The general condition for optimality (cf. equation (1.3)) is that the *marginal increase* in the first term should equal the *marginal decrease* in the second term, not that the terms themselves should be equal. It just happens in this case that the latter property holds when the former does.

Readers familiar with inventory theory will note the structural equivalence of the objective function (1.5) to the objective function in the classical economic-lot-size problem and the resulting similarity between the formula for $\tilde{\mu}^*$ and the economic-lot-size formula.

1.1.2 Extensions and Exercises

1. *Constraints on the Service Rate.* Suppose the service rate is constrained to lie in an interval, $\mu \in [\underline{\mu}, \bar{\mu}]$. Characterize the optimal service rate, μ^*,

in this case. Do the same for the case where the feasible values of μ are discrete: $\mu \in \{\mu_1, \mu_2, \ldots, \mu_m\}$.

2. *Nonlinear Waiting Costs.* Suppose in the above model that the customer's waiting cost is a nonlinear function of the time spent by that customer in the system: $h \cdot t^a$, if the time in system equals t, where $a > 0$. (Note that for $a < 1$ the waiting cost $h \cdot t^a$ is concave in t, whereas for $a > 1$ it is convex in t.) Set up and solve the problem of choosing μ to minimize the expected steady-state total cost per unit time, $C(\mu)$. For what values of a is $C(\mu)$ convex in μ?

3. *General Service-Time Distribution.* Consider an *M/GI/1* model, in which the generic service time $\boldsymbol{S}$ has mean $\mathrm{E}[\boldsymbol{S}] = 1/\mu$ and second moment $\mathrm{E}[\boldsymbol{S}^2] = 2\beta/\mu^2$, where $\beta \geq 1/2$ is a given constant and μ is the decision variable. (Thus the coefficient of variation of service time is given by $\sqrt{var(\boldsymbol{S})}/\mathrm{E}[\boldsymbol{S}] = \sqrt{2\beta - 1}$, which is fixed.) In this case the Pollaczek-Khintchine formula yields

$$W(\mu) = \frac{1}{\mu} + \frac{\lambda\beta}{\mu(\mu - \lambda)} \ .$$

Set up the problem of determining the optimal service rate μ^*, with linear waiting cost rates. For what values of β is $C(\mu)$ convex? If possible, find a closed-form expression for μ^* in terms of the parameters, λ, c, h, and β. (The easy cases are when $\beta = 1$ (e.g., exponentially distributed service time) and $\beta = 1/2$ (constant service time, $\boldsymbol{S} \equiv 1/\mu$).)

1.2 Optimal Arrival Rate

Now consider a *FIFO M/M/1* queue in which the service rate μ is fixed and the arrival rate λ is a decision variable.

Examples

1. A machine center: at what rate λ should incoming parts (or subassemblies) be admitted into the work-in-process buffer?

2. A communication system: at what rate λ should messages (or packets) be admitted into the buffer before a communication channel?

Performance Measures and Trade-offs

As λ increases, the throughput (number of jobs served per unit time) increases. (For $\lambda < \mu$, the throughput equals λ; for $\lambda \geq \mu$, the throughput equals μ.) This is clearly a "good thing." On the other hand, the congestion also increases as λ increases, and this is just as clearly a "bad thing." Again a simple linear model offers one way of capturing the trade-off between the two performance measures.

1.2.1 A Simple Model with Deterministic Reward and Linear Waiting Costs

Suppose there is a deterministic reward r per entering customer and (as in the previous model) a waiting cost per customer which is linear at rate h per unit time in the system. Let $B(\lambda)$ denote the expected steady-state net benefit per unit time. Then

$$B(\lambda) = \lambda \cdot r - h \cdot L(\lambda) \ , \tag{1.6}$$

where $L(\lambda)$ is the steady-state expected number of customers in the system, expressed as a function of the arrival rate λ. As in the previous section, we have $L(\lambda) = \lambda W(\lambda)$, where $W(\lambda)$ is the steady-state expected waiting time in the system, and (assuming a first-in, first-out (*FIFO*) queue discipline) $W(\lambda)$ is given by

$$W(\lambda) = \frac{1}{\mu - \lambda} \ , \ 0 \leq \lambda < \mu \ ,$$

with $W(\lambda) = \infty$ for $\lambda \geq \mu$. Again it follows from standard results in descriptive queueing theory that the long-run average cost equals the expected steady-state cost, if steady state exists (which is true if and only if $\lambda < \mu$). Otherwise the long-run average cost equals ∞. Therefore, without loss of generality we assume $\lambda < \mu$.

For the *M/M/1* model, the problem thus takes the form:

$$\max_{\{\lambda \in [0,\mu)\}} r \cdot \lambda - h \cdot \left(\frac{\lambda}{\mu - \lambda} \right) \ . \tag{1.7}$$

The presence of the constraint, $\lambda \geq 0$, makes this problem more complicated than the example of the previous section. Since $B(\lambda) \to -\infty$ as $\lambda \uparrow \mu$, we do not need to concern ourselves about the upper limit of the feasible region. But we must take into account the possibility that the maximum occurs at the lower limit, $\lambda = 0$.

Let λ^* denote the optimal arrival rate. Note that

$$B''(\lambda) = \frac{-2h\mu}{(\mu - \lambda)^3} < 0 \ , \text{ for all } \mu > \lambda \ ,$$

so that $B(\lambda)$ is strictly concave and differentiable in $0 \leq \lambda < \mu$. Therefore its maximum occurs either at $\lambda = 0$ (if $B'(0) \leq 0$) or at the unique value of $\lambda > 0$ at which $B'(\lambda) = 0$ (if $B'(0) > 0$).

It then follows from (1.6) that λ^* is the unique solution in $[0, \mu)$ to the following conditions:

$$\text{(Case 1)} \qquad \lambda = 0 \qquad \text{, if } \ r \leq hL'(0) \ ; \tag{1.8}$$

$$\text{(Case 2)} \quad r = hL'(\lambda) \quad \text{, if } \ r > hL'(0) \ . \tag{1.9}$$

Now for the *M/M/1* queue,

$$L'(\lambda) = \frac{\mu}{(\mu - \lambda)^2} \ ,$$

so that $B'(0) \leq 0$ if $r \leq h/\mu$ and $B'(0) > 0$ if $r > h/\mu$. Therefore

$$\text{(Case 1)} \qquad \lambda^* = 0 \ , \qquad \text{if } r \leq h/\mu \ ;$$

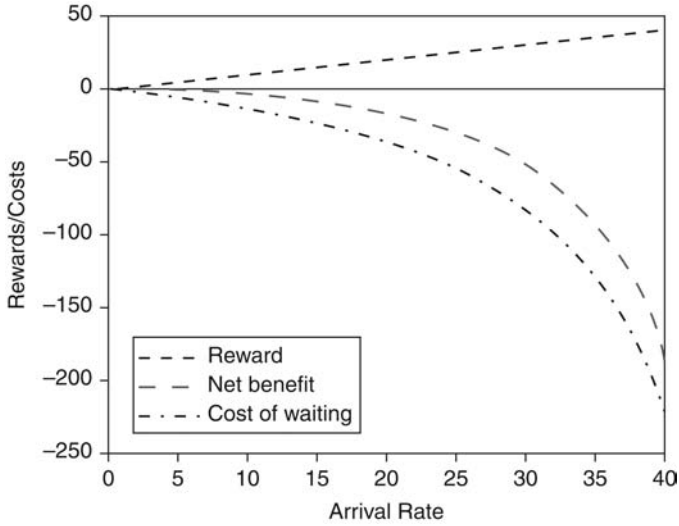


Figure 1.2 *Optimal Arrival Rate, Case 1:* $r \leq h/\mu$

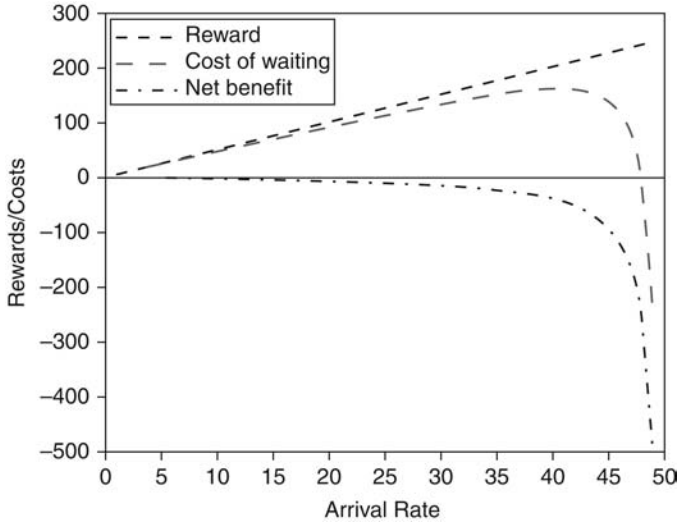


Figure 1.3 *Optimal Arrival Rate, Case 2:* $r > h/\mu$

$$\text{(Case 2)} \quad \lambda^* = \mu - \sqrt{\mu h/r}\,, \quad \text{if } r > h/\mu\,;$$

The two cases are illustrated in Figures 1.2 and 1.3, respectively.

Since $\mu - \sqrt{\mu h/r} > 0$ if and only if $r > h/\mu$, we can combine Cases 1 and 2 as follows:

$$\lambda^* = \left(\mu - \sqrt{\mu h/r}\right)^+ ,$$

where $x^+ := \max\{x, 0\}$. Note that in Case 1 we have $h/\mu \geq r$; that is, the expected waiting cost is at least as great as the reward even for a customer who enters service immediately. Hence it is intuitively clear that $\lambda^* = 0$: there is no economic incentive to admit any customer. If $r > h/\mu$, then it is optimal to allocate λ so that the surplus capacity, $\mu - \lambda$, equals the square root of $\mu h/r$.

1.2.2 Extensions and Exercises

1. *Constraints on the Arrival Rate.* Suppose the feasible set of values for λ is the interval, $[\underline{\lambda}, \bar{\lambda}]$, where $0 \leq \underline{\lambda} < \bar{\lambda} \leq \infty$. The problem now takes the form:

$$\max_{\{\lambda \in [\underline{\lambda}, \bar{\lambda}]\}} \{\lambda \cdot r - hL(\lambda)\} \ . \tag{1.10}$$

   Since $B(\lambda) = -\infty$ for $\lambda \geq \mu$, we can rewrite the problem in equivalent form as

$$\max_{\{\lambda \in [\underline{\lambda}, \min\{\bar{\lambda}, \mu\}]\}} \left\{\lambda \cdot r - h\left(\frac{\lambda}{\mu - \lambda}\right)\right\} \ . \tag{1.11}$$

   (Note that the feasible region reduces to $[\underline{\lambda}, \mu)$ when $\bar{\lambda} \geq \mu$.) Characterize the optimal arrival rate, λ^*, for this problem.

2. *General Service-Time Distribution.* Consider an *M/GI/1* model, in which the generic service time $\boldsymbol{S}$ has mean $\mathrm{E}[\boldsymbol{S}] = 1/\mu$ and second moment $\mathrm{E}[\boldsymbol{S}^2] = 2\beta/\mu^2$, where $\beta \geq 1/2$ is given. The Pollaczek-Khintchine formula yields

$$W(\lambda) = \frac{1}{\mu} + \frac{\lambda\beta}{\mu(\mu - \lambda)} \ .$$

   Set up the problem of determining the optimal arrival rate, λ^*, with deterministic reward and linear waiting cost. Show that λ^* is again characterized by (1.8) and (1.9), and use this result to derive an explicit expression for λ^*, in terms of the parameters, μ, β, r, and h.

1.2.3 An Upper Bound on the Optimal Arrival Rate

Note that

$$B(\lambda) = \lambda r - h\lambda W(\lambda) = \lambda(r - hW(\lambda)) \ , \tag{1.12}$$

so that $B(\lambda) > 0$ for positive values of λ such that $r > hW(\lambda)$ and $B(\lambda) \leq 0$ for values of λ such that $r \leq hW(\lambda)$. If $r \leq hW(0)$ then $r \leq hW(\lambda)$ for all $\lambda \in [0, \mu)$, since $W(\cdot)$ is an increasing function. In this case $\lambda^* = 0$. Otherwise, we can restrict attention, without loss of optimality, to values of λ such that $r > hW(\lambda)$. In the *M/M/1* case, $W(\lambda) = 1/(\mu - \lambda)$, so that $r \leq hW(0)$ if and only if $r \leq h/\mu$. Moreover, $r = hW(\lambda)$ if and only if $\lambda = \mu - h/r$. These observations motivate the following definition.

Define $\bar{\lambda}$ by:

$$\text{(Case 1)} \qquad \bar{\lambda} = 0 \,, \qquad \text{if } r \le h/\mu \,; \tag{1.13}$$

$$\text{(Case 2)} \quad \bar{\lambda} = \mu - h/r \,, \quad \text{if } r > h/\mu \,; \tag{1.14}$$

Since $B(\lambda) \ge 0$ for $0 \le \lambda \le \bar{\lambda}$, and $B(\lambda) \le 0$ for $\bar{\lambda} < \lambda < \mu$, it follows that $\bar{\lambda}$ is an upper bound on λ^*. Moreover, in some contexts $\bar{\lambda}$ can be interpreted as the *individually optimal* (or *equilibrium*) arrival rate, as we shall see presently.

1.2.4 Social vs. Individual Optimization

In our discussion of performance measures and trade-offs, we have been implicitly assuming that the decision maker is the operator of the queueing facility, who is concerned both with maximizing throughput and minimizing congestion. Our reward/cost model assumes that each entering customer generates a benefit r to the facility and that it costs the facility h per unit time per customer in the system. In this section we offer alternative possibilities for who the decision maker(s) might be. But first we must resolve another issue.

We have also been implicitly assuming that the decision maker (whoever it is) can freely choose the arrival rate λ from the interval $[0, \mu)$. How might such a choice be implemented? Here is one possibility.

Suppose that potential customers arrive according to a Poisson process with mean rate Λ ($\Lambda \ge \mu$). A potential customer joins (or is accepted) with probability a and balks (or is rejected) with probability $1 - a$. The accept/reject decisions for successive customers are mutually independent, as well as independent of the number of customers in the system. That is, it is not possible to observe the contents of the queue before the accept/reject decision is made. As a result, customers enter the system according to a Poisson arrival process with mean rate $\lambda = a\Lambda$.† Moreover, a customer who enters with probability a when the arrival rate equals λ receives an expected net benefit equal to

$$a(r - hW(\lambda)) + (1 - a)0 = a(r - hW(\lambda)) \,.$$

Now let us consider the possibility that the decision makers are the customers themselves, rather than the facility operator. We discuss this possibility in the next two subsections.

1.2.4.1 Socially Optimal Arrival Rate

Suppose now that benefits and costs accrue to individual customers and the decision maker represents the collective of all customers. In this case, a reasonable objective for the decision maker is to maximize the expected net benefit received per unit time by the collective of all customers: $B(\lambda) = \lambda(r - hW(\lambda))$. This is precisely the objective function that we have been considering. In this

† Note that the assumption that $\Lambda \ge \mu$ ensures that the feasible region for λ is the interval $[0, \mu)$, as in our original formulation.

context, our probabilistic interpretation of the choice of λ still makes sense. That is, the decision maker, acting on behalf of the collective of all customers, admits each potential arrival with probability $a = \lambda/\Lambda$.

The optimal arrival rate λ^* can now be interpreted as *socially optimal*, since it maximizes *social welfare*, that is, the expected net benefit received per unit time by the collective of all customers, namely $B(\lambda)$. To emphasize this interpretation, we shall henceforth write "λ^s" instead of "λ^*". In the *M/M/1* case, then, the socially optimal arrival rate is given by

$$\lambda^s = (\mu - \sqrt{\mu h/r})^+ \ . \tag{1.15}$$

The system controller can implement λ^s by admitting each potential arrival with probability $a^s := \lambda^s/\Lambda$ and rejecting with probability $1 - a^s$.

1.2.4.2 Comparison with Individually Optimal Arrival Rate

This interpretation of λ^s as the socially optimal arrival rate suggests the following question: how does the socially optimal arrival rate compare to the *individually optimal* arrival rate that results if each individual potential arrival, acting in its own interest, decides whether or not to join?

Suppose (as above) that potential customers arrive according to a Poisson process with arrival rate Λ ($\Lambda \geq \mu$) and each joins the system with probability a and balks with probability $1-a$. Each customer who enters the system when the arrival rate is λ receives a net benefit $r - hW(\lambda)$. A customer who balks receives nothing. As is always the case with design (static control) models, we assume that the decision ($a = 0, 1$) must be made without knowledge of the actual state of the system, e.g., the number of customers present.

Now, however, the criterion for choice of a is purely selfish: each customer is concerned only with maximizing its own expected net benefit. Since a single individual's action has a negligible effect on the system arrival rate λ, each potential customer can take λ as given. For a given λ, the individually optimizing customer seeks to maximize its expected net benefit,

$$a(r - hW(\lambda)) + (1 - a) \cdot 0 \ ,$$

by an appropriate choice of a, $0 \leq a \leq 1$. Thus, the customer will join with probability $a = 1$, if $r > hW(\lambda)$; join with probability $a = 0$, if $r < hW(\lambda)$; and be indifferent among all a, $0 \leq a \leq 1$, if $r = hW(\lambda)$.

Motivated by the concept of a *Nash equilibrium*, we define an *individually optimal* (or *equilibrium* arrival rate, λ^e (and associated joining probability $a^e = \lambda^e/\Lambda$), by the property that no individual customer trying to maximize its own expected net benefit has any incentive to deviate unilaterally from λ^e (a^e). From the above observations, it follows that $\lambda^e = 0$ ($a^e = 0$) if $r \leq hW(0)$ (Case 1), whereas if $r > hW(0)$ (Case 2) then $\lambda^e = a^e\Lambda$ is the (unique) value of $\lambda \in (0, \mu)$ such that

$$r = hW(\lambda) \ . \tag{1.16}$$

To see this, first note that in Case 1 the expected net benefit from choosing a

positive joining probability, $a > 0$, is $a(r - hW(0))$, which is less than or equal to zero, the expected net benefit from the joining probability $a^e = \lambda^e/\Lambda = 0$. Hence, in Case 1 there is no incentive for a customer to deviate unilaterally from $a^e = 0$. In Case 2, since $r - hW(\lambda^e) = 0$, the expected net benefit is

$$a(r - hW(\lambda^e)) + (1 - a) \cdot 0 = 0 \ ,$$

and hence does not depend on the joining probability a. Thus, customers are indifferent among all joining probabilities, $0 \leq a \leq 1$, so that once again there is no incentive to deviate from $a^e = \lambda^e/\Lambda$.

Since $W(\lambda) = 1/(\mu - \lambda)$ in the *M/M/1* case, we see that the individually optimal arrival rate λ^e coincides with $\bar{\lambda}$ as defined by (1.13) and (1.14). But we have shown that $\lambda^* = \lambda^s \leq \bar{\lambda} = \lambda^e$. In other words, the *socially optimal* arrival rate, λ^s, is less than or equal to the individually optimal arrival rate, λ^e.

The following theorem summarizes these results:

Theorem 1.1 *The socially optimal arrival rate is no larger than the* individually optimal *arrival rate:* $\lambda^s \leq \lambda^e$. *Moreover,* $\lambda^s = \lambda^e = 0$, *if* $r \leq h/\mu$, *and* $0 < \lambda^s < \lambda^e$, *if* $r > h/\mu$.

A review of our arguments above will show that this property is not restricted to *M/M/1* systems and is in fact quite general. In fact, this theorem is valid for *any* system (for example, a *GI/GI/1* queue) in which the following conditions hold:

1. $W(\lambda)$ is strictly increasing in $0 \leq \lambda < \mu$;
2. $W(\lambda) \uparrow \infty$ as $\lambda \uparrow \mu$;
3. $W(0) = 1/\mu$.

1.2.5 Internal and External Effects

Suppose $r > h/\mu$. It follows from (1.12) that

$$B'(\lambda) = r - [h \cdot W(\lambda) + h \cdot \lambda W'(\lambda)] \ ,$$

and that λ^s is found by equating $h \cdot W(\lambda) + h \cdot \lambda W'(\lambda)$ to r, whereas (cf. (1.16)) λ^e is found by equating $h \cdot W(\lambda)$ to r. We can interpret $h \cdot W(\lambda)$ as the *internal effect* and $h \cdot \lambda W'(\lambda)$ as the *external effect* of a marginal increase in the arrival rate. The quantity $h \cdot W(\lambda)$ is the waiting cost of the marginal customer who joins when the arrival rate is λ. It is "internal" in that it is a cost borne only by the customer itself. On the other hand, the quantity $h \cdot \lambda W'(\lambda)$ is the marginal increase in waiting cost incurred by all the customers as a result of a marginal increase in the arrival rate. It is "external" to the marginal joining customer, since it is a cost which that customer does not incur. The fact that $\lambda^s \leq \lambda^e$ (that is, customers acting in their own interest join the system more frequently than is socially optimal) is due to an individually optimizing customer's failure to take into account the external effect of its decision to enter. The formula for λ^e only takes into account the internal effect of the decision to enter, that

is the customer's own waiting cost, $hW(\lambda)$. By contrast, the formula for λ^s takes into account both the internal effect, $hW(\lambda)$, and the external effect, $h\lambda W'(\lambda)$.

It follows that individually optimizing customers can be induced to behave in a socially optimal way by charging each entering customer a fee or *congestion toll* equal to the external effect, $h\lambda W'(\lambda)$. In this way arrival control can be *decentralized*, in the sense that each individual customer can be left to make its own decision. (Again, note that these results hold for any system in which $W(\lambda)$ is a well defined function satisfying conditions (1)–(3). See Chapter 2 for further analysis and generalizations.)

1.3 Optimal Arrival Rate and Service Rate

Now let us consider an *M/M/1* queue in which both the arrival rate λ and the service rate μ are decision variables. We shall use a reward/cost model that combines the features of the models of the last two sections. There is a reward r per entering customer, a waiting cost h per unit time per customer in the system, and a service cost c per unit time per unit of service rate. The objective function (to be maximized) is the steady-state expected net benefit per unit time, $B(\lambda, \mu)$, that is,

$$B(\lambda, \mu) = \lambda \cdot r - h \cdot L(\lambda, \mu) - c \cdot \mu \ , \ 0 \leq \lambda < \mu \ ,$$

with $B(0, 0) = 0$. (Note that $B(\lambda, \mu)$ has a discontinuity at $(0, 0)$.) If $c \geq r$, then obviously the optimal solution is $\lambda^* = \mu^* = 0$, with net benefit $B(0, 0) = 0$, since for all $0 \leq \lambda < \mu$ we have $B(\lambda, \mu) < 0$. Henceforth we shall assume that $c < r$, in which case we can exclude the point $(0, 0)$ and restrict attention to the region $\{(\lambda, \mu) : 0 \leq \lambda < \mu\}$, since it contains pairs (λ, μ) for which $B(\lambda, \mu) > 0$. Note that $B(\lambda, \mu)$ is continuously differentiable over this region.

Following the program of the previous two sections, let us use the first-order optimality conditions to try to identify the optimal pair, (λ^*, μ^*). Differentiating $B(\lambda, \mu)$ with respect to λ and μ and setting the derivatives equal to zero leads to the equations,

$$\begin{aligned} \frac{\partial}{\partial \lambda} B(\lambda, \mu) &= r - h \cdot \frac{\partial}{\partial \lambda} L(\lambda, \mu) = 0 \ , \\ \frac{\partial}{\partial \mu} B(\lambda, \mu) &= -h \cdot \frac{\partial}{\partial \mu} L(\lambda, \mu) - c = 0 \ . \end{aligned}$$

Since $L(\lambda, \mu) = \lambda/(\mu - \lambda)$, for $0 \leq \lambda < \mu$, we have

$$\frac{\partial}{\partial \lambda} L(\lambda, \mu) = \frac{\mu}{(\mu - \lambda)^2} \ , \ \frac{\partial}{\partial \mu} L(\lambda, \mu) = \frac{-\lambda}{(\mu - \lambda)^2} \ ,$$

from which we obtain the following two simultaneous equations for λ and μ,

$$\frac{h \cdot \mu}{(\mu - \lambda)^2} = r \ ,$$

$$\frac{h \cdot \lambda}{(\mu - \lambda)^2} = c \,,$$

the unique solution to which is

$$\lambda = \frac{h \cdot c}{(r-c)^2} \,,\ \mu = \frac{h \cdot r}{(r-c)^2} \,. \tag{1.17}$$

Note that this solution is feasible (that is, $\lambda < \mu$) since $c < r$.

To recapitulate, under the assumption that $c < r$, we have identified a unique interior point of the feasible region ($0 < \lambda < \mu$) that satisfies the first-order optimality conditions. Surely this must be the optimal solution. After all, we have simply brought together the two models and analyses of the previous sections, in which μ and λ, respectively, were decision variables and in the course of which we verified that our objective function, $B(\lambda, \mu)$, is both concave in λ and concave in μ. What we have not verified, however, is joint concavity in (λ, μ). Without joint concavity, we cannot be sure that a solution to the first-order optimality conditions is a local (let alone a global) maximum.

In fact $B(\lambda, \mu)$ is *not* jointly concave in (λ, μ), because $L(\lambda, \mu) = \lambda/(\mu - \lambda)$ is not jointly convex . To check for joint convexity, we must evaluate

$$\Delta := \left(\frac{\partial^2 L}{\partial \lambda^2}\right)\left(\frac{\partial^2 L}{\partial \mu^2}\right) - \left(\frac{\partial^2 L}{\partial \lambda \partial \mu}\right)^2$$

and check whether Δ is nonnegative. Since

$$\begin{aligned}
\frac{\partial^2 L}{\partial \lambda^2} &= \frac{2\mu}{(\mu-\lambda)^3} \,, \\
\frac{\partial^2 L}{\partial \mu^2} &= \frac{2\lambda}{(\mu-\lambda)^3} \,, \\
\frac{\partial^2 L}{\partial \lambda \mu} &= \frac{-(\lambda+\mu)}{(\mu-\lambda)^3} \,,
\end{aligned}$$

we have

$$\begin{aligned}
\Delta &= \left(\frac{2\mu}{(\mu-\lambda)^3}\right)\left(\frac{2\lambda}{(\mu-\lambda)^3}\right) - \left(\frac{-(\lambda+\mu)}{(\mu-\lambda)^3}\right)^2 \\
&= \frac{1}{(\mu-\lambda)^6}\left[4\lambda\mu - (\lambda^2 + 2\lambda\mu + \mu^2)\right] \\
&= \frac{1}{(\mu-\lambda)^6}\left[-(\lambda^2 - 2\lambda\mu + \mu^2)\right] \\
&= \frac{1}{(\mu-\lambda)^6}\left[-(\mu-\lambda)^2\right] \\
&= \frac{-1}{(\mu-\lambda)^4} < 0
\end{aligned}$$

Thus $L(\lambda, \mu)$ is not jointly convex and therefore $B(\lambda, \mu)$ is not jointly concave in (λ, μ).

It follows that the stationary point (1.17) identified by the first-order conditions does not necessarily yield the global maximum net benefit. To gain further insight, let us evaluate $B(\lambda, \mu)$ at this stationary point. Substituting the expressions from (1.17) into the formula for $B(\lambda, \mu)$ and simplifying, we obtain (after simplifying)

$$B(\lambda, \mu) = -\frac{h \cdot c}{r - c} < 0 = B(0, 0) \; .$$

So the proposed solution in fact yields a negative net benefit! It is therefore dominated by the point $(0, 0)$ (do nothing) and we know that we can do even better than that when $c < r$.

To see how much better, let us examine the problem from a slightly different perspective. Define the traffic intensity ρ (as usual) by $\rho := \lambda/\mu$ and rewrite the net benefit as a function of λ and ρ:

$$\tilde{B}(\lambda, \rho) := r \cdot \lambda - \frac{h \cdot \rho}{1 - \rho} - \frac{c \cdot \lambda}{\rho} \; .$$

Now fix a value of ρ such that

$$\frac{c}{r} < \rho < 1 \; .$$

Then we have

$$\tilde{B}(\lambda, \rho) = \lambda \cdot (r - \frac{c}{\rho}) - \frac{h \cdot \rho}{1 - \rho} \; .$$

The second term is constant and the first term is positive and can be made arbitrarily large by choosing λ sufficiently large. Thus $B(\lambda, \rho) \to \infty$ as $\lambda \to \infty$ and hence there is no finite optimal solution to the problem. Rather, one can obtain arbitrarily large net benefit by judiciously selecting large values of both λ and μ.

Of course these observations raise serious questions about the realism of our model. We shall address these questions later (in Chapter 5). In the meantime, we need to understand what went wrong with our approach based on finding a solution to the first-order optimality conditions.

As we saw, the net-benefit function in this model fails to be jointly concave because it contains a congestion-cost term that is proportional to $L(\lambda, \mu)$, the expected steady-state number of customers in the system, which fails to be jointly convex. This congestion-cost term can be written as

$$h \cdot L(\lambda, \mu) = \lambda(h \cdot W(\lambda, \mu)) \; ,$$

where $W(\lambda, \mu)$ is the expected steady-state waiting of a customer in the system. In other words, we have a congestion cost per unit time that takes the form

$$\text{(no. customers arriving per unit time)} \times \text{(congestion cost per customer)} \; .$$

While the congestion cost per customer (in this case, $h/(\mu - \lambda)$) is jointly convex, the result of multiplying by λ is to destroy this joint convexity.

As we shall see in later chapters, this type of congestion cost and its associated non-joint-convexity are not an anomaly but in fact are typical in queueing optimization models. As a result one must be very careful when applying classical economic analysis based on first-order optimality equations. It is not enough to simply assume that the values of the parameters are such that there exists a finite optimal solution in the interior of the feasible region, which then must satisfy the first-order conditions (because they are necessary for an interior maximum). We have seen in the present example that there may be no such interior optimal solution, no matter what the parameter values are. Moreover, there may be an easily identified solution to the first-order conditions which one is tempted to identify as optimal but which may in fact be far from optimal.

The literature contains a surprising number of examples in which these kinds of mistakes have been made.

1.4 Optimal Arrival Rates for a Two-Class System

Now suppose we have an *M/M/1* queue in which there are two classes of customers. The service rate μ is fixed but the arrival rates of the two classes (denoted λ_1 and λ_2) are decision variables. Customers are served in order of arrival, regardless of class, so that the expected steady-state waiting time in the system is the same for both classes and is a function, $W(\lambda)$, of the total arrival rate, $\lambda := \lambda_1 + \lambda_2$. Recall that in the *M/M/1* case $W(\lambda)$ is given by

$$W(\lambda) = \frac{1}{\mu - \lambda} \ , \ \lambda < \mu \ ; \ W(\lambda) = \infty \ , \ \lambda \geq \mu \ . \tag{1.18}$$

We shall assume a reward/cost model like that of Section 1.2, but with class-dependent rewards and waiting cost rates. Specifically, there is a reward r_i per entering customer of class i, and a waiting cost h_i per unit time per customer of class i in the system. The objective is to maximize the steady-state expected net benefit per unit time:

$$\begin{aligned} \max_{\{\lambda,\lambda_1,\lambda_2\}} \quad & B(\lambda_1,\lambda_2) = r_1\lambda_1 + r_2\lambda_2 - (\lambda_1 h_1 + \lambda_2 h_2)W(\lambda) \\ \text{s.t.} \quad & \lambda_1 + \lambda_2 = \lambda \\ & \lambda_1 \geq 0 \ , \ \lambda_2 \geq 0 \end{aligned}$$

As in the single-class model considered in Section 1.2, if all rewards and costs accrue to the customers, a solution $(\lambda_1^s, \lambda_2^s)$ to this optimization problem will be *socially optimal*, in the sense of maximizing the aggregate net benefit accruing to the collective of all customers. Moreover, if potential customers of class i arrive according to a Poisson process with mean rate $\Lambda_i \geq \mu$, then a socially optimal allocation can be implemented by admitting each class-i arrival with probability $a_i^s = \lambda_i^s/\Lambda_i$.

The following Karush-Kuhn-Tucker (*KKT*) first-order conditions are *necessary* for $(\lambda_1, \lambda_2, \lambda)$ to be optimal for this problem (see, e.g., Bazaraa et

al. [16]):

$$r_i = h_i W(\lambda) + \delta \text{ and } \lambda_i > 0 \tag{1.19}$$

$$\text{or } r_i \leq h_i W(\lambda) + \delta \text{ and } \lambda_i = 0 \tag{1.20}$$

for $i = 1, 2$, and

$$\lambda = \lambda_1 + \lambda_2 \ , \tag{1.21}$$

$$\delta = (\lambda_1 h_1 + \lambda_2 h_2) W'(\lambda) \ . \tag{1.22}$$

Now consider this system from the perspective of individual optimization. Suppose a fixed, arbitrary toll, δ, is charged to each entering customer. Each customer of class i takes $W(\lambda)$ as given and chooses the probability a_i of joining to maximize

$$a_i \cdot (r_i - h_i W(\lambda) - \delta) + (1 - a_i) \cdot 0 \ , \ a_i \in [0, 1] \ .$$

In other words, a class-i customer who joins receives the net benefit, $r_i - h_i W(\lambda)$, minus the toll, δ, paid for the use of the facility. A customer who balks receives (pays) nothing. Then it is easy to see that arrival rates, $\lambda_i = a_i \cdot \Lambda_i$, that satisfy equations (1.19) and (1.20) will be individually optimal for the customers of both classes. Moreover, for the given toll δ, a solution to (1.19), (1.20), and (1.21) is a Nash equilibrium.

As expected, equation (1.22) reveals that the socially optimal toll is just the *external effect*, defined (as usual) as the marginal increase in the total delay cost incurred as a result of a marginal increase in the flow, λ. By charging this socially optimal toll, the system operator can induce individually optimizing customers to behave in a socially optimal way, thereby making the Nash-equilibrium allocation coincide with the socially optimal allocation $(\lambda_1^s, \lambda_2^s, \lambda^s)$ (cf. Section 1.2).

1.4.1 Solutions to the Optimality Conditions: the M/M/1 *Case*

Let us now examine the properties of the solution(s) to the *KKT* conditions, using the explicit expression (1.18) for $W(\lambda)$ for an *M/M/1* system. The problem of finding a socially optimal allocation of flows takes the form

$$\begin{aligned} \max_{\{\lambda_1, \lambda_2\}} \quad & r_1 \lambda_1 - \frac{h_1 \lambda_1}{\mu - \lambda_1 - \lambda_2} + r_2 \lambda_2 - \frac{h_2 \lambda_2}{\mu - \lambda_1 - \lambda_2} \\ \text{s.t.} \quad & \lambda_1 + \lambda_2 < \mu \\ & \lambda_1 \geq 0 \ , \ \lambda_2 \geq 0 \end{aligned}$$

Without loss of generality, we may assume that $\mu = 1$. (Equivalently, measure flows in units of fraction of the service rate μ.) Let $a := r_1/h_1$, $b := r_2/h_2$, $c := h_1/h_2$. Then an equivalent form for the above problem is

$$\begin{aligned} \max_{\{\lambda_1, \lambda_2\}} \quad & c \left(a\lambda_1 - \frac{\lambda_1}{1 - \lambda_1 - \lambda_2} \right) + b\lambda_2 - \frac{\lambda_2}{1 - \lambda_1 - \lambda_2} \\ \text{s.t.} \quad & \lambda_1 + \lambda_2 < 1 \end{aligned} \tag{1.23}$$

$$\lambda_1 \geq 0 \ , \ \lambda_2 \geq 0$$

For an interior optimal solution, equation (1.19) must be satisfied for $i = 1, 2$. The unique solution to these equations is given by

$$\begin{aligned} \tilde{\lambda}_1 &= \frac{b(c-1)}{(ca-b)^2} - \frac{1}{c-1} \\ \tilde{\lambda}_2 &= \frac{c}{c-1} - \frac{ca(c-1)}{(ca-b)^2} \end{aligned}$$

It can be shown that this pair $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ is an interior point ($\tilde{\lambda}_1 > 0$, $\tilde{\lambda}_2 > 0$, $\tilde{\lambda}_1 + \tilde{\lambda}_2 < 1$) if the parameters satisfy the following conditions:

$$\begin{aligned} & b > a > 1 \ ; \\ & c > \frac{b-1}{a-1} \ ; \\ & a < \frac{(ca-b)^2}{(c-1)^2} < b \ . \end{aligned}$$

So, for an *M/M/1* system in which the parameters satisfy these conditions, we have established that the first-order optimality conditions have a unique interior-point solution. This result tempts us to conclude that this solution is indeed optimal. But the model of Section 1.3, in which the unique interior-point solution to the optimality conditions turned out to be nonoptimal, should serve as a warning to proceed more cautiously. The question remains whether there are other, non-interior-point solutions to the *KKT* conditions and whether one of these could yield a higher value of the objective function. Put another way: are the *KKT* conditions sufficient as well as necessary for an optimal solution to our problem?

1.4.2 Are the KKT *Conditions Sufficient?*

To answer this question, let us return to the problem in its original form. The objective function takes the following form (after substituting for λ from the equality constraint),

$$B(\lambda_1, \lambda_2) = r_1\lambda_1 + r_2\lambda_2 - f(\lambda_1, \lambda_2) \ ,$$

where $f(\lambda_1, \lambda_2) := (\lambda_1 h_1 + \lambda_2 h_2) W(\lambda_1 + \lambda_2)$. That is, $f(\lambda_1, \lambda_2)$ is the total delay cost per unit time expressed as a function of λ_1 and λ_2. The *KKT* conditions will be sufficient for social optimality if $B(\lambda_1, \lambda_2)$ is jointly concave in (λ_1, λ_2), which is true if and only if $f(\lambda_1, \lambda_2)$ is jointly convex in (λ_1, λ_2). It is easily verified that $f(\lambda_1, \lambda_2)$ is convex in λ_1 and convex in λ_2. To check for joint convexity, we evaluate

$$\Delta := \left(\frac{\partial^2 f}{\partial \lambda_1^2}\right)\left(\frac{\partial^2 f}{\partial \lambda_2^2}\right) - \left(\frac{\partial^2 f}{\partial \lambda_1 \partial \lambda_2}\right)^2$$

and find that $\Delta = -((h_1 - h_2)W'(\lambda_1 + \lambda_2))^2$, which is strictly negative unless $h_1 = h_2$, that is, unless the customer classes are homogeneous with respect to their sensitivity to delay. Thus $f(\lambda_1, \lambda_2)$ is *not* in general a jointly convex function of λ_1 and λ_2. Indeed, the conditions for joint convexity fail at *every* point in the feasible region if the customer classes are heterogeneous, that is, if $h_1 \neq h_2$. It follows that $B(\lambda_1, \lambda_2)$ fails to be jointly concave unless $h_1 = h_2$.

Remark 1 Note that we did not use the specific functional form (1.18) of $W(\lambda)$ in our demonstration of the nonconvexity of $f(\lambda_1, \lambda_2)$. The only properties that we used were that the delay $W(\lambda)$ for each customer is an increasing, convex, and differentiable function of the sum of the flows, and that the delay cost per unit time for each class i is the product of the flow, λ_i, and the delay cost per customer, $h_i W(\lambda)$. All these properties are weak and hold for many queueing models, not just for the *M/M/1* case. As we shall see in Chapters 4 and 5, nonconvexity is a widely encountered phenomenon in models for the design of queues with more than one decision variable.

The nonconcavity of the objective function, $B(\lambda_1, \lambda_2)$, leads one to suspect that the first-order *KKT* conditions, (1.19)–(1.22), may not be sufficient for an optimal allocation. In particular, an interior-point solution to these conditions – such as the one found in the previous subsection – might not be optimal. Let us now examine that question. First observe that such a solution must lie on the line $\lambda_1 + \lambda_2 = \lambda$, where λ satisfies

$$r_1 - h_1 W(\lambda) = r_2 - h_2 W(\lambda) \; . \tag{1.24}$$

Along this line both the total flow λ and the net benefit, $B(\lambda_1, \lambda_2)$, are constant: $B(\lambda_1, \lambda_2) = B$, say. In particular, the two extreme points on this line, namely, $(\lambda, 0)$, and $(0, \lambda)$, share this net benefit; that is,

$$B(\lambda, 0) = B(0, \lambda) = B \; .$$

But

$$\begin{aligned} B(\lambda, 0) &\leq B(\lambda_1^*, 0) \; , \\ B(0, \lambda) &\leq B(0, \lambda_2^*) \; , \end{aligned}$$

where λ_i^* is the optimal flow allocation to class i when only that class receives positive flow $(i = 1, 2)$.

Thus we see that any interior solution to the first-order *KKT* conditions is dominated by *both* the optimal single-class allocations. In other words, the system achieves at least as great a net benefit by allocating all flow to a single class, *regardless of which class*, than by using an interior allocation satisfying the first-order conditions!

Our next observation has to do with external effects, congestion tolls, and equilibrium properties. First note that charging each user a toll δ (per unit of flow) equal to the external effect, that is,

$$\delta = (\lambda_1 h_1 + \lambda_2 h_2) W'(\lambda_1 + \lambda_2) \; ,$$

makes $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ a Nash equilibrium for individually optimizing customers: no customer of either class has an incentive to deviate from this allocation, assuming that all other customers make no change. Thus, we see that, even by charging the "correct" toll (namely, a toll equal to the external effect), we cannot be certain that the customers will be directed to a socially optimal flow allocation. Rather, the resulting allocation, even though it is a Nash equilibrium, may be dominated by both of the optimal single-class allocations.

Thus we have a dramatic example of the pitfalls of marginal-cost pricing (that is, pricing based on first-order optimality conditions) when the customer classes are heterogeneous in their sensitivities to congestion.

As an example, let us return to the *M/M/1* example of Section 1.4.1. Let $a = 4$, $b = 9$, and $c = 4$. In this case, the solution to the first-order conditions is

$$\tilde{\lambda}_1 = 0.218 \; ; \; \tilde{\lambda}_2 = 0.354 \; .$$

The optimal single-user flow allocations are $\lambda_1^s = 0.500$ and $\lambda_2^s = 0.667$. The objective function values of these three flow allocations are:

$$\begin{aligned} B(\tilde{\lambda}_1, \tilde{\lambda}_2) &= 3.81 \\ B(\lambda_1^s, 0) &= 4.00 \\ B(0, \lambda_2^s) &= 4.00 \end{aligned}$$

Thus we have an illustration of the general result derived above: the interior-point equilibrium flow allocation is dominated by both optimal single-user allocations.

For this example, Figure 1.4 and Figure 1.5 show, respectively, a contour plot and graph of the response surface of the net benefit function, $B(\lambda_1, \lambda_2)$.

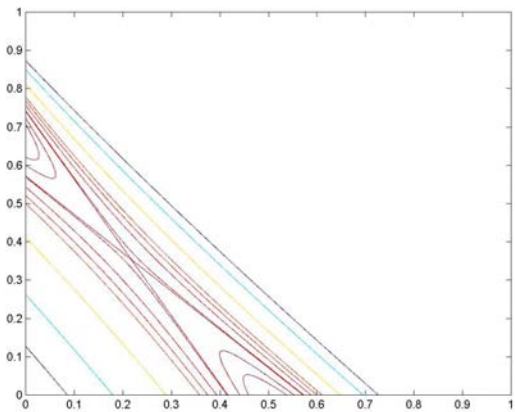


Figure 1.4 *Net Benefit: Contour Plot*

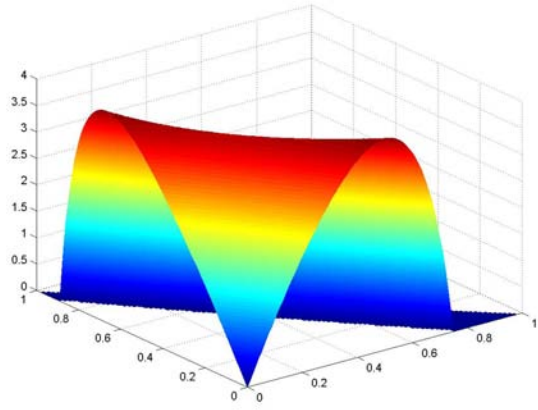


Figure 1.5 *Net Benefit: Response Surface*

1.5 Optimal Arrival Rates for Parallel Queues

Now let us consider n independent *M/M/1* queues, with service rates μ_j and arrival rates λ_j, $j = 1, \ldots, n$. Suppose that the μ_j are fixed and that the λ_j are design variables. Our objective is to minimize the steady-state expected number of customers in the system, subject to a constraint that the total arrival rate should equal a fixed value, λ. Thus the problem takes the form

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} \frac{\lambda_j}{\mu_j - \lambda_j} \\ \text{s.t.} \quad & \sum_{j=1}^{n} \lambda_j = \lambda \\ & 0 \le \lambda_j < \mu_j \ , \ j = 1, \ldots, n \ . \end{aligned} \tag{1.25}$$

We can interpret this problem as follows. Suppose customers arrive to the system according to a Poisson process with mean arrival rate λ. We must decide how to split this arrival process among n parallel exponential servers, each with its own queue. The splitting is to be done probabilistically, independently of the state and past history of the system. That is, each arriving customer is sent to queue j with probability $a_j = \lambda_j/\lambda$, so that the arrival process to queue j is Poisson with mean arrival rate λ_j.

We shall use a Lagrange multiplier to eliminate the constraint on the total arrival rate. The Lagrangean problem is:

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} \frac{\lambda_j}{\mu_j - \lambda_j} - \alpha \sum_{j=1}^{n} \lambda_j \\ \text{s.t.} \quad & 0 \le \lambda_j < \mu_j \ , \ j = 1, \ldots, n \ . \end{aligned} \tag{1.26}$$

The solution is parameterized by α, which can be interpreted as the imputed reward per unit time per unit of arrival rate. Problem (1.26) is separable, so we can minimize the objective function separately for each facility. For facility j, the problem takes the form of the single-facility arrival-rate-optimization problem of Section 1.2, with $r = \alpha$, $h = 1$. The solution is:

$$\lambda_j = \lambda_j^s(\alpha) := (\mu_j - \sqrt{\mu_j/\alpha})^+ \ , \ j = 1, \ldots, n \ . \tag{1.27}$$

This solution will be optimal for the original problem if α is chosen so that $\sum_{j=1}^n \lambda_j^s(\alpha) = \lambda$.

Thus an optimal allocation satisfies the following conditions ($j = 1, \ldots, n$):

$$L_j'(\lambda_j) = \frac{\mu_j}{(\mu_j - \lambda_j)^2} = \alpha \ , \text{ if } \lambda_j > 0 \ , \tag{1.28}$$

$$L_j'(\lambda_j) = \frac{1}{\mu_j} \geq \alpha \ , \text{ if } \lambda_j = 0 \ , \tag{1.29}$$

for some α such that $\sum_{j=1}^n \lambda_j = \lambda$.

These results can be used to solve the original problem (1.25) graphically. First, plot each $\lambda_j^s(\alpha)$ as a function of α, as shown in Figure 1.6. Define

$$\lambda^s(\alpha) := \sum_{j=1}^n \lambda_j^s(\alpha) \ ,$$

so that $\lambda^s(\alpha)$ is the total arrival rate in an optimal solution of problem (1.26) corresponding to Lagrange multiplier α. We can now find the optimal solution to the original problem for a particular value of λ by drawing a horizontal line from the vertical axis at level λ and finding its intersection with the graph of $\lambda^s(\alpha)$, then drawing a vertical line to the α axis. Where this line intersects the graph of $\lambda_j^s(\alpha)$, we obtain $\lambda_j^s = \lambda_j^s(\lambda)$, the optimal value of λ_j for the original problem with total arrival rate λ.

We can derive an explicit solution for the λ_j^s in terms of the parameter λ (denoted $\lambda_j^s(\lambda)$, $j = 1, \ldots, n$) in the following way. First, order the μ_j so that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. From (1.27) it can be seen that $\lambda^s(\alpha)$ is a continuous, strictly increasing function of α, for $\alpha \geq \mu_1^{-1}$. In this range, therefore, $\lambda^s(\alpha)$ has an inverse, which we denote by $\alpha(\lambda)$. We solve for $\alpha(\lambda)$ separately over the intervals induced by $\mu_1^{-1} \leq \alpha \leq \mu_2^{-1}$, $\mu_2^{-1} \leq \alpha \leq \mu_3^{-1}$,.... In particular, for $\mu_1^{-1} \leq \alpha \leq \mu_2^{-1}$,

$$\begin{aligned} \lambda_1^s(\alpha) &= \mu_1 - \sqrt{\mu_1/\alpha} \ , \\ \lambda_j^s(\alpha) &= 0 \ , \ j = 2, \ldots, n \ . \end{aligned}$$

Thus $\lambda_1^s(\alpha) = \lambda$ in this range, so that

$$\sqrt{\frac{1}{\alpha}} = \frac{\mu_1 - \lambda}{\sqrt{\mu_1}} \ , \tag{1.30}$$

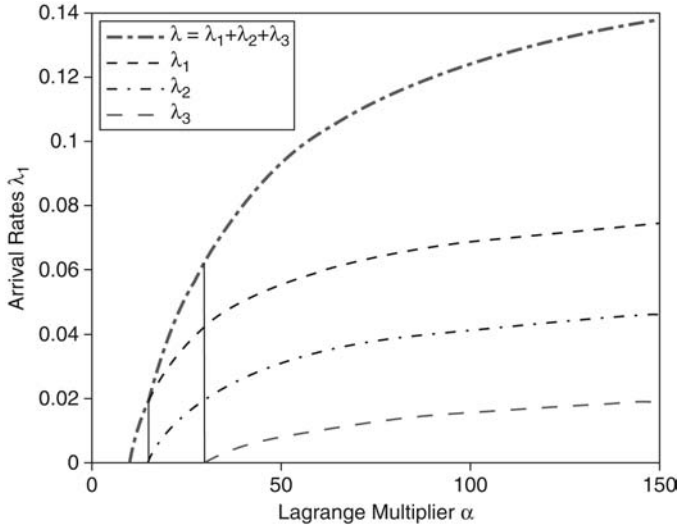


Figure 1.6 *Arrival Control to Parallel Queues: Parametric Socially Optimal Solution*

and hence

$$\lambda_1^s(\lambda) = \mu_1 - \frac{\sqrt{\mu_1}}{\sqrt{\mu_1}}(\mu_1 - \lambda) = \lambda \ .$$

But it follows from (1.30) that $\mu_1^{-1} \leq \alpha \leq \mu_2^{-1}$ if and only if $0 \leq \lambda \leq \mu_1 - \sqrt{\mu_1\mu_2}$.

Summarizing, for $r_1 := 0 \leq \lambda \leq r_2 := \mu_1 - \sqrt{\mu_1\mu_2}$, we have

$$\begin{aligned} \lambda_1^s(\lambda) &= \lambda \ , \\ \lambda_j^s(\lambda) &= 0 \ , \ j = 2, \ldots, n \ . \end{aligned}$$

Continuing this argument, we can deduce the general form of the solution for $\lambda_j^s(\lambda)$, $j = 1, \ldots, n$. In general, define $r_k := \sum_{i=1}^{k}(\mu_i - \sqrt{\mu_i\mu_k})$, $k = 1, \ldots, n$, $r_{n+1} := \sum_{i=1}^{n} \mu_i$. Then, for $k = 1, \ldots, n$, if $r_k \leq \lambda \leq r_{k+1}$,

$$\begin{aligned} \lambda_j^s(\lambda) &= \mu_j - \left(\frac{\sqrt{\mu_j}}{\sum_{i=1}^{k}\sqrt{\mu_i}}\right)\left(\sum_{i=1}^{k}\mu_i - \lambda\right) \ , \ j = 1, \ldots, k \ , \\ &= 0 \ , \ j = k+1, \ldots, n \ . \end{aligned}$$

Note that each λ_j^s is piecewise linear in λ. Figure 1.7 gives a typical illustration. Note that, once $\lambda_j^s(\lambda)$ is positive, its rate of increase is nonincreasing in λ (thus $\lambda_j^s(\lambda)$ is concave in $\lambda \geq r_j$) and that the rates of increase of the $\lambda_j^s(\lambda)$ for fixed λ are nondecreasing in j.

Individually Optimal Allocation

The allocation described above assumes that the allocation of total "demand," λ, to the various facilities is made in accordance with the system-wide

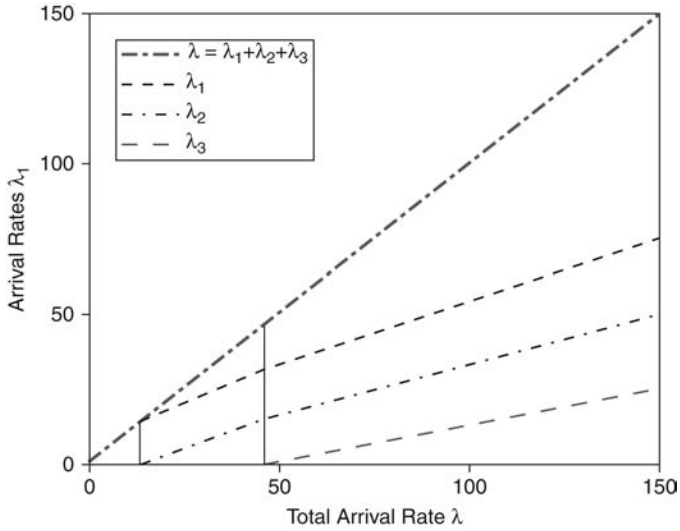


Figure 1.7 *Arrival Control to Parallel Queues: Explicit Socially Optimal Solution*

objective of minimizing the total rate of waiting per unit time: $\sum_{j=1}^{n} L_j(\lambda_j) = \sum_{j=1}^{n} \lambda_j/(\mu_j - \lambda_j)$. An equivalent way of viewing this problem is to visualize each arriving customer having a probability, $a_j = \lambda_j/\lambda$, of joining facility j, $j = 1, \ldots, n$, where the $a_j's$ are to be chosen (by an omnipotent system designer) to minimize the steady-state expected waiting time of an arbitrary customer:

$$\sum_{j=1}^{n} \left(\frac{\lambda_j}{\lambda}\right)\left(\frac{1}{\mu_j - \lambda_j}\right) = \frac{1}{\lambda}\sum_{j=1}^{n} L_j(\lambda_j)$$

Now let us consider an allocation $(\lambda_1, \ldots, \lambda_n)$ (equivalently, a set of joining probabilities $(a_1, ..., a_n)$) from the point of view of an individual customer who wishes to minimize his expected waiting time. Under the allocation in question, an arriving customer chooses facility j with probability $a_j = \lambda_j/\lambda$; conditional on joining facility j, the expected waiting time is $(\mu_j - \lambda_j)^{-1}$. (As is always the case in design models, we assume that the fixed mean service rates μ_j and the arrival rates λ_j associated with the given allocation are known and the system is in steady state, but the exact number of customers at each facility cannot be observed.) The customer's unconditional expected waiting time is therefore $\sum_{j=1}^{n} a_j(\mu_j - \lambda_j)^{-1}$. As usual we call an allocation $(\lambda_1, \ldots, \lambda_n)$ (or a set of joining probabilities $(a_1, \ldots, a_n)$) *individually optimal* if no customer, acting in its own interest, has an incentive to deviate unilaterally from the allocation. This will be the case if and only if $(\mu_j - \lambda_j)^{-1} = (\mu_k - \lambda_k)^{-1}$ for all j, k such that $\lambda_j > 0$ and $\lambda_k > 0$, and $(\mu_j - \lambda_j)^{-1} \leq \mu_k^{-1}$, if $\lambda_j > 0$ and $\lambda_k = 0$. Otherwise, e.g., if $(\mu_j - \lambda_j)^{-1} > (\mu_k - \lambda_k)^{-1}$ for some j, k such that $\lambda_j > 0$, an arriving customer could strictly reduce its expected waiting time

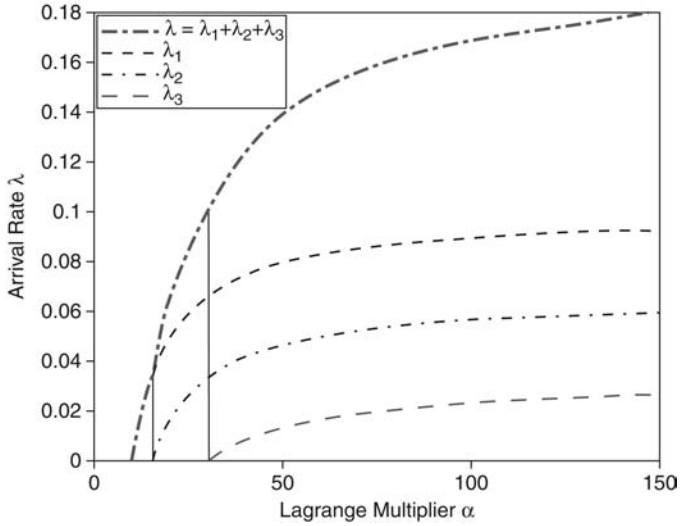


Figure 1.8 *Arrival Control to Parallel Queues: Parametric Individually Optimal Solution*

by joining facility j with probability $a'_j := 0$ and facility k with probability $a'_k := a_j + a_k$, rather than $a_j = \lambda_j/\lambda$ and $a_k = \lambda_k/\lambda$, respectively.

In other words, an individually optimal allocation satisfies the following conditions, for $j = 1, \ldots, n$:

$$W_j(\lambda_j) = \frac{1}{\mu_j - \lambda_j} = \alpha \text{ , if } \lambda_j > 0 \text{ ;} \tag{1.31}$$

$$W_j(\lambda_j) = \frac{1}{\mu_j} \geq \alpha \text{ , if } \lambda_j = 0 \text{ ;} \tag{1.32}$$

for some $\alpha > 0$ such that $\sum_{j=1}^{n} \lambda_j = \lambda$.

We would like to compare such an allocation, denoted $\lambda_j^e(\alpha)$, or $\lambda_j^e(\lambda)$, to the *socially optimal* allocation, $\lambda_j^s(\alpha)$, or $\lambda_j^s(\lambda)$. First observe from (1.31) and (1.28) that an individually optimal allocation equates *average costs*, $1/(\mu_j - \lambda_j)$ (internal effects), whereas a socially optimal allocation equates *marginal costs*, $\mu_j/(\mu_j - \lambda_j)^2 = 1/(\mu_j - \lambda_j) + \lambda_j/(\mu_j - \lambda_j)^2$ (internal plus external effects), at all open facilities j.

In terms of α, the individually optimal allocation can be written as

$$\lambda_j^e(\alpha) = (\mu_j - 1/\alpha)^+ \text{ , } j = 1, \ldots, n \text{ .}$$

Figure 1.8 illustrates the behavior of $\lambda_j^e(\alpha)$, assuming $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Now α must be chosen so that $\sum_{j=1}^{n} \lambda_j^e(\alpha) = \lambda$, in order to find $\lambda_j^e(\lambda)$, $j = 1, \ldots, n$. This can be done in the same way as for socially optimal allocations. (The details are left to the reader.) In general, define $s_k := \sum_{i=1}^{k} (\mu_i - \mu_k)$, $k = 1, \ldots, n$, $s_{n+1} := \sum_{i=1}^{n} \mu_i$. Then the individually optimal allocation is as

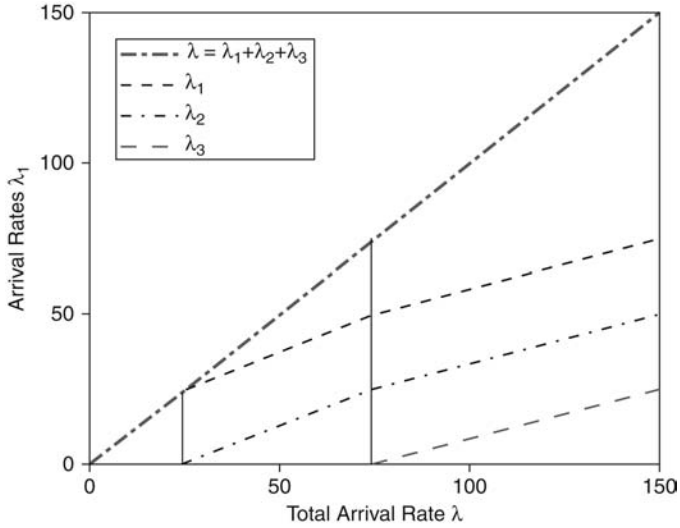


Figure 1.9 *Arrival Control to Parallel Queues: Explicit Individually Optimal Solution*

follows: for $k = 1, \ldots, n$, if $s_k \leq \lambda \leq s_{k+1}$, then

$$\begin{aligned} \lambda_j^e(\lambda) &= \mu_j - [\sum_{i=1}^{k} \mu_i - \lambda]/k \ , \ j = 1, \ldots, k \ , \\ &= 0 \ , \ j = k+1, \ldots, n \ . \end{aligned}$$

Figure 1.9 illustrates the behavior of the individually optimal facility arrival rates as a function of the total arrival rate. Note that the positive $\lambda_j^e(\lambda)$ are piecewise linear in λ, with nonincreasing slope. The slopes of all positive $\lambda_j^e(\lambda)$ are equal in this case.

In Figure 1.10, the individually optimal allocation is superimposed on the socially optimal allocation, for purposes of comparison. As a general observation, we can say that the individually optimal allocation assigns more (fewer) customers to faster (slower) servers than the socially optimal allocation. More specifically, for the example in Figure 1.10, the individually optimal allocation always assigns more arrivals to facility 1, the fastest one, and fewer arrivals to facility 3, the slowest one, than the socially optimal allocation does. As λ increases, facility 2 first receives fewer, then more, arrivals in the individually optimal than in the socially optimal allocation. Thus, facility 2 plays the role of a "slower" server in light traffic and a "faster" server in heavy traffic.

1.6 Endnotes

Over the past forty years, there have been a number of survey papers and books that discuss optimal control of queues, including Sobel [181], Stid-