Pricing and Admission Control for Shared Computer Services Using the Token Bucket Mechanism

by

Opher Baron

B.Sc., The Technion Israel Institute of Technology (1998)
MBA, The Technion Israel Institute of Technology (1998)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2003

Signature of Author .....
Sloan School of Management
8 May 2003

Certified by .....
Gabriel R. Bitran
Deputy Dean; Nippon Telephone and Telegraph Professor of Management
Research Head

Accepted by .....
Birger Wernerfelt
Professor of Management Science; Chair PhD Program

Pricing and Admission Control for Shared Computer Services Using the Token Bucket Mechanism

by

Opher Baron

Submitted to the Sloan School of Management
on 8 May 2003, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Management

Abstract

This dissertation presents and analyzes token-bucket pricing schemes for shared resources. This research is motivated by the computer services industry, where services are provided mostly on a dedicated basis. However, leading computer companies such as HP and IBM forecast that external service providers will share resources between customers, in order to realize economies of scale. Two of the challenges faced by providers and consumers of shared services are admission control and pricing. In order to allow sellers to guarantee service levels, we recommend that pricing schemes for shared resources include admission controls. The implementation of such schemes requires understanding of buyers' and sellers' actions and a characterization of the admission control.

This dissertation reviews the computer services supply-chain and proposes a five-step procedure for analyzing the pricing of shared services. Then it extends the usage of token-bucket and token-bucket-with-rate-control admission controls to pricing schemes. We show that for the token-bucket (token-bucket-with-rate-control) mechanism the bucket level behaves as a two- (one-) sided regulated random walk. Thus, the performance analysis (loss sales or backlog) is identical to the analysis of threshold crossing probabilities of regulated random walks. This dissertation's main contribution is an upper bound on the probability of a two-sided regulated random walk being on its "rare" boundary.

Using the bounds developed, we solve constrained or relaxed versions of the buyer's problem. For the token-bucket-with-rate-control pricing scheme and exponential demand the buyer's problem can be solved in closed form. Moreover, numerical experiments show that the approximate solutions for the normal demand case are within 1% of optimal. Similar results hold for the token-bucket mechanism.

Finally, we characterize the output stream of these admission controls (the jumps of one- or two-sided regulated random walks). We use a Brownian motion approximation for the bucket level process, but still consider the actual demand and arrival processes. Moreover, we enhance the performance of this approach by relating fill rates with the percentage of periods with losses. Numerical results show that in both mechanisms, when demand is exponential or normal, the approximated first two moments of the output stream are, typically, within the 99% confidence intervals.

Research Head: Gabriel R. Bitran
Title: Deputy Dean; Nippon Telephone and Telegraph Professor of Management

Thesis Supervisor: Dirk Beyer
Title: Project Manager, HP Labs

Thesis Supervisor: Stephen C. Graves
Title: Abraham J. Siegel Professor of Management

Thesis Supervisor: Yashan Wang
Title: Assistant Professor Management Science

Contents

List of Figures

List of Tables

Acknowledgments

First, there is my advisor and mentor Professor Gabriel Bitran, whom I would like to thank for being the person he is. I hope to continue learning from his spirit on both the personal and professional levels. His support throughout these years cannot be measured. I am grateful to Professor Yashan Wang for teaching me rigorous thinking, as well as for introducing me to the subject of large deviation. I owe thanks also to Dr. Dirk Beyer of HP Labs, whose comments regarding each aspect of this work were extremely valuable. I hope that in the future I will have similar qualities to him, so I will be able to model and analyze real-world problems and understand the limits of such models and analysis. Thanks are also due to Professor Steve Graves for his time and guidance throughout my years at MIT. I owe additional thanks to all these people for their instructive comments on this work.

I am grateful to Jörn Altman, Shailendra Jain, Sharad Singhal, and Alex Zhang of HP labs, for the continuos feedback and insightful remarks they gave me all along this research project. I thank Professors Avishai Mandelbaum, David Simchi-Levi, and Lawrence Wein. The energy and research enthusiasm of these three are an inspiration to me.

I thank my mother Bracha Baron who throughout my life has given me the motivation to excel in what I am doing despite the inevitable difficulties. I thank Itzchak Klein who taught me the importance of being patient in turbulent times. I thank my father Zvika and his wife Hana Baron for always being willing to listen and give advice. I thank my sister Meyrav Wolfson and her husband Mordechai. Meyrav always understands my feelings, even when they are wrong; I am proud to call her my friend. I thank my grandmother Rita Baron for teaching me the real joy of life. I am also thankful to Dror Baron, to Niv Kochav, who is like a brother, and to Ester and Yehezkel Yehezkel for being there for me. I thank Ela, Bruno, and Lior Fainaru who, along with all the people above, teach me every day to keep my priorities right.

I thank my friends at MIT, those who have gone through a similar path in the last years, people without whom I would not have been able to complete the studies here: Hasan A., Damian B., Felipe C., Amit D., Juan-Carlos F., Paulo G., Jay J., Paulo O., Oded R., and Hazhir R. I am happy that I have had the opportunity to meet them, and I hope that in the future we will keep being friends. I thank Anna Piccolo, Marguerite Baty, and Sharon Cayley for helping me to find my way along the corridors of MIT and the English language.

I dedicate this work to Iris and Ethan. Without Iris's love, continuous support and encouragement this dissertation would never have been written. Iris and Ethan are the fountain of life for me.

Chapter 1

Shared Computer Services and the Challenges in Pricing Them

1.1 Introduction

Today, many businesses outsource their computing needs (web-hosting, record keeping, databases) to external service providers that usually install dedicated computer resources and charge for hardware, software, and support for these resources. Forecasts of leading computer companies such as IBM, HP and Sun Microsystems are that in the future, such service providers are likely to share resources between customers to realize economies of scale. In such a case, pricing computing services might need to be usage-based. In this chapter we present the pricing problem that faces Application Service Providers (ASPs) when they consider the provisioning of services based on shared computer resources. One of the characteristics of this problem is its focus on pricing for a relatively small number of customers, which allows for a fine market segmentation by provisioning a highly tailored service. Thus, pricing computer services differs from traditional pricing problems, which focus on pricing of standard "tangible" products to a large number of customers.

An additional factor that raises the interest in managing shared resources is the increased interest in grid computing, or a utility data center. Grid computing is defined (Gridcomputing, 2003) as "a type of parallel and distributed system that enables the sharing, selection, and aggregation of geographically distributed 'autonomous' resources dynamically at runtime depending

on their availability, capability, performance, cost and users' quality-of-service requirements." The notion of grid computing has been an active research area in leading universities for over a decade. In the last few years this notion has enjoyed the attention of leading computer companies (HP, IBM, and Sun Microsystems). For the interested reader many articles and white papers on the subject of grid computing can be found on the Internet by searching for the term "grid computing," at the sites of the above companies, at (Gridforum, 2003), (Sharcnet, 2001), (Gridcomputing, 2003), and at sites and papers referenced therein.

This dissertation presents the application of token bucket admission controls to pricing of shared resources (as explained in Chapter 2). The present chapter introduces the concept of provisioning shared computer services and the challenges in pricing them. Chapter 2 presents token bucket admission controls and their potential uses as pricing schemes for shared services. We consider the idea of using token bucket admission controls as pricing schemes as one of the main contributions of this work. Chapter 2 also outlines the research questions that are addressed in the later chapters, all of which focus on using token bucket pricing schemes for the pricing of a single shared resource. Chapter 3 analyses the performance of token bucket admission controls, where performance is considered as resources availability. One of the highlights of this dissertation, presented in this chapter, is a new bound developed for the loss probability of a two-sided regulated random walk. Chapter 4 uses the results developed in Chapter 3 to solve the problem faced by buyers of a service when it is priced using token bucket pricing schemes. The main tool used in order to find bounds on the solution to the buyer's optimization problem is a proof of convexity of a constrained version of the buyer's problem. Chapter 5 characterizes the output stream from the token bucket admission controls. The aggregation of this stream among many buyers is the effective demand faced by the seller, when using a token bucket pricing scheme. This chapter's results, which are based on a new attitude towards characterizing the output process from token bucket admission controls, are another major contribution of this work. In Chapter 6 we summarize this dissertation, and discuss additional problems that need to be solved in order to make the token buckets pricing schemes applicable.

This objective of this chapter is to facilitate our understanding of the pricing of shared computer resources problem. It provides background on the computer services industry by

describing the five layers that form the computer services supply chain: providers of computer hardware, communication hardware and software; providers of Internet Data Centers (IDC); computer services providers, such as Application Service Providers, Internet Service Providers (ISP), and others; companies and organizations that consume computer services; and the final customers who, for example, surf the web or run a tailored software products. We outline the services provided and list the computer resources required for providing these services.

After learning where the pricing problem of shared services is located in the computer services supply chain, it is important to understand the difference between provisioning of dedicated service and provisioning of shared service. Subsection 1.2.3 discusses the advantages of shared services, which are mainly the resource savings due to smoother demand, and its disadvantages, such as the need to assure privacy and security of data. Section 1.3 provides a five-step framework for analyzing pricing of shared resources. It includes a short survey of the literature and practice of pricing of computer services, and a qualitative comparison of existing pricing schemes.

Finally, we summarize this chapter, concluding that the pricing problem faced by providers of shared computer services is real and important, and that there is a place for a pricing scheme that is tailored towards pricing of shared resources.

1.2 Background on Computer Services

This section describes the computer services supply chain and points to the location of our pricing problem in this chain. In addition, it presents the relevant computer services and the resources required to provide them.

1.2.1 The Supply Chain of Computer Services

The computer service supply chain can be crudely partitioned into five main layers of parties involved, as shown in Figure 1-1:

- First layer – Infrastructure providers. This layer includes providers of infrastructure for communication and for computers. The communication infrastructure providers are the providers of bandwidth (cables) companies such as Enron, and Internet backbone

companies[1] such as AT&T and CompuServ. The providers of computer infrastructure are companies such as HP, IBM, and SAN Microsystems, which provide hardware; and Microsoft, Netscape, SAP, and I2, which provide software.

- Second layer – Internet data center providers. This layer includes companies such as Exodus and Qwest. These companies build Internet Data Centers (IDCs). They buy the hardware and software required in order to provide computer services, connect the IDCs to the Internet backbone, and provide the required infrastructure, such as the huge air-conditioned computer farms, and basic network infrastructure to connect their computers to one another.

- Third layer – Internet Service Providers (ISPs) or Application Service Providers (ASPs), such as Oracle Business On-line and Breakaway Solutions. These companies operate the computer farms that were built by the second-layer companies. They do all the required configuration (computers and network) and are responsible for installation and disconnection of service.

- Fourth layer – Companies and organizations that provide service to end customers (the fifth layer participant). This layer includes companies such as "Amazon.com" and organizations such as MIT. The services these companies use include services that are given to the end customers, database maintenance, and computer power that can be used to run dedicated software, such as logistics or payroll software.

- Fifth layer – The end users of the Internet and other computer resources. These participants in the chain typically use computer services for purchasing on-line, surfing the web, and the running of dedicated software.

As in any supply chain, such a crude partition is helpful for the conceptual understanding of the industry, but is not an accurate representation of the "real" computer resources supply chain. For example, it ignores the vertical integration of firms. A recent example of this is Exodus, which is bankrupt and now attempting to operate as an ASP instead of only an IDC provider.

[1] In (Austin, 2001), Robert Austin writes: "'Backbone Providers' own the very large data transmission lines

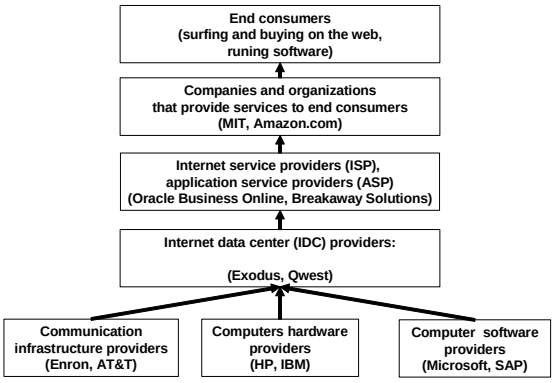


Figure 1-1: The Supply Chain of Computer Services

Looking at the above partition of the computer services supply chain, one sees that all members of layers one to four have pricing problems. We focus on the pricing problem faced by ASPs. For simplicity, we will call ASPs "sellers" and the companies purchasing their services "buyers."

In order to investigate the pricing problem sellers face, we should first understand the services they provide.

1.2.2 Computer Services

There are many different categories of computer service providers, which have won the name XSP for X service providers, e.g., DSP for database service provider, CSP for computing service provider, ASP for application service provider, etc. For the remainder of this chapter we will refer to all of these as ASPs.

A typical ASP has 20 to 50 customers in each IDC, and each one of these customers uses about 100 computers. It is expected that the size of IDC will increase and numbers of about 500 customers and 50,000 computers might be realistic. Such an increase will make the operation

via which large quantities of data are moved long distances."

of IDCs much more complicated.

The main services offered by ASPs are hosting services. We will concentrate on three hosting services: database hosting, web hosting, and application hosting. A common characteristic of these services is that they require a combination of three computer resources (storage, bandwidth, and processing). However, each hosting service requires different proportions of these resources.

Besides the hosting services there is a wide range of other services provided by ASPs. The following is a non-exhaustive list of such services: Internet Access - Providing connection to the Internet based on a dedicated or, more commonly, shared bandwidth. Load Balancing - Smoothing the data traffic across few servers. Complex Clustering - Dividing different servers to back and front groups and allowing for Data Replication and Data Mirroring. Hosting for developers - Providing ASP developers with the advanced servers and software required to allow running state-of-the-art web-based applications.

There are different methods of providing these services, and we give a possible taxonomy for these methods in Appendix A.1.

1.2.3 Advantages and Disadvantages of Shared Services

The main premise of shared resource services is to increase usage rates. Such an increase will allow ASPs to provide better service at lower costs. Yet, the danger of overloading resources in periods of peak demand should be considered. There are two main reasons that shared resources make possible an increase in usage rates: the integrality of computer resources and the aggregation of demand. We illustrate these reasons using a simple example.[2]

Assume seven identical customers require computer power equal to half a CPU seven days a week. Moreover, the first customer requires additional computer power of a full CPU on Monday; the second customer requires additional computer power of a full CPU on Tuesdays; and so on, until the seventh customer, who requires additional computer power of a full CPU on Sunday.

If each customer purchases a dedicated computer service, the total number of computers required is 7x2=14. But in a shared service there are opportunities to decrease the number of

[2]For simplicity, this example ignores the extra computer power required to manage a shared pool of CPUs

computers required. First, if the ASP considers the maximum number of computers required to satisfy customers peak demand at the same time, 1.5x7=10.5→ 11 computers will be required. These savings are due to the integrality of CPUs. The second, and more important, resource saving can be realized once one considers the aggregated demand of all customers. In our example, 0.5x7+1=4.5→ 5 computers will be enough to satisfy all customers' demand. This saving follows from the demand aggregation. The benefit of demand aggregation of different customers is a general property of addition of random variables and is a consequence of the law of large numbers. As long as different customers' demands are not fully correlated, the coefficient of variation (the ratio of demand's standard deviation to its mean) of their aggregated demand is decreasing. Thus, if we preclude perfect correlation, the ratio of average to peak demand increases as the ASP aggregates the demand of a larger number of customers.

To summarize, shared services help ASPs to increase the utilization of their computer resources, serving more customers using fewer resources (per customer). This, in turn, leads to a win-win situation in which ASPs can reduce their service cost and increase their provided service level at the same time. Note that the challenges faced by ASPs today are similar to the challenges faced by system administrators in the pre-personal-computer era, when time sharing of computers was necessary.

Despite the above advantages, provisioning of shared computer resources has some disadvantages that cannot be ignored. We mention four of them.

First, shared services require a good resource allocation mechanism to assure the best performance for the system as a whole. Unfortunately, due to the continuous and dynamic nature of the use of computer resources, this allocation needs to be done frequently. Therefore, the allocation of resources to different users is a complicated task. Moreover, the dynamic management of the resources pool requires a large investment in software that can manage this pool. This software itself will consume some computer resources, decreasing the resource savings from the shared services.

Second, shared services require technology that assures security and privacy for the different clients, despite their usage of the same resources pool. Historically, privacy and security of business data is one of the major reasons against outsourcing of computer services. Clearly, these issues cause an even larger concern if the outsourced service is shared. Thus, it is clear

that security and privacy must be guaranteed, or else this concern will prevent the provisioning of shared services.

Third, provisioning of shared service means that the service level provided for each user does not depend only on its own usage. Thus, users might face a lower service level than they expect or need. In the case of a dedicated resource, a low service level can be explained by the fact that the buyer needs to buy (or rent) more resources, but in the case of shared service, such an explanation is unacceptable. Therefore, shared service increases the responsibilities of the service providers. In order to allow ASPs to provide the required service levels, they need to operate an admission control, such that they can prevent one buyer from consuming all of the resources in a certain time, and thereby blocking the other buyers. Thus, the operation of admission control is an additional challenge faced by buyers and sellers of shared computer resources.

Finally, the pricing of shared resources is difficult. It makes sense to charge larger users more than smaller users, using some usage-based pricing mechanism. But, as will be shown in Chapter 1.3, there are many difficulties in implementing such pricing schemes. One of the difficulties is that most buyers are interested in knowing ahead of time how much they are going to be charged for their computer services. Moreover, when resources are shared, some of them might be idle but still require costly maintenance, a cost that should be shared in some manner among all buyers.

Despite these disadvantages, we believe that the huge cost savings resulting from demand aggregation are large enough to justify the adoption of shared services by ASPs.

1.3 Framework for Analyzing Pricing Schemes for Shared Resources

We consider a five-step framework for analyzing pricing schemes for shared resources. The first step is gaining familiarity with pricing schemes in the literature and in practice. The second step is learning the important attributes required of a pricing scheme for shared resources from both buyers' and sellers' points of view. The third step is comparing existing pricing schemes along these attributes. The fourth step is analyzing this comparison in order to recommend

good pricing schemes. The recommendation can be either to choose from existing pricing schemes or to suggest new ones. The fifth step is surveying practitioners' and theoreticians' views regarding the usefulness of any recommended pricing scheme. Clearly, after such a survey it might be necessary to repeat steps two to five.

This chapter presents a short version of the first four steps. It paves the way for this dissertation that focuses on the analysis of novel token bucket pricing schemes. This research has gone through several cycles of the above framework, but still cannot give a conclusive result about the best pricing scheme to use for shared services in general or specifically for shared computer services. However, it concludes that the token bucket pricing scheme might be a good pricing scheme for shared resources.

In this chapter, we first enrich our knowledge of pricing schemes by surveying the pricing in the literature and in practice in subsection 1.3.1. The only computer resource whose pricing is discussed in the literature is communication; thus, we present the *Smart Market* approach (MacKie-Mason and Varian, 1994a) as an example of congestion pricing. In smart market pricing, as with other congestion pricing, the main objective is to manage network congestion. In contrast to congestion pricing, the most common pricing in the computer services markets, as we have found from the advertisements of hosting companies on the web and a short series of interviews, is fairly simple: most service providers charge a fixed fee for different usage packages, but there are some innovators who use different kinds of usage-based pricing, mainly in the market for database products. The pricing for large customers might be a little more elaborate, but is still rather simple.

As a second step, we list the important attributes for a pricing scheme for shared computer resources, from both the buyers' and the sellers' points of view. This list is based on the author's interpretation of a short field study and is by no means a definitive list of these important attributes.

For the third step we compare four pricing schemes, according to the attributes listed in subsection 1.3.3. This is only a qualitative comparison, and so it is prone to favor the author's point of view. The fourth step, analyzing this comparison, concludes that there is a place for a new pricing scheme tailored toward pricing of shared computer resources, and that it will be beneficial to include admission control in such a pricing scheme.

1.3.1 Survey of Pricing Schemes

Pricing Computer Resources in the Literature

This subsection summarizes some of the pricing schemes suggested in the literature, and the next subsection reviews pricing in practice. An important review of the main ideas in pricing in computer networks is given in (Shenkar et al., 1995).

The vast majority of the literature we encounter deals with the pricing of bandwidth usage. The literature expects an efficient network pricing scheme to achieve three goals. The first is congestion control, which can be achieved by increasing the price of bandwidth when system congestion is high. The second target is social welfare, which means that users with a higher assessment for the service will get service when congestion occurs, whereas users with a lower assessment will not get service in such cases. This second goal might be replaced by maximizing the service provider's revenue (or profit). Finally, it is expected that sellers will recover the huge fixed cost of maintaining the network and that sellers will be encouraged to increase the amount of available bandwidth, as necessary.

Most of the methods we examine suggest some kind of dynamic pricing scheme, which looks unrealistic since users cannot be expected to monitor their usage and preferences continually. Thus, researchers (Courcoubetis et al., 1998) suggest that buyers use an Intelligent Agent that will learn their preferences and then help in implementing them.

We discuss congestion pricing and its weaknesses using the example of the smart market pricing scheme presented in (MacKie-Mason and Varian, 1994a).

Congestion Pricing Congestion Pricing has attracted a fair amount of research. [see (Chung and Sirbu, 2001)(Courcoubetis et al., 1998)(Kelly and Tan, 1998)(MacKie-Mason and Varian, 1994b)(MacKie-Mason and Varian, 1994a).] The main goal of congestion pricing is to reduce the congestion on the Internet, and it can probably be generalized to managing usage levels of other resources.

One approach suggested for pricing of network services is "smart market," which is presented in (MacKie-Mason and Varian, 1994a). It requires users to decide on the price they are willing to pay for the transfer of each packet. In cases of no congestion, all the packets are sent without charge; however, in cases of congestion, the packets are sent according to their price.

Thus, during congestion, the packet that would be dropped is the cheapest one, and the price for non-dropped packets is the price attached to the most expensive packet that was dropped. Like the Vickery auction, which is also known as a second-price-auction, the smart market mechanism has the incentive compatibility and truthful revealing properties. The incentive compatibility means that the network users and network administrators will find it profitable to use the network. The truthful revealing property assures that users are induced to price packets according to their true evaluation of them; thus the network administrator knows the value of an incremental increase in the network's capacity.[3]

The advantages of congestion pricing are twofold. First, because they are being charged a higher price when resources are busy, customers are expected to balance their demand, and by doing so, to help the seller to satisfy the required quality-of-service using lower resources' capacity. A second advantage results from an economic analysis of congestion pricing schemes, where the "right price" for the resources usage will be charged and the "right amount" of investment in increasing resource capacity will be induced. (For a justification of this logic with an example, see (MacKie-Mason and Varian, 1994b).)

However, a major disadvantage of dynamic pricing for shared resources is that it prevents buyers and sellers from predicting the cost (or revenues) during a period. It turns out that predicting these expenses (revenues) is valuable for both parties.

Non-congestion Pricing An exception in the literature for Internet pricing schemes is the Flexible Service Plan (Altmann and Chu, 2001a). This scheme, is based on the findings of the University of California at Berkeley INternet Demand EXperiment (INDEX) project, which took place during April 1998 - December 1999, and measured the connection between demand for different bandwidths and the willingness of end users to pay for it. The flexible service plan suggests charging a fixed price for a fixed bandwidth and allowing end-users to purchase higher bandwidth on demand, for an additional cost (according to number of bits sent or minutes used). Such a pricing scheme is common in pricing of cell phone service.

An additional exception is (Maglaras and Zeevi, 2003), which discusses revenue management of shared resources. Their main result is that, for economic reasons, the operation of a shared

[3] This value is the value of the packets that were dropped.

pool of resources should be managed by a heavy traffic regime.

Pricing Computer Resources in Practice

This section presents the results of a field study that took place between November 2001 and January 2002. This study included interviews with practitioners and academics, as well as web surfing. We divide the results of this study into pricing for small users and pricing for large ones.

One of our conclusions from this study is that the demand for different resources has a "bursty" pattern. For example, the database update with new sales is usually done once in each period (say once a week), thus buyers would like to use bandwidth (for sales updating) for only a few hours during Saturday, and not to use bandwidth (for this purpose) during the rest of the week. Therefore, buyers do not pay for the continuous use of computer resources, rather they pay for a burst of use followed by a period of low usage. This fact is one of the main contributors to the low average usage rate of dedicated computer resources, and we should be aware of it when we consider different pricing schemes for shared computer services. Moreover, this explains why users of computer resources are willing to pay for a service with no service level guarantees.

Pricing for Small Users In order to learn about the current practices in services provided by Application Service Providers and their pricing, we conducted a survey using public information available on the web (Datadomo, 2001)(Exodus, 2001)(genuity, 2001)(Hosting, 2001)(Intersystems, 2001)(Nic, 2001)(Qwest, 2001)(Superb, 2001). We observed the pricing methods of many companies, and present here a few representative pricing schemes. The Application Service Providers included here are: Hosting, Superb, Alentus (formerly Pacific Netware), and NIC. In general, companies advertised prices for small consumers and not for large companies. Thus, this information is not representative of the pricing schemes that exist for large companies. The latter are discussed in the next subsection. As of January 2002, we did not find any proposals of shared service to large consumers.

Hosting is an American company with headquarters in Waltham, Massachusetts. Its parent company is Allegiance Telecom. Alentus (formerly Pacific Netware) is an international company

with branches in Edmonton, California and Singapore. Superb is an international company with branches in Washington (D.C.), Vancouver, and New Delhi. All three companies provide web, database, and application hosting, as well as dedicated Internet access, and some of them provide other related services that we do not mention here. The pricing scheme used by these companies (for the shared web and application hosting) is a set-up fee plus a monthly fee according to packages of disk space, monthly data transfer, and number of E-mail accounts. They price their dedicated Internet connections in the same manner, when the packages are defined according to the connectivity speed. All companies allow for additional memory, monthly data transfer, and e-mail accounts for additional charge.

Nielsen International Communication (NIC) is a European company, whose headquarters is in Brussels. It provides similar services in addition to web design and development, database integration and tailor-made database solutions. The pricing scheme offered by NIC is similar to the pricing of the other hosting companies. Yet NIC probably uses a usage-based fee for the traffic per month, since its published traffic fee is the only one charged at the end of each month.

We include information about the pricing of database software by InterSystems Corp., Oracle, and Microsoft since they implement a usage-based pricing scheme for their products. Both Microsoft and Oracle, as well as IBM, were selling their database software based on the number of users allowed to use the system. They all changed their pricing scheme to a usage based one according to a universal power unit, which measures the total power (in MHz cycles) of the computers using the software. The advantage of this pricing scheme, for companies that use the software through the web, is that they do not need to know the number of users of a web application, which is hard to measure. Moreover, due to the high availability of database software on the web, the number of users during a period might vary a lot, and change the associated costs for buyers. The disadvantage of the universal power unit pricing scheme is that when computers are upgraded, their power increases, which, in turn, results in an increase in the software's price.

InterSystems sells its database software according to a per-transaction fee or according to the number of users. Their idea is to allow small companies, the use of their database without a large initial expense.

Another pricing scheme is offered by DataDomo, which is an American company located in Okemus Michigan. It provides a service of building data-intensive web pages using its software. Its pricing includes a set-up fee to cover the fixed database cost and a monthly fee for maintaining the web and database servers as well as to cover a number of their software usage. An additional monthly fee is charged for additional disk space and for more runs of the software. Thus, DataDomo differs from the hosting companies, since it provides only one service and there are no different packages for different prices.

Pricing for Large Users In pricing for large users, similar ideas to the ones used for pricing services to small customers are used. The pricing includes a set-up cost, a fixed monthly fee (for a fixed amount of network usage), and the operation of the dedicated hardware. For a web connection, which is provided in a shared manner, there is also an additional cost for quantities over the fixed usage. The two main subjects of negotiation are the amount of usage that buyers are willing to commit themselves to, and the cost for each unit of usage within this fixed quantity. In general, the fixed usage fee is decreasing in the quantity buyers are willing to commit to (a quantity discount on all units). Surprisingly, in most cases this fixed unit-price is also the unit-price for usage above the committed usage level. This shows that sellers have enough capacity to serve higher demand, and that their main concern is to assure a fixed income base.

Two ways are used to measure the monthly usage level. The less common one, which is used by 5%-15% of the sellers, is to measure the total usage during the month. The main drawback of this measure is that it ignores the usage variation, and therefore makes service level guarantees and sellers' resources planning complex tasks.

The second, and more common, way is known as the 95/5 pricing. In this method, a month is divided into about 8200 intervals of 5 minutes. The usage (number of packets sent) in each such interval is recorded and at the end of the month the different usage records are ordered from the lowest to the highest. The monthly usage is charged according to the 95^{th} percentile of this order. The 95/5 scheme gives buyers an incentive to decrease their usage deviation. However, it still does not bound the usage a buyer can ask for in a period, and does not simplify the tasks of service level guarantees and sellers' resources planing. Moreover, a

wise usage management by buyers, i.e. concentrating a lot of demand in the 500 intervals that they do not pay for, can result in a very low payment for an actual high usage. This possibility is more realistic than it looks, since companies' weekly working hours are about 50 hours out of 168, so they use communication less than one-third of the time anyway.[4] Add to this the possibility of using different sellers, and the 95^{th} percentile can be driven down substantially.

1.3.2 Important Attributes of Pricing Schemes

We list here the important attributes of a pricing scheme for shared services, from both buyers' and sellers' perspectives.

Important Attributes from the Buyer's Perspective

From the buyer's point of view, the most important attribute of a pricing scheme for shared resources is to allow buyers cheap computer services. Recall that the traditional, dedicated service results in high costs due to high usage variability of computer resources. Moreover, today communication service, which is shared, is provided without service level guarantees. Thus, we consider the attributes of allowing for high deviations in demand, service level guarantees, and a low cost as the most important ones for a pricing scheme for shared resources.

Due to the high variability of demand, many companies do not have a complete characterization of their demand. Therefore, a pricing scheme that requires low information on the demand and its variability is preferable.

Another important objective of buyers is to have an accurate prediction of the charges for the computer services. This is important for budget planning and as a psychological hedge against the unknown. The importance of this attribute is further discussed in (McKnight and Boroumand, 2000), and empirical evidence for it, in the case of Internet usage, was given by the INDEX project. (See, for example, (Altmann and Chu, 2001b).) In a similar manner, a good pricing scheme should allow buyers to change their demand with small overhead costs. For example, a company that predicts a higher usage rate in the period before Christmas should be able to get additional resources at an appropriate cost.

[4] In fact, the peak hours for network usage are around noon to 2:00 p.m. every working day, which is less than 5%.

Finally, it is important that the pricing scheme be simple, so buyers will be able to track the cost they pay and the source for this cost, as well as to understand the pricing scheme.

Important Attributes from the Seller's Perspective

From the seller's point of view, the most important attribute of a pricing scheme for shared resources is to allow for resource planning and providing of service level guarantees. Otherwise, sellers will be unable to compete with dedicated resources providers, despite cheaper prices, and they will not be able to broaden their customer base. To achieve this, sellers need to give buyers incentives to smooth their demand (so the peak demand could be served from the fixed resource pool) as well as to shift demand from expensive (or congested) resources to cheap (or less congested) ones (thus decreasing the chances of excess demand for the expensive resource). Good resource planning also requires that the pricing scheme will give buyers incentives to report truthfully their predicted usage level.

In the survey of pricing in practice we noted that the per-unit price charged from buyers of Internet services does not increase if the buyer uses more than what they predicted. This proves the importance of a fixed revenue for the sellers, since they are more interested in promising themselves a fixed income than in gaining a small extra profit.

Finally, sellers need the pricing scheme to be simple to understand and to operate.

1.3.3 Qualitative Comparison among Pricing Schemes

This subsection offers a qualitative comparison among the different pricing schemes mentioned in terms of the important attributes discussed earlier. The results are summarized in Table 1-2. We consider the lower cost attribute as one related to the shared resource services, and therefore we ignore it in this comparison. Furthermore, since this is only a qualitative comparison, we rank the different pricing schemes as best, very good, good, medium, and worse. These ranks should be recognized as informs judgements rather than actual performances on each attribute. Finally, it is important to emphasize that these rankings are from the author's point of view and are based on a discussion with a small group of practitioners and academics. We think that this ranking is reasonable; however, we strongly recommend more detailed study of the subject. Such study can start using the Questions for Industry's Representatives in Appendix

Comparison of Pricing Schemes

	Fixed cost	Smart market	Flexible ser.plan	95/5
Supports demand variability	Good	Best	Very Good	Good
Supports service level guarantees	Worse	Best	Good	Medium
Supports resource planning	Worse	Very Good	Very Good	Medium
Gives incentives	Worse	Best (*)	Good	Medium
Easy to understand and operate	Best	Worse	Best	Best
Information requirements	Best	Worse	Good	Very Good
Known costs / revenues	Best	Worse	Good	Medium
Cost of demand's profile changes	Best	Worse	Best	Best

Figure 1-2: A qualitative comparison between the different pricing schemes, in terms of the important attributes for a shared resources scheme. (*) is not in seller's control.

A.2.

As can be seen from the Table, none of the pricing schemes is perfect (or even just the best) in each of the required attributes. The remainder of this subsection briefly discusses the ranking in Table 1-2.

For the attribute of allowing high demand variability, the smart market scheme performs best, since the price that buyers pay is independent of their demand and its variability. We consider the flexible service plan scheme as very good for this attribute, since, using it, buyers can easily purchase a higher level of resources, according to their needs. In fact, an integral part of this scheme is the consideration of demand variability. We rank both the fixed cost and the 95/5 schemes as having good performances as well, since in both the price for the additional demand does not vary greatly.

For the attribute of supporting service level guarantees, the best scheme is the smart market, because it guarantees the right service level to all buyers, according to their evaluation of the services provided. The flexible plan can also provide good service level guarantees, but since buyers can change their demand profile without early notice, we only rank it as good. For this attribute the 95/5 and fixed price schemes are medium and worse, correspondingly, since in both there are no constraints on buyers' demand variability, yet the first gives a cost incentive

to buyers not to substantially vary their demand too often.

For the attribute of supporting resource planning both the smart market and the flexible service plan are ranked as very good. The smart market gives buyers the right value for resource expansion; however, it does so only after the demand realization. The flexible service plan confines buyers' average demand and its peak. The 95/5 gives only a partial support for resource planning, since the 95^{th} percentile of buyers' predicted demand should be known by sellers, but the peak demand is ignored. The fixed cost scheme is the weakest scheme by this measure.

For the attribute of giving buyers the right incentives, the smart market mechanism does best; however, these incentives are not controlled by the seller. The flexible service plan gives incentives to buyers to truly report their expected (predicted) demand, as well as to smooth it, yet it is hard to relate the usages of different resources using this scheme. Again, the 95/5 gives buyers some incentives, which is better than what is done using the fixed price scheme.

For the attribute of simplicity to understand and operate, we rank all schemes except the smart market as best.

For the attribute of information requirements, the fixed cost requires only expected demand to decide on price, and therefore it has the lowest requirements, and is ranked first. The information requirement is increasing when we go to the 95/5 scheme, which also requires a knowledge of the 95^{th} percentile. A further increase is required for a good implementation of the flexible service plan (buyers should know when they have higher demand). Clearly, the demand information required by the smart market scheme is the highest.

For the attribute of fixed costs and revenues, our ranking is clear.

For the attribute of cost of demand profiles changes, again all schemes but the smart market got a good score (low cost of changes). In this case the smart market mechanism is deficient since the cost of such a change is unknown by both sellers and buyers; hence it is ranked last.

1.3.4 Analysis of the Comparison

From the short discussion in this chapter, we have two conclusions. The first is that none of the pricing schemes presented here is perfect for pricing of shared resources. On the one hand, the pricing schemes common in practice are very simple. These pricing schemes were good enough

to start up the industry; however, as the industry matures, some more sophisticated pricing schemes need to be put in place. Moreover, it looks as if the transformation of the computer industry to shared services is the right time to implement such a change. It is important to recall here that since the market we consider is of selling services to large companies, the pricing scheme used does not have to be as simple as the ones used today. On the other hand, congestion pricing and dynamic pricing look too cumbersome to be implemented as pricing for shared computer services. Their main drawback is that the total expense associated with such pricing depends on usage characteristics of other buyers, and cannot be obtained at the beginning of a contract.

The flexible service plan looks to be the best candidate, of the ones compared in Table 1-2, to be used as a pricing scheme for shared computer resources. However, even this scheme's performance can be improved. The three main attributes whose performance can be improved are the ones for which the smart market mechanism get a "best" score. A further investigation of why the smart market mechanism does so well in these attributes suggests that it is related to the elaborate admission control that the smart market mechanism includes. Thus, our second conclusion is that it will be beneficial to include admission control in a pricing scheme for shared services.

1.4 Summary

In this chapter we presented the pricing problem of application service providers, when they consider the provisioning of services based on shared computer resources. We described the supply chain of computer services, surveyed the different services provided, and mentioned the main computer resources required for provisioning of computer services. We list shared services' advantages, which are mainly the resource savings due to smoother demand, and its disadvantages, such as the need to assure privacy and security of data.

Based on the discussion in this chapter, we conclude that the main catalyst for provisioning of shared services is the high cost of computer resources. This high cost is a result of the low average usage rates of traditional, dedicated computer services. Thus, we believe that provisioning of shared computer services, or utility-like computing, will occur in the foreseeable

future. Furthermore, one of the important challenges of provisioning of such services is pricing them. We think that a good pricing scheme for shared services should provide buyers incentives to smooth their demand and not to over-use the common resources pool. Thus, the pricing scheme needs to have a usage-based component, i.e., buyers' expenses should be tied to their usage levels.

This chapter also presents a framework for analyzing pricing of shared resources. It gives a brief survey of the pricing schemes for computer resources: pricing schemes that are used in practice, and those that were suggested in the literature. It is important to note that both surveys are far from being complete, and it is recommended that further investigation of both be considered, as we suggest in Chapter 6.

The study reported here shows that pricing in practice is different between large and small users, and it is very simple. In addition, the literature of pricing computer resources is confined mostly to pricing of network usage, which is focused on managing network congestion. We also present a list of important attributes for a pricing scheme for shared resources, from both buyers' and sellers' points of view. Finally, we compare four pricing schemes with regard to these attributes. We consider such a comparison as a building block for choosing pricing schemes for shared resources; yet we emphasize that the results reported here are based primarily on the author's point of view.

The analysis of this comparison led to two conclusions. The first was that none of the pricing schemes compared is perfect for pricing of shared resources. It looks as if the pricing schemes in practice are overly simple whereas the ones in the literature are, in general, too detailed. The second conclusion is that the good theoretical performance of the pricing schemes in the literature can be attributed to the admission control included in these schemes. The importance of combining admission control and pricing is that this allows the seller to guarantee the required service level to buyers. Based on these two observations, it appears that there is room for a new pricing scheme that is tailored towards pricing of shared computer resources. Moreover, we think that it will be beneficial to include admission control in such a pricing scheme.

The next chapter presents token bucket admission controls, which have been extensively discussed in the literature of telephony and computer networks. Chapter also proposes to use

these admission controls as pricing schemes for shared resources. As a first step towards a novel pricing scheme, this dissertation provides a stylized model that allows for an analysis of the token bucket mechanisms. For example, in the rest of the dissertation we assume that buyers' demand distribution is known. Such an assumptions is helpful for the modeling of buyers' behavior; however, they might be too simple for practical purposes. We hope that this analysis will help to establish the token bucket as a basis for pricing schemes, and therefore help to solve the shortcomings of the pricing schemes discussed earlier in this chapter.

Chapter 2

The Token Bucket Pricing Schemes

2.1 Introduction

One of the common admission controls in the network and telephony literature is the *Token Bucket* (TB) method. The application of TB admission control to network flow allows for flexibility in the definition of the demand process and for differentiation between customers. Moreover, it is helpful in provisioning service level guarantees: packets that arrive at a source that has no tokens will be marked as packets "out of the spec," and if they arrive at a congested router, they will be dropped. Thus, if the network is uncongested, all packets will arrive at their destinations; but if the network is congested, sellers promise to serve only packets that are "in spec," a promise they can fulfill, given their resource levels.[1] We recall that both flexibility in the definition of the demand process and provisioning service level guarantees are important attributes for a pricing scheme for shared services.

In Section 2.2 we present the TB admission control and its extension, the token bucket with rate control (TBwRC). In Subsection 2.2.2, we explain the use of the TB admission controls as pricing schemes. Section 2.3 explains the problems needing to be solved in order for TB pricing schemes to be implemented and outlines the three main research questions this dissertation focuses on.

We consider the idea of using TB admission controls as pricing schemes for shared resources

[1] The resource level should, of course, be planned accordingly.

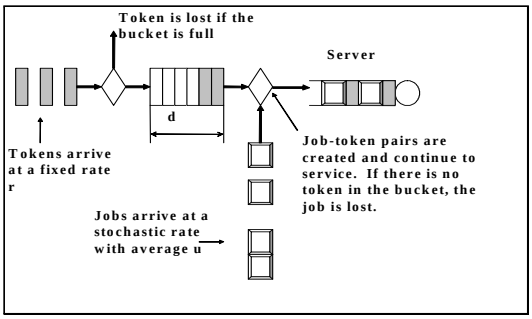


Figure 2-1: Diagram of the token bucket control mechanism.

as one of the major contributions of this work.

2.2 Token Bucket Pricing Schemes

2.2.1 Token Bucket Admission Controls

Token bucket admission control is discussed in the network literature; see (Berger, 1991),(Berger and Whitt, 1994), (Pancha and Zarki, 1995), (Dovrolis et al.,) and references therein.

It uses two parameters in order to define the demand for a network's resources: the token rate, denoted r, and the bucket depth, denoted d. Every source gets tokens at a rate r (not necessarily an integer) and has to have a token in order to send a packet. Tokens that are not used can be accumulated until level d is reached; and if a token arrives at a full bucket, it is lost. Thus, the average send rate is not larger than r and the size of the maximal burst is limited to d. Figure 2-1 illustrates this control mechanism. Using this method, a user who needs to send large files once in a while can request a low rate but a big bucket, while a user with low demand variability can request a small bucket.

For our purpose, we interpret the TB mechanism as an admission control for any resource. We need to define a **work unit** as being analogous to a packet. Then, each token represents such a unit, and a job that requires processing of a few work units is analogous to a file. A buyer that sends jobs to be processed is analogous to a source that sends files. Finally, we establish a **basic time period** in which tokens enter the bucket. It is important to choose this time period such that the processing of any job takes no more than one period.

With the definitions above, we can think of the buyer as having a finite bucket, with size d, full with tokens. Every time the buyer wishes to send a job to the seller (just as occurs with the network), the buyer attaches a token to each unit of work of the job. If there are no tokens in the bucket and the buyer sends the unit of work, it will not be processed. One can think of a unit of work and a token as a letter and a stamp.

Traditionally, a major disadvantage of TB admission controls is that there is no known methodology to relate demand characteristics to the depth and rate parameter choices. Other limitations of the TB admission control are discussed in (McKnight and Boroumand, 2000).

The Token Bucket with Rate Control Admission Control

Another version of the TB mechanism, which is common in the network literature, adds a jobs buffer that allows jobs to wait in it until the arrival of tokens. However, this practical extension does not change the analysis of service level presented in this dissertation. This extension holds because, Theorem 1 of (Berger, 1991), which requires that the arrival of jobs be a Markovian process and that the arrival of tokens be an independent renewal process, states that the probability of loss in a system with job and token buffers depends only on the combined capacities of these two buffers. Both requirements are reasonable and are assumed to hold throughout this dissertation.

Extending the idea above, let us think about a control where the jobs queue is infinite. Such a control is studied in the network literature (Berger, 1991). Using this control, when the bucket is empty, jobs are not lost but their processing is delayed, and therefore this control is called a rate control. This control models cases in which buyers can smooth their demand such that it will not be lost, by delaying the submission of jobs until tokens arrive.

2.2.2 Token Bucket as Pricing Schemes

To implement the TB as a pricing scheme, the seller sets the price per token and per rental of token storage space (bucket depth) for tokens. We will denote the token cost as R and the depth rental fee as D. Having these costs in hand, the buyer has incentives to order the minimal quantities of rate and depth such that her service requirements will be satisfied. Clearly, the same definitions can serve for the implementation of the TBwRC admission control as a pricing

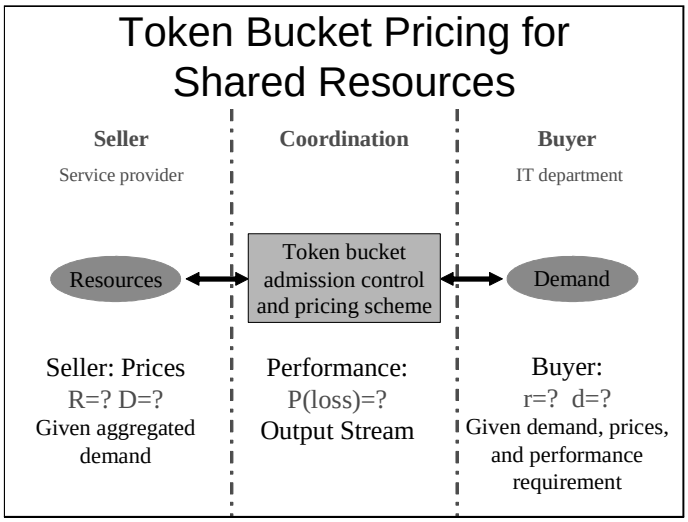


Figure 2-2: The implementation of a token bucket pricing scheme for shared resources requires solutions to a few problems. This figure indicates the problems we consider in this dissertation and their boundaries.

scheme. Moreover, a more sophisticated method of pricing token and depth, say, with a quantity discount, can be easily implemented as well. Yet in this work we confine ourselves to the simplest linear pricing mechanism.

2.3 Road Map of Dissertation

This section presents the questions addressed by this dissertation. There are additional issues that must be considered in order to facilitate the use of TB pricing schemes; but we defer a discussion of these questions to Chapter 6.

Figure 2-2 presents this dissertation's domain. We aim to facilitate the coordination between buyers and a seller of a shared resource, in a free market. The coordination between these parties will be based on TB admission controls and pricing schemes, as presented earlier in this chapter. We consider both the buyer's demand and the seller's resources as exogenous inputs to the model. Moreover, the buyer's demand is aggregated from many users, since we consider the buyer as a large company.

For our purpose, any demand that was accepted by the TB admission controls will be

served by the seller. An alternative method to formulate this is to assume that the seller can guarantee a known service level to any demand that is accepted by the admission control. In this method, the buyer's service level requirement, as described in the rest of the dissertation, already considers this degrading performance by the seller.

Within this domain, four main problems need to be solved in order to facilitate the use of TB pricing schemes.

2.3.1 The Buyer's Problem

Viewing the TB pricing from the buyer's perspective, it is reasonable to assume that the buyer wishes to minimize her expenditures, subject to some service level constraint. For now, we treat this service level as availability of resources, and we will better define this service level in the next chapter. Most likely, buyer's demand (which we assume to be known) is stochastic, and therefore the service level requirement will be probabilistic in nature. With this view, the buyer tries to balance the risk of low availability of resources with the risk of overpaying for tokens. In a reversed view, the buyer tries to balance the risk of losing tokens (paying for unused tokens) with the risk of losing jobs (due to unavailability of the resources). The buyer's problem addresses the traditional problem in implementing TB admission controls: the problem of choosing parameters that will result in satisfactory performances. Moreover, it adds the aspect of minimizing the cost of this rate and depth choice; thus, it requires the prices of the depth and rate as inputs.

To summarize, the buyer's problem will be a stochastic optimization problem, that of minimizing expenditures for rate and depth, subject to a probabilistic constraint. The inputs to the buyer's problem are the demand, service level requirement, and admission control chosen by the buyer; the prices of rate and depth, chosen by the seller; and the probabilistic dependence of the service level on the rate and depth choices. The outputs of the buyer's problem are the rate and depth choices of the buyer as a function of these inputs.

2.3.2 Performance of Token Bucket Admission Controls

The buyer's problem of choosing the rate and depth parameters can be solved if a description of the admission controls' performance is available. Thus, we consider such a description

as a major challenge in the implementation of TB pricing schemes. In looking for these performances, we rely on our assumption that demand is known. From the TB pricing schemes' points of view, such a description is a major advantage, because it will help to satisfy some of the important attributes of a pricing scheme for shared resources. (These attributes are discussed in section 1.3.)

To summarize, the performance analysis problem is to find a description for the performance of the admission control. This description will serve as an input to the buyer's problem. The inputs to the performance of the admission controls problem are rate and depth parameters, the demand characterization, and the chosen admission control. The output of this problem is the service level assured by the admission control as a function of its parameters.

2.3.3 The Seller's Pricing Problem

Viewing TB pricing from the seller's perspective, it is reasonable to assume that the seller wishes to maximize profits or revenues. The way we define the domain of our problem, with the seller resources planning outside its scope, constrains us to look only at the revenue aspect of the seller's problem.

There is a need to emphasize that the seller's problem is not a traditional revenue management problem. One of the major reasons for this is that in the TB case the sold "good" is an intangible service. In addition, in most traditional revenue management problems, the buyers' reactions to prices are simple. One common way to model buyers' behavior is by assuming that their evaluations for the "good" are drawn from a given reservation price distribution (usually iid). Such an assumption is reasonable when the number of potential buyers is large. However, in the TB pricing schemes, not only is the number of prospective buyers small, but their reactions are also not simple. In our case, buyers are strategic players as they solve their own expenditures minimization problems. Due to the importance of a known revenue and expenses for the seller and the buyers (respectively), the seller can consider a fairly long horizon (traditional contracts last six months). Thus, the seller's pricing problem does not include elements of dynamic pricing, making it simpler than some traditional revenue management problems.

Since optimal prices are a function of the demand faced by the seller, an essential input to the seller's problem is the output from TB admission controls. This problem is discussed in

the next subsection.

To summarize, the seller's pricing problem is determining prices to charge in order to maximize revenues in steady state. This problem differs from the seller's profit maximizing problem, which involves resources planning (and therefore is outside the scope of this dissertation). The inputs to the seller's problem are a pricing mechanism (recall, we assume that pricing is linear, but this is an arbitrary choice), and the effective demand the seller faces as a function of the prices she chooses. This last input is a complex combination of each buyer's demand characteristics and pricing scheme choices. The outputs of this optimization problem are the prices the seller will charge the different buyers.

2.3.4 Characterizing the Output Stream of the Token Bucket Admission Controls

Based on the seller's pricing problem, the characterization of the output stream from a TB or a TBwRC admission controls is to find the probabilistic nature (as the probability density function) of this stream in a steady state. We use the term "effective demand" for the output stream from the admission controls, since it contributes one stream to the "effective demand" seen by the seller. These characterization problems can be divided into two parts. The first one is to characterize the effective demand from an admission control of one buyer. The second part is to characterize the aggregation of buyers' effective demand from many buyers, in order to get the effective demand faced by the seller. If we assume that the aggregation arises from many independent buyers, we could use the central limit theorem and approximate the effective demand faced by the seller, using only a characterization of the first two moments of buyer's effective demand. Since the characterization of the effective demand is an essential input to the seller's pricing problem, we look for solutions that depend on the rate and depth choice by the buyer.

To summarize, the problem of characterizing the output stream from TB mechanisms is to find a probabilistic description of the effective demand faced by the seller; this effective demand results from the aggregation of the effective demands of each buyer. The inputs to the characterization of the effective demand problem are the buyer's demand, service level requirements, and depth and rate choice. The output is a probabilistic description of the

buyer's effective demand, such as the probability density function and the first moments of this demand.

2.3.5 Research Focus

Based on the input and output relations between the problems discussed above, we can think of a game-theoretic representation for the problem of implementing TB pricing schemes for shared resources. In the first phase, there is a cooperative game in which a buyer and the seller choose a TB pricing scheme and a pricing mechanism (linear prices, quantity discounts, etc.). Then, there is a non-cooperative game, in which the seller moves first and decides on prices in order to maximize her revenues. Based on this choice, each buyer tries to minimize her expenditures subject to her demand characterization and service level requirements. In such a game the seller has the advantage of moving first, since she can foresee buyers' strategy (rational choices) as reactions to her strategy.

This game theory framework is complicated by different levels of information sharing between the parties. Additional factors that increase the complexity are 1) that at the second phase the seller plays "against" many buyers, and the problems of 2) performance analysis and 3) the characterization of the effective demand.

In this dissertation we focus on solving the buyer's problem, the performance of TB admission controls, and the characterization of the effective demand of a buyer. We solve these problems under the assumption that all information is publicly available.

We do not consider the seller's problem because for TB pricing schemes to be more practical, a token market should be operated. Such a market will allow buyers to purchase tokens when required. We further discuss this problem in Chapter 6, where we point to future research problems. Clearly, such a market will change the buyer's strategy, and therefore there is less of a point in solving the seller's problem without considering this extra tool.

2.4 Summary

Chapter 1 concluded that the pricing of shared resources is an important and applicable problem and that the existing pricing schemes are not perfect for pricing of shared resources, and

therefore, a pricing scheme that is tailored toward pricing of shared resources is required. It also pointed out that the combination of admission control with pricing might be helpful in looking for new pricing schemes.

Accordingly, this chapter has presented the TB and TBwRC admission controls, which are well known in the computer network field. A main contribution of this dissertation, presented in this chapter, is the idea of using TB admission controls as pricing schemes for shared resources.

This chapter has also described a road map for the rest of this dissertation that focuses on the analysis of TB pricing schemes. This analysis addresses of three issues: the buyer's problem, the performance of TB admission controls, and the characterization of the effective demand of a buyer. We will solve these problems in the next three chapters.

A traditional disadvantage of TB admission controls is that there is no known methodology to relate demand characteristics to the depth and rate parameters choice. Thus, **??** addresses and solves the problem of the performance of TB admission controls. It combines results from large deviation theory with original results, and provides bounds and asymptotics on the service level provided to a user of the TB admission controls, given her demand and rate and depth choice.

These results are used in Chapter 4 to provide approximations for the optimal parameter choices by the buyer. The analysis of the buyer's problem is given for both the TB (loss of jobs) and the TBwRC (backlog of jobs). In order to gain insight into the resource level required by a seller in order to satisfy all its buyers' demand, we consider two extreme cases. If the seller has only one customer, the resource level required in order to assure no losses is the largest burst that this buyer can use in a period, i.e., d. However, if the seller has a very large (infinite) number of customers, all require a rate r (without a full correlation between their demands), and the resource level required to assure no losses is r times number of customers independent of the different $d's$ (due to the law of large numbers). In this second case the seller fully enjoys the statistical aggregation of demand and saves significant resources by the provisioning of shared resource service.

Chapter 5 combines results from the theories of large deviations and Brownian motion. It presents the analysis of the effective demand (demand that is admitted by TB admission control) from one buyer. In contrast to results from inventory theory, where it is typically

simpler to assume backlog rather than losses of sales, this analysis is far more complicated in the backlog (the TBwRC) case, then in the loss of sales (the TB) case.

In Chapter 6 we summarize this dissertation and discuss directions for future research both: for implementing the TB pricing schemes, and for a better understanding of the subject of pricing shared computers.

Chapter 3

Performance Analysis of the Token Bucket Admission Controls

3.1 Introduction

In this chapter we analyze the performances of the two TB admission controls that were presented in Chapter 2. Thus, we tackle here the traditional problem of implementing TB admission controls: the problem of choosing the rate and depth parameters of the TB in order to satisfy performance requirements.

Recall that we only focus on admission control of one resource (such as CPU time, storage, or communication). This is the second problem discussed in Chapter 2. As is mentioned there, the results of this analysis will be an input to the buyer's problem discussed in Chapter 4.

Section 3.2 describes the demand process considered in this dissertation. We also define the two most common service level measures: percentage of periods with loss, and percentage of jobs processed (fill rate). It turns out that the service level realized from TB admission controls is not easy to measure. Thus, an important result of this dissertation is that this service level can be approximated using large deviations techniques.

This chapter focuses on the analysis of the percentage of periods with loss (backlog) as a service level for TB admission controls. We show that the bucket level (without rate control) behaves as a two-sided regulated random walk. This random walk can be bounded (from below) with the one-sided regulated random walk that represents the bucket level with rate

control. We then develop bounds and asymptotics to the service level provided to buyers that use the TBwRC admission control, based on known results from the theory of large deviations and sequential analysis.

One of the main contributions of this dissertation is given in subsection 3.4.2. We provide, in Proposition 5, an upper bound on the shortages probability, which is the probability that a two-sided regulated random walk is on its "rare" threshold. This result is based on intuition gained from a sample path analysis of the differences between a one-sided and a two-sided regulated random walk. This intuition suggests a counting procedure that could be carefully implemented and would lead to the requested bound. The application of this bound requires additional approximations, which are described on subsection 3.4.3.

Finally, section 3.5 discusses the fill rate as a service measure and provides a method to approximate it based on known results regarding the distribution of a two-sided regulated random walk.

3.2 The Model

3.2.1 The Demand Model and Sequence of Events

The usage level denoted by u in any period is drawn from a distribution $F(u)$, with the non-negative line as its support, and letting $G_u(s) = E_u(e^{su})$ be its Moment Generating Function, which is assumed to exist at a neighborhood of $s = 0$. Studies on network performance ((Balakrishnan et al., 1998), (Paxson, 1997), and (Paxson, 1999)), show that network throughput is steady over intervals of minutes (and sometimes hours). This makes our model a good approximation for network usage, when periods are of a few minutes' length.

We further assume that the usage level in each period is independent and identically distributed (iid). This assumption does not hold for communications usage, which has been shown to have a self-similar nature in such studies as (Leland et al., 1994), (Paxson and Floyd, 1995), and (Barford, 2001). (For a definition of "self-similar" see (Leland et al., 1994).) For network usage, self-similarity means that the aggregated traffic along long intervals (hours) has a similar covariance to that of the aggregated traffic along small intervals (seconds). However, the observed demand in our case is aggregated over both time (five to fifteen minutes) and

users (since the buyer is assumed to be a large company); thus demand is more likely to be iid. Moreover, the demand process for CPU cycles is considered to be Gaussian and iid across intervals of minutes and many users. Therefore, our assumption is reasonable for the demand process the sellers face.

We do not claim that the assumptions described above hold for any computer resource; however, it seems as if these assumptions are good enough for an initial performance analysis of TB admission controls.

The sequence of events in the TB case is as follows: At the start of each period the bucket level increases by r (so the highest bucket level a period can begin with is: $r + d$); then the usage occurs according to the description above; and tokens are consumed up to their number in the bucket. If the usage in a period is lower than the token rate, no more than d tokens are carried to the next period. Note that if there was a loss in a period, the bucket level carried to the next period is zero; thus, if the next period's demand is larger than r, loss will reoccur.

In the TBwRC we follow the same sequence of events. However, if the bucket is empty, jobs are backlogged rather than lost; thus the bucket might have a negative number of tokens.

3.2.2 Service Level

The two most common service level definitions for the TB admission control are "percentage of periods with loss" and "fill rate," which is the percentage of work processed. Under both definitions, buyers want the probability of loss to be smaller than a threshold. Parallel service level definitions for the TBwRC are "percentage of periods with backlogs" and "percentage of periods when demand is satisfied in the period it arrives." In this case we can also measure the average delay of a job and the average delay of a job given that it was backlogged.

To elaborate on the first definition, consider an indicator that gets the value 1 if no losses (or backlogs, in the TBwRC case) occur during a period, and zero otherwise. The buyer wants the percentage of periods with an indicator value of zero to be less than or equal to a threshold. For the second definition, each job carries such an indicator. Both definitions require a steady state loss probability to exist. Indeed, the TB control guarantees the existence of a steady state, since the bucket level holds values in the closed set $[0, d]$. For steady state distribution to exist in the TBwRC case, we need to further assume that $r > E(u)$. In this chapter we

assume that this holds even in the TB case and that (without loss of generality) the first period starts with a full bucket. In this dissertation we confine ourselves mostly to the percentage of periods with loss performance measure. A discussion of the fill rate as a service level is given in section 3.5.

To analyze the service level of percentage of periods with loss in the TB case, we recall that if a period had a loss, the bucket level carried to the next period is equal to zero. For simplicity, we assume that every time the bucket level is zero, losses happen (this assumption holds when demand is continuous). Denoting the bucket level in period i, as $\widetilde{L}_i$, we define the bucket level at steady state as $\widetilde{L} \equiv \lim_{n\to\infty} \widetilde{L}_n$, and we also use this notation ($\widetilde{L}$) for the **bucket level** process, when no confusion arises. We write the service level requirement for the bucket level case $P\left\{\widetilde{L}=0\right\} \leq 1-\alpha$, where α (the service level) is typically higher than 90%.

To analyze the service level of percentages of periods with backlogs in the TBwRC, we recall that if a period had backlogs, the bucket level carried to the next period is lower or equal to zero. Denoting the bucket level with rate control in period i, as L_i, we define the bucket level minus jobs in the queue as the bucket level with rate control, and its distribution at steady state is noted as $L \equiv \lim_{n\to\infty} L_n$. We also use this notation (L) for the **bucket level with rate control** process, when no confusion arises. We write the service level requirement for the bucket level with rate control case $P\{L \leq 0\} \leq 1-\alpha$.

3.3 The Bucket Level as a Random Walk

This section analyzes the bucket level, $\widetilde{L}$ process, and the bucket level with rate control process L and shows that they behave as a two-sided and a one-sided regulated random walk respectively. In addition, it presents the **shortfall process** Y. We show that these one-sided regulated random walks can represent the level of the bucket with rate control and provide upper bounds on the service levels of the TB admission control mechanism. To help the reader understand the evolution of the different random walks, Figure 3-1 gives an example of these different processes. In this example, $d=1$, $u \sim U[0,1]$.

Using the notation u_i as the usage level for the i^{th} period, and recalling our assumption

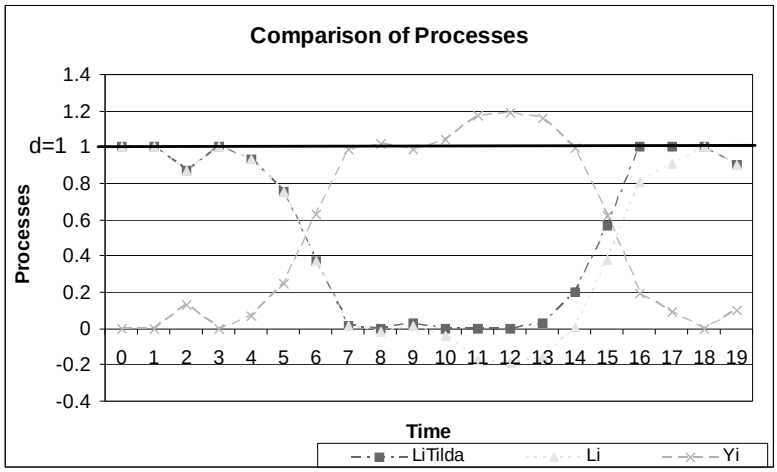


Figure 3-1: The different processes of the bucket level $\widetilde{L}$, the bucket level with rate control L, and the shortfall Y. In this example, $d = 1, u \sim U[0,1]$. Note that $L = \widetilde{L}$ from time 0 until the first time $\widetilde{L}$ becomes 0, at period 7. Then $\widetilde{L} > L$, which holds until the next time L is at d, at period 18.

$\widetilde{L}_0 = d$, we can write:

$$\begin{aligned} \widetilde{L}_1 &= \min\{d, \max[0, d + r - u_1]\} \\ \widetilde{L}_2 &= \min\left\{d, \max\left[0, \widetilde{L}_1 + r - u_2\right]\right\} \\ &\vdots \\ \widetilde{L}_{i+1} &= \min\left\{d, \max\left[0, \widetilde{L}_i + r - u_i\right]\right\}. \end{aligned} \tag{3.1}$$

Looking at the equation for $\widetilde{L}_1$, the minimum is taken between the event where the usage rate was lower than r (thus the bucket level is d), and the event where the usage level was higher than r, (thus the bucket level can be reduced to $d + r - u_0$). However, since in this case the bucket level cannot decrease below zero, one needs to add the maximum operator. The extensions of this procedure for any period are given in (3.1).

We define $x_i \equiv u_i - r$, and note that the Moment Generating Function of x is given by:

$$G_x(s) = e^{-sr} G_u(s). \tag{3.2}$$

Recall that we assume that the token rate r is chosen such that $E(u) < r$, leading to $E(x) < 0$.

Plugging x_i instead of $r - u_i$ into (3.1) shows that the bucket level process behaves as a positive drift random walk that is regulated between 0 and d. Moreover, this representation also describes an inventory problem, in which the control mechanism is a base stock level, where shortages become lost sales.

Computing the probability of losses in steady state is equivalent to computing the probability of a two-sided regulated random walk to be on its lower boundary. Unfortunately, we do not have such a probabilistic description in hand. However, in Figure 3-1 we notice that $L = \widetilde{L}$ from time zero until (not including) the first time $\widetilde{L}$ becomes zero; then $\widetilde{L} > L$, which holds strictly until the next time at which L is back at d. From then on $L = \widetilde{L}$ again and the above relations between the two processes are regenerated. In Figure 3-1, both processes coincide at times $t = 0$ and $t = 18$. Thus, we can bound the loss probability by looking at the steady state threshold crossing probability of a one-sided regulated random walk that evolved in a similar way. This bound should be effective, since we are interested in a high service level, so there is a low probability for the original random walk to be on its lower boundary, and both random walks should behave "similarly."

In fact, in (Ioannis and Yong, 2002), it is shown that, when $d \to \infty$, the exponential term in which the probability of loss sales decreases is equal to the exponential term in which the percentage of periods with backlogs decreases. Moreover, the required one-sided regulated random walk is described by the TBwRC. For the same sequence of events as above, L_i is given by:

$$L_{i+1} = \min\{d, L_i - x_i\} \tag{3.3}$$

Thus, the evolution of the L process is regulated only at its maximum value d, omitting the Max operator from the evolution equations for the $\widetilde{L}$ process. In addition, the L process has a steady state distribution due to its positive drift (recall our assumption $E(u) < r$).

We define "cycle" according to the L process:

Definition 1 *Cycle: Let* $s^1 = \inf_{i \geq 0}\{L_i \leq 0\}$, *and* $T^1 = \inf_{i \geq s^1}\{L_i = d\}$. *Then* T^1 *ends the first cycle. We now treat* T^1 *as time zero for the second cycle and define recursively, for* $j > 1$*:* $s^j = \inf_{i \geq T^{j-1}}\{L_i \leq 0\}$, *and* $T^j = \inf_{i \geq s^j}\{L_i = d\}$, *then* $T_j - T_{j-1}$ *is the length of the* j^{th} *cycle.*

In fact, s^1 is the stopping time for the first downcrossing of zero, T^1 the stopping time for

return to d after a downcrossing. In a similar manner, for $j > 1$, s^j is the stopping time for the first downcrossing of zero after T^{j-1}, and T^j is the stopping time for the first return to d after s^{j-1}.

In words, a cycle is the duration from the return of L to a level d after losses happened until its first return to level d after **at least one more period with loss occurs**. Figure 3-1 includes one cycle in periods zero until 18, when the next cycle starts. This regenerative structure is a key observation that we use in subsection 3.4.2, to provide a tight bound on the loss probability in the bucket level process.

Note that the length of cycles is finite, with probability one, since despite the positive drift the event of downcrossing zero will happen (the regulator at d gives us an infinite number of trials to downcross) and afterwards L returns to d, with probability one due to its positive drift.

Let us summarize this discussion. As can be seen in Figure 3-1, and is shown in (Ioannis and Yong, 2002) $\widetilde{L} \geq L$ and therefore:

$$P\left(\widetilde{L} = 0\right) \leq P\left(L \leq 0\right). \tag{3.4}$$

The L process is equivalent to an inventory problem in which the control mechanism is a base stock level, and shortages are backlogged. Using this last equivalence, we follow the derivation in (Glasserman, 1997) and obtain a bound on the steady state probability that the L process is lower than a threshold. Recall, that this threshold is the base stock level in the inventory case, and the bucket depth in the TBwRC case. Defining $Y_i = d - L_i$ as the shortfall, i.e., how far the bucket level is from d, we obtain from 3.3, $Y_{i+1} = \max\{0, Y_i + x_i\}$, which is the Lindely Recursion. Thus, Y, the **shortfall process,** is distributed as a maximum value of a random walk with a negative drift ($E(x) < 0$) and independent and identically distributed increments. We note this random walk process as S, where $S_0 = 0$, and $S_i = \sum_{j=1}^{i} x_i$. Lemma 2 summarizes this result:

Lemma 2 *The steady state probability $P\left(Y \geq d\right)$ of the shortfall process exceeding a threshold is equal to this threshold-crossing-probability by a random walk with a negative drift $E\left(u\right) - r$:*

$$P\left(Y \geq d\right) = P(S_N \geq d) \tag{3.5}$$

where $N = \inf\{i = 0...\infty | S_i \geq d\}$*, and* $P(S_N \geq d)$ *is therefore, the probability of this random walk ever crossing* d*.*

Note that, N is the stopping time in which the random walk crosses the level d.

The proof of Lemma 2 can be found in chapter 7 of (Gallager, 1996). Using these processes, (3.4), and (3.5), the buyer's service level can be bounded by:

$$P\left(\widetilde{L} = 0\right) \leq P(L \leq 0) = P\left(Y \geq d\right) = P(S_N \geq d). \tag{3.6}$$

This equation allows us to bound the steady state loss probability of the TB control; it is central to our approximation of the service level obtained from the TB mechanism.

3.4 Bounds and Asymptotics on the Service Levels

To simplify the results in this subsection, we use the generic processes' names: one-sided regulated random walk rather than TBwRC, and two-sided regulated random walk rather than the shortfall process (TB). We will return to the TB terminology in the next section.

This section gives upper and lower bounds and asymptotics to the service level provided by a two-sided regulated random walk. It starts with citing bounds and asymptotics for the one-sided regulated random walk, and then presents our approach for tightening the upper bound for the two-sided regulated random walk. We consider this tight bound as one of the major contributions of this dissertation. Subsection 3.4.3 describes additional approximations that make this bound useful. Finally, this section presents lower bounds on the loss probability for both random walks.

3.4.1 Upper Bounds on Backlog Probability for a One-sided Regulated Random Walk

Lemma 3 (below) cites results from different sources regarding the threshold crossing probability of a random walk and the waiting time in a D/GI/1 queue. These approximations are necessary because the exact distribution of the regulated random walks in steady state cannot be determined. The analysis in the preceding subsection, combined with the results of Lemma

3, gives bounds and asymptotics of the service level provided to the buyer for each (d, r) pair.

Lemma 3 *Threshold crossing probability of a random walk. For a random walk of an iid random variable x, with a negative drift, i.e., $E(x) < 0$, let $G_x(s) \equiv E_x\left(e^{sx}\right)$ be the moment generating function of x, and assume that $G_x(s) = 1$ for some $s^* > 0$ (we named s^*: the conjugate point of X), then for any $d > 0$:*
Part A: (See equation 20 of Chapter 7 of (Gallager, 1996)):

$$P(S_N \geq d) \leq e^{-s^* d}, \tag{3.7}$$

where $N = \inf\{i = 0..\infty \,|S_i \geq d\}$.
Part B: (see Chapter 8 of(Siegmund, 1985)):

$$\lim_{d\to\infty} \frac{1}{d} \ln P(S_N \geq d) = -s^* \ln \left[\frac{1 - E_{s^*}\left[e^{-s^* x^+}\right]}{s^*\left(E_{s^*}\left(x^+\right)\right)} \right], \tag{3.8}$$

where x^+ is a positive element drawn from the distribution of x. The expectation is taken after a change of measure to the conjugate distribution (noted as $E_{s^*}(\cdot)$). Note that the argument of the ln*, on the right-hand side, is one over the moment generating function of the residual life of x^+ evaluated at the point s^*.*
Part C: The probability that the waiting time Y for a D/GI/1 queue (with r as the interarrival process and u_i as the service requirements, and $x_i \equiv u_i - r$) is higher than d can be bounded (see (Ross, 1974)) is:

$$\left\{ \sup_{0\leq a} E\left[e^{s^*(x-a)} \,|x > a\right] \right\}^{-1} e^{-s^* d} \leq P(Y \geq d) \leq \left\{ \inf_{0\leq a} E\left[e^{s^*(x-a)} \,|x > a\right] \right\}^{-1} e^{-s^* d}. \tag{3.9}$$

Remark 1: Given our analysis in the last section, the probability discussed in part C of the Lemma is equal to the threshold crossing probability for a one-sided regulated random walk.

Remark 2: The coefficients of the exponential term in part C are strictly between zero and one; thus the upper bound in part C is never worse than the one in part A.

Remark 3: When the usage in each period follows the memoryless exponential distribution (with parameter λ), the argument inside the ln in the right-hand side of part B (as well as in

the upper and lower bounds in Part C) is given by: $1 - s^*/\lambda$. Thus, the probability of loss, for the one-sided regulated random walk, is known exactly for any $d > 0$.

Remark 4: More elegant results for the bounds are available regarding distributions that satisfy conditions of New Better (Worse) than Used, in both (Glasserman, 1997) and (Ross, 1974). However, we do not mention these results here.

In section 3.5, we discuss bounds and asymptotics for the fill rate service level using large deviation results, as well as approximations for this service level based on Brownian motion. These approximations use results from (Harrison, 1985) and can be further refined based on (Berger and Whitt, 1992).

In the remainder of this paper we assume that x is not a constant and that $G_x(s)$ satisfies the requirements of Lemma 3.

3.4.2 Upper Bound on Loss Probability for a Two-sided Regulated Random Walk

Here we suggest a tighter upper bound on the loss probability when using a TB control. This approach is new and, to the best of our knowledge, is the first description of a consistent method of estimating the difference between the loss probability of a one-sided regulated random walk and that of a two-sided regulated one.

We consider the events of loss in a one-sided regulated random walk (with "backlog") $P(L \leq 0)$, and in a two-sided sided regulated random walk (with "lost sale") $P\left(\widetilde{L} = 0\right)$. Note that both processes have a positive drift.

We note the cycle lengths (from Definition 1) as T with expectation $E(T)$. Focusing on the level of a one-sided regulated random walk, there might be a few occasions, during each cycle, in which loss occurs. Each such occasion is defined as a subcycle that includes the duration of time in which $L \leq 0$, i.e., the time from just after downcrossing zero to just before upcrossing zero. We emphasize that lengths of different subcycles are not iid. A subcycle's length depends on the undershoot at the beginning of the subcycle. This undershoot is not necessarily independent between the first and the second subcycles (or any other two for that matter). Indeed, each k^{th} subcycle is iid with the other k^{th} subcycles but is not iid with subcycles that are not the k^{th} ones. However, in the exponential demand case the undershoot

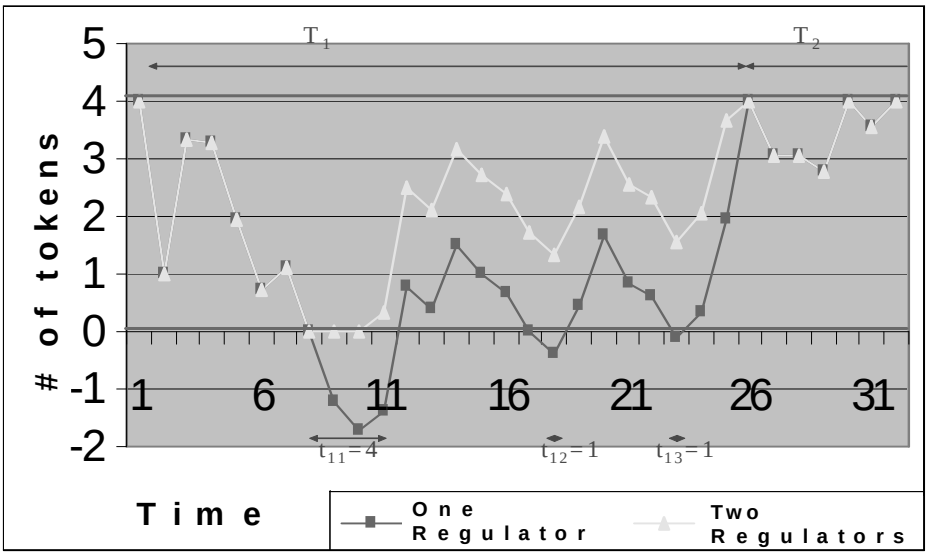


Figure 3-2: One and a "half" cycles of losses for a one-sided (with squares) and a two-sided (with triangles) random walk. Both walks have a positive drift and are regulated at $d = 4$; the two-sided regulated is also regulated at 0.

of each subcycle is identically distributed due to the memoryless property of the exponential distribution. Thus, in this case, subcycles form a renewal process, and their lengths are iid. We define t as a generic subcycle and $E(t)$ as the expected length of such a subcycle. (A concise definition for subcycles is given in Appendix A.5.)

In Figure 3-2 the first cycle begins at time zero and ends at time 26 when the second cycle begins. This second cycle continues after the end of the figure. The first subcycle is from time 7 (when the bucket level at the one-sided regulated random walk is below zero) and ends at time 10, before the upcrossing. The length of this subcycle is 4 periods. The second and third subcycles are in periods 17 and 22, and each of them has a length of one period.

For ease of exposition, we call the number of periods with losses in a one-sided regulated random walk the **periods with loss in an L cycle.** The number of periods with losses in a cycle for a two-sided regulated random walk will be called the **periods with losses in an $\widetilde{L}$ cycle,** despite the fact that cycles are defined using only the one-sided regulated random walk.

Thus, we can write the backlog probability in a one-sided regulated random walk as follows:

$$P(L \leq 0) = E\left(\frac{\text{periods with loss in an } L \text{ cycle}}{\text{number of periods in a cycle}}\right). \tag{3.10}$$

In a similar manner, we write the probability of loss in a two-sided regulated random walk:

$$\begin{aligned} P\left(\widetilde{L} = 0\right) &= E\left(\frac{\text{periods with loss in an } \widetilde{L} \text{ cycle}}{\text{number of periods in a cycle}}\right) \\ &\leq E\left(\frac{\text{periods with loss and usage} > r \text{ in an } L \text{ cycle}}{\text{number of periods in a cycle}}\right) \\ &\leq P(L \leq 0). \end{aligned} \tag{3.11}$$

The first inequality holds, since any period with loss in the two-sided regulated random walk has a usage higher than r and is also a period with backlog in the one-sided regulated walk, which is always lower or equal to the two-sided regulated random walk. However, there might be periods with usage higher than r, in which there are backlogs in the one-sided random walk, but there are no losses in the two-sided regulated one. In Figure 3-2, three such periods are at the downcrossing of zero at the beginning of the second and third subcycles, and period 10 is the last period in the first subcycle. The second inequality holds, since the numerator in (3.11) is smaller than the one in (3.10) due to the additional restriction on the usage in each period in (3.11).

Based on (3.11), we can bound the probability of losses in the two-sided regulated random walk using the expected number of periods with losses and usage higher than r during a cycle in a one-sided regulated random walk.

We define: $n \equiv$number of subcycles within a cycle and t_k the length of the k^{th} subcycle (if such exist -or zero otherwise). Then, using the expectations of n and t_k, we get:[1]

Lemma 4 *The expected number of periods with backlogs and usage higher than r during a cycle in a one-sided regulated random walk, noted as M, is given by:*

$$M = E(n) + \overline{F_u}(r) \sum_{k=1}^{\infty} E(t_k)$$

[1] We want to thank Yashan Wang for his help in the proof of this Lemma.

Lemma 4 can be interpreted as follows: If we assume independence between the usage rate in a subcycle's period and the location of this period within a subcycle, the expected number of periods with losses and usage higher than r during a cycle could be written $M = \overline{F_u}(r) \sum_{k=1}^{\infty} E(t_k)$. However, such an assumption ignores the first period in each subcycle, a period for which we know that its usage was higher than r (since the one-sided regulated random walk declined during this period). Thus, we need to add the correction factor of one period per subcycle, which is, in expectation, the expected number of subcycles per cycle, to get $M = E(n)\left(1 - \overline{F_u}(r)\right) + \overline{F_u}(r) \sum_{k=1}^{\infty} E(t_k) = E(n) + \overline{F_u}(r) \sum_{k=1}^{\infty} E(t_k - 1)$. However, this ignores the extra information that the other periods within this subcycle have not ended the subcycle yet. This additional information means that these periods did not increase the random walk enough to cancel the undershoot; it suggests that the usage during these periods is higher than the usage in regular ones. Lemma 4 shows that –in expectation– the $t_k - 1$ periods within a subcycle (excluding the downcrossing period) have probability of usage higher than r just as if they are t_k periods with a regular usage.

Lemma 4 is proved in Appendix A.5. Using this lemma we provide a tighter (than the ones summarized in Lemma 3) bound on the loss probability in a two-sided regulated random walk:

Theorem 5 *The percentages of periods with losses in a two-sided regulated random walk,* $P\left(\widetilde{L} = 0\right)$*, is related to the percentages of periods with losses in a one-sided regulated random walk* $P(L \leq 0)$ *via:*

$$P\left(\widetilde{L} = 0\right) \leq P(L \leq 0)\left(\bar{F}_u(r) + \frac{1}{E(t)}\right) \tag{3.12}$$

The intuition behind Theorem 5, is similar to the one for Lemma 4. To understand why the factor to the right of $P(L \leq 0)$ is smaller than one, as we expect from (3.11), we observe that the expected length of a subcycle is larger than the expected number of trials until a demand lower than r is realized ($1/F_u(r)$). This is true because the time until the first increase of the random walk is distributed geometrically with the above probability and is shorter than a subcycle (since the first move up need not be large enough to cover for the undershoot at the beginning of a period). Thus, $E(t) \geq 1/F_u(r)$, from which $\bar{F}_u(r) + 1/E(t) \leq 1$ follows.

Proof. of Theorem 5:

Let us look at the renewal defined by the cycles, where renewal epochs are beginnings of cycles, and attach to it a reward function with a value one whenever the level of the random walk is lower than zero, and zero otherwise. This is a legitimate reward function since the reward in each interarrival (cycle) depends only on what happened during this interarrival time. Thus, the probability of backlogs in a one-sided regulated random walk can be interpreted as the reward rate,[2] and we can use the key renewal theorem (see, for example, (Gallager, 1996)) to write:

$$\begin{aligned} P(L \leq 0) &\equiv E\left(\frac{\sum_{k=1}^{\infty} t_k}{T}\right) && (3.13)\\ &= \frac{E\left(\sum_{k=1}^{\infty} t_k\right)}{E(T)} && (3.14)\\ &= \frac{E(n)\,E(t)}{E(T)}, && (3.15) \end{aligned}$$

where we add the second equality since the expected number of periods with losses in a one-sided regulated random walk during a cycle ($E\left(\sum_{k=1}^{\infty} t_k\right)$) is the expected number of subcycles multiplied by the expected length of a subcycle.

Equation (3.11), the definition of M, and the key renewal theorem, can be stated as:

$$P\left(\widetilde{L}=0\right) \leq \frac{M}{E(T)}.$$

Substituting M from Lemma 4, we can bound the loss probability in the two-sided regulated random walk, $P\left(\widetilde{L}=0\right)$, by:

$$P\left(\widetilde{L}=0\right) \leq \frac{M}{E(T)} = \frac{E(n) + \overline{F_u}(r)\sum_{k=1}^{\infty} E(t_k)}{E(T)}$$

Using (3.14), we can replace the summation divided by $E(T)$, with $P(L \leq 0)$. Moreover, using

[2] We could also represent this system as a D/GI/1 queue (arrivals every time unit, with service requirement distributed as U_i/r), and attach a reward 1 for each customer that had to wait more than d.

(3.15) we can replace $E(n)$ over $E(T)$, with $P(L \leq 0)$ over $E(t)$, to get:

$$\begin{aligned} P\left(\widetilde{L}=0\right) &\leq \frac{P(L \leq 0)}{E(t)} + \overline{F_u}(r) P(L \leq 0) \\ &= P(L \leq 0)\left[1/E(t) + \overline{F_u}(r)\right], \end{aligned}$$

concluding the proof. ■

Theorem 5 bounds the probability of losses for a two-sided regulated random walk. However, the expected length of a generic subcycle cannot be found in a closed-form. Thus, in order to use Theorem 5 we need to approximate this expected length. In the next subsection we describe a procedure that allow us to estimate the expected subcycle length.

3.4.3 Estimating the Expected Subcycle Length

The estimation of the expected length of a subcycle will follow from applying Wald's equality (given next in Equation 3.16) to subcycles, relating the expected distance traveled by a one-sided regulated random walk, from the start to the end of a subcycle, to the expected length of a subcycle. The expected distance traveled by the random walk is the level of the undershoot (once downcrossing zero) plus the level of the overshoot (once upcrossing d). Using the asymptotics from sequential analysis (given in Part B of Lemma 3) we can approximate the expected overshoot and undershoot. In this subsection we give a detailed explanation of this procedure and demonstrate it for the normal and exponential usage cases.

The Procedure

Here we give a procedure to estimate the expected subcycle length $E(t)$, for non arithmetic distributions, and demonstrate it for the normal and exponential demand cases. We use ideas and results borrowed from sequential analysis as given in (Siegmund, 1985); therefore, we state the results using the shortfall process. We also recall Wald's equality (see, for example Chapter 3 of (Gallager, 1996)), which states that for a stopping time N and a function $g(x_i)$ of the iid random variable x_i, such that $E[g(x_i)] = G$,

$$E\left[\sum_{i=1}^{N} g(x_i)\right] = GE(N). \tag{3.16}$$

Consider the events that lead to a subcycle: first, the shortfall process had an upcrossing of d to a height of $d+a$ with an overshoot $a \geq 0$. Then, there were $t \geq 1$ periods in which the shortfall stayed above d; finally, there was a period in which d was downcrossed to a level $d-b$, with an undershoot $b > 0$. Thus, we can consider t as a stopping time of a random walk with steps $x_i = u_i - r$ (a negative drift), going down from its overshoot level a to zero.

By Wald's equality (3.16), applied to the subcycles

$$E(t_k) = E(S_{t_k})/E(x),$$

due to the linearity of expectation we can find $E(t_k)$ by adding the expected overshoot and undershoot of subcycles. Closed-form approximations for the overshoots and undershoots can be found, as we explain below, using the asymptotics of sequential analysis, presented in part C of Lemma 3. However, our approximation of $E(S_{t_k})$ is independent of k, and therefore our approximation for $E(t)$ is less accurate.

In (Siegmund, 1985) it is shown that for non-arithmetic distributions, as a threshold increases to infinity (decreases to infinity) its overshoot (undershoot) is distributed as the residual life of the positive (negative) part of x, noted as x^+ (x^-). Thus, the expected overshoot can be approximated by

$$\frac{E\left((x^+)^2\right)}{2E(x^+)} \tag{3.17}$$

and the expected undershoot is given by the same expression, substituting x^+ with x^-.

Note that using this approximation for the overshoot of the first subcycle is reasonable, since a high service level requires a high depth. However, using it for the overshoots of other subcycles, as well as for the undershoots of a, is less accurate.

In the case of exponential demand this approximation is completely accurate for the overshoots (due to the memorylessness of x^+, which is exponentially distributed). However, even then the approximation for $E(S_{t_k})$ is not accurate for the undershoots, since the distribution of x^- is not exponential.

Finally, our approximation for the expected subcycle length is

$$E(t) = \left(\frac{E\left((x^+)^2\right)}{2E(x^+)} + \frac{E\left((x^-)^2\right)}{2E(x^-)}\right)\frac{1}{E(x)}. \tag{3.18}$$

Remark: The approximation of the expected undershoot and overshoot can be replaced by simpler expressions $E(u\,|u > r\,) - r$, and $r - E(u\,|u < r\,)$, respectively. For further discussions of these approximations, see the description of the enhanced approximation presented in Chapter 5.

Normal Demand

We now demonstrate the computation of $E(t)$, according to (3.18) for the normal (u, σ) demand case.

Here x is normally $(u - r, \sigma)$ distributed. Moreover, the conjugate distribution (which is "responsible" for the overshoot) is normally $(r - u, \sigma)$ distributed, and thus the overshoot and undershoots are symmetric problems and we can compute any one of them. However, an accurate computation for either one of them is not available and we will approximate them assuming that the distance until the overshoot is large.

The following analysis is in line with chapter 8 of (Siegmund, 1985). Combining his equations (8.39) and (8.44) we get:

$$E\,(a) = \exp\left[\sum_{n=1}^{\infty} \frac{1}{n} P\,(S_n \leq 0)\right],$$

where $S_n \equiv \sum_{i=1}^{n} x_i$. Now, standardizing the distribution to a standard normal one, we can write:

$$\begin{aligned} E\,(a) &= \exp\left[\sum_{n=1}^{\infty} \frac{1}{n} P\left(\frac{S_n - n\,(r-u)}{\sqrt{n}\sigma} \leq -\frac{\sqrt{n}\,(r-u)}{\sigma}\right)\right] \\ &= \exp\left[\sum_{n=1}^{\infty} \frac{1}{n} \Phi\left(-\frac{\sqrt{n}\,(r-u)}{\sigma}\right)\right], \end{aligned}$$

where Φ is the Cumulative Distribution Function of a standard normal random variable. In equation (4.37) (Siegmund, 1985) defines:

$$V\,(\mu) \equiv \frac{1}{2\mu^2} \exp\left[-2 \sum_{n=1}^{\infty} \frac{1}{n} \Phi\left(-\frac{\sqrt{n}\,|\mu|}{2}\right)\right],$$

which can be numerically computed. However, to simplify our solution we use the approximation given in (4.38) of (Siegmund, 1985):

$$V(\mu) \approx e^{-0.583\mu},$$

as $\mu \to 0$. Thus, using this notation, we can approximate (substituting $\mu = \frac{2(r-u)}{\sigma}$):

$$\begin{aligned} E(a) &= \frac{\sigma}{2(r-u)\sqrt{2V\left(\frac{2(r-u)}{\sigma}\right)}} \\ &\approx \frac{\sigma}{2(r-u)\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}. \end{aligned}$$

Symmetrically, we can approximate $E(b)$ with the same expression. Plugging these approximations into (3.16), we get:

$$E(t) \approx \frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}} - 1.$$

Thus, for the normal case, we can rewrite (3.12) to get the following approximated bound on the loss probability in the bucket process:

$$P\left(\widetilde{L}=0\right) \lesssim p(L \le 0)\frac{1+\bar{F}_u(r)\left[\frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}-1\right]}{\frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}-1}.$$

Exponential Demand

We now demonstrate the computation of $E(t)$, according to (3.18), for the exponential (λ) demand case.

Let us estimate the expected undershoots, $E(b)$. The distribution of x^- is the distribution of $-(\tilde{u}-r)$, where $\tilde{u}$ is distributed as an exponential (λ) given that it is smaller than r. Thus,

the density of $\tilde{u}$ is given by

$$f(\tilde{u}) = \frac{f(u)}{p(u \leq r)} = \begin{cases} \frac{\lambda e^{-\lambda u}}{1-e^{-\lambda r}} & \text{when } 0 \leq u \leq r \\ 0 & \text{otherwise} \end{cases}.$$

Consequently, the density of x^- is given by:

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda(-x+r)}}{1-e^{-\lambda r}} & \text{when } 0 \leq x \leq r \\ 0 & \text{otherwise} \end{cases}$$

The first and second moments of the undershoot can now be computed and substituted into (3.17) to get the undershoot's expected value

$$E(b) = \frac{e^{r\lambda}\left(2 - 2r\lambda + r^2\lambda^2\right) - 2}{2\lambda\left[1 + e^{r\lambda}\left(r\lambda - 1\right)\right]}. \tag{3.19}$$

Since the distribution of x^+ is exponential, the expected overshoot is just the distribution's mean: $E(a) = 1/\lambda$. Combining this with (3.19) and (3.16), we get:

$$E(t) = \frac{r^2\lambda e^{r\lambda}}{2\left[1 + e^{r\lambda}\left(r\lambda - 1\right)\right]\left(r - E(u)\right)} - 1$$

Thus, for the exponential case, we can rewrite (3.12) to get the following bound on the loss probability in the bucket process:

$$\begin{aligned} P\left(\widetilde{L} = 0\right) &\lesssim p\left(L \leq 0\right) \frac{1 + \bar{F}_u(r)\left[\frac{r^2\lambda e^{r\lambda}}{2\left[1+e^{r\lambda}(r\lambda-1)\right](r-E(u))} - 1\right]}{\frac{r^2\lambda e^{r\lambda}}{2\left[1+e^{r\lambda}(r\lambda-1)\right](r-E(u))} - 1} \\ &= p_s \frac{1 - e^{-r\lambda} + \frac{r^2\lambda}{2\left[1+e^{r\lambda}(r\lambda-1)\right](r-1/\lambda)}}{\frac{r^2\lambda e^{r\lambda}}{2\left[1+e^{r\lambda}(r\lambda-1)\right](r-1/\lambda)} - 1} \\ &= P_s\left(\frac{2\left(1 - e^{-r\lambda}\right)\left[1 + e^{r\lambda}\left(r\lambda - 1\right)\right]\left(r\lambda - 1\right) + r^2\lambda^2}{r^2\lambda^2 e^{r\lambda} - 2\left[1 + e^{r\lambda}\left(r\lambda - 1\right)\right]\left(r\lambda - 1\right)}\right). \end{aligned}$$

3.4.4 Lower Bounds on the Loss Probability for One- and Two-sided Regulated Random Walks

We rewrite the loss probability in a two-sided regulated random walk $P\left(\widetilde{L}=0\right)$, by conditioning on its specific level. When the level of this random walk at the end of a period is l, a loss will happen in the next period if its usage is higher than $l+r$, that is, with probability $1-F_U\left(r+l\right)\equiv\overline{F}_U\left(r+l\right)$. Thus, noting the cumulative distribution function of $\tilde{L}$ as $F_{\tilde{L}}\left(l\right)$ we can rewrite this loss probability:

$$P\left(\widetilde{L}=0\right)=\int_0^d\overline{F}_U(r+l)dF_{\tilde{L}}\left(l\right). \tag{3.20}$$

Based on (3.20), we can bound:

$$\int_0^d\overline{F}_U(r+l)dF_{\tilde{L}}\left(l\right)\geq\overline{F}_U(r+d),$$

since $\overline{F}_U\left(u\right)$ is a decreasing function. We name this bound the "loose lower bound." The interpretation of this bound is that the loss probability in practice is never lower than the loss probability when the bucket is full. A better bound can be found assuming that the service level constraint is active, $P\left(\tilde{L}=0\right)=1-\alpha$. Then

$$P\left(\widetilde{L}=0\right)\geq\left(1-\alpha\right)\bar{F}_U\left(r\right)+\alpha\bar{F}_U\left(r+d\right),$$

which we name the "lower bound".

We note that both bounds also bound the backlog probability for the corresponding one-sided regulated random walk.

3.5 Fill Rate as the Service Level

Here we extend our analysis to a service level definition of fill rate, using similar results to those given in Lemma 3. The expression for the fill rate is:

$$FR_{Bucket}=\frac{E\left(u\right)-E\left[\max\{u-\widetilde{L}-r,0\}\right]}{E\left(u\right)} \tag{3.21}$$

where the second expectation in the nominator is the expected units lost per period. The maximum operator ensures that when no losses occur, i.e., $u - \widetilde{L} - r < 0$, no unit losses are considered. Using the same logic as in (3.4), we see that the fill rate for the bucket level is higher than the fill rate for the bucket level with rate control process (where the fill rate for the latter process is defined as the jobs that are served in the period they arrive). The fill rate for the bucket level with rate control is identical to the fill rate of the shortfall process, which can be computed by:

$$FR_{Shortfall} = \frac{E(u) - E[\max\{0, \min\{Y - r - d + u, u\}\}]}{E(u)}$$

where the second expectation in the nominator is the expected units lost. The additional minimum operation ensures that when the shortfall at the end of the period is larger than u (due to very large shortfall at the beginning of the period), the backlog counted is not more than u. This is the case when token supply $(d + r - y)$ is much lower than $-u$, an event that cannot happen in the original bucket level process.

Results similar to those in Lemma 3 are given in (Glasserman, 1997): for any fixed r, $1 - FR_{Shortfall}$ is proportional to $C/[E(u)\, s^*]\left(1 - e^{s^* r}\right) e^{-s^* d}$.

3.5.1 Approximating the Fill Rate Using a Brownian Motion

Using the first two moments of the demand process one can approximate the bucket level as a Brownian motion. It is a common practice to approximate random walks and regulated random walks using Brownian motion and regulated Brownian motions. Such approximations are more accurate in the heavy traffic limit, i.e., when the TB rate is close to the mean demand. Thus, such approximations are better when depth is cheaper.

Here we present the conditions under which such an approximation is reasonable, and the lines to the proof of convergence of the bucket level distribution to a corresponding regulated Brownian motion.

It is well known that, with a proper scaling, a continuous version of the random walk described by a sum of iid random variables with a zero mean and a unit standard deviation converge, in distribution, to a Wiener process (i.e., to a standard Brownian motion). To state

this mathematically, we use $\lfloor t \rfloor$ for the largest integer smaller than t. Then, let x be a random variable drawn from an iid distribution with a zero mean and a unit standard deviation. Let Y be the following continuous process: $Y_t = \sum_{i=1}^{\lfloor t \rfloor} x_i + (t - \lfloor t \rfloor)\, x_{\lfloor t \rfloor + 1}$. Note that at the beginning of each period Y has the value of the random walk $\sum_{i=1}^{N} x_i$; and that during each period the Y process advances at a fixed rate to the value of the random walk at the beginning of the next period (instead of jumping by x_{i+1} at the end of the period, as does the random walk). We further define a series of scaled processes $Z_t^n = Y_{tn}/\sqrt{n}$ (for each such process we scale up time by n and scale down the state space by $\sqrt{n}$). Then:

$$\lim_{n \to \infty} Z_t^n \stackrel{d}{=} W_t,$$

where W_t is a Wiener process and $\stackrel{d}{=}$ denotes convergence in distribution.

An intuitive explanation for this convergence can be as follows. We recall that a standard Brownian motion has iid increments that are *normally* $(0, t)$ distributed. It turns out that the above scaling creates, in distribution, a process with these two properties.

To understand the independent increments property, consider two adjacent increments of Z^n. Two such increments are dependent since they include a portion from the same period (i.e., from the part in which the Y process advances in a fixed rate). Clearly, when n increases the number of periods that are included in each increment of Z^n grows rapidly, and therefore the influence of the single joint period diminishes. To motivate the *normally* $(0, t)$ distribution of Z^n's increments we look at integer times. It is seen that at time 1: $Z_1^n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i$, which as $n \to \infty$ is distributed *normally* $(0, 1)$, by the law of large numbers. Thus, $\lim_{n\to\infty} Z_1^n \sim$ *normally* $(0, 1)$. In a similar manner $\lim_{n\to\infty} Z_2^n = \lim_{n\to\infty} \frac{1}{\sqrt{2n}} \sum_{i=1}^{2n} x_i \sim$ *normally* $(0, 2)$, and so on. However, the process $\lim_{n\to\infty} Z^n$ is not necessarily continuous; thus we cannot claim that it is a Brownian motion, only that its distribution is the same as that of a Brownian motion.

Using this equality in distribution, it has been shown that a one-sided (two-sided) regulated random walk can be approximated by a one-sided (two-sided) regulated Brownian motion. Details of such proofs can be found in Section 4 of (Berger and Whitt, 1992), and in Chapter 8 of (Kushner and Dupuis, 2001). Section 4 of (Berger and Whitt, 1992) also presents the

technical conditions of a heavy traffic assumption, conditions that are less restrictive than the ones we have assumed so far.

There are additional methods that can be used to approximate the random walk we are investigating. One such method is to estimate the random walk by applying the queueing analogue, which follows from the Lindely recursion that describes the bucket level evolution. There are closed-form results for an M/M/1/C queue, and other finite-buffer single-server queues can be analyzed or well approximated in some cases. In (Berger and Whitt, 1992) results based on such analysis are presented. A different methodology is to construct a finite dimension Markov chain and analyze it. This last method requires a discretization of the state space in the range zero to d and will be more accurate as the number of states increases. However, the computation of the solution of such a chain is also typically more complicated as the number of states increases. We choose to use the Brownian motion approximation mainly due to its tractability and adequately accurate predictions.

The reason for the tractability of the Brownian motion, as demonstrated in (Berger and Whitt, 1992), is that there are closed-form solutions for its density. Moreover, a similar regenerative structure to the one present in the two-sided regulated random walk can be found in a two-sided regulated Brownian motion. (This regenerative structure is the cycles mentioned in Definition 1.)

Thus, we can approximate the bucket level using an analogous Brownian motion. With such an approximation we can use Proposition 5 (page 90) of (Harrison, 1985) that gives the distribution of a two-sided regulated Brownian motion as

$$f_{\widetilde{L}}\left(\widetilde{l}\right) = \begin{cases} \frac{-\theta e^{\theta \widetilde{l}}}{1-e^{\theta d}} & \widetilde{l} \in [0,d] \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta = 2\left(r - E\left(u\right)\right)/\sigma^2 > 0$. This proposition also shows that the average control per time unit that is required to keep the bucket level positive is given by:

$$\frac{E\left(u\right) - r}{\left(1 - e^{\theta d}\right)}. \tag{3.22}$$

Since the average work per unit of time is give by $E\left(u\right)$, (3.22) can be divided by the mean

demand and we can approximate the work fill rate, denoted FR:

$$FR = 1 - \frac{E(u) - r}{E(u)(1 - e^{\theta d})}. \tag{3.23}$$

Thus, we can use the above expression as an approximation for the work fill rate realized from the TB admission control. A similar expression can be found for the TBwRC case.

3.6 Summary

This chapter analyzes the performances of the two TB admission controls that were presented in Chapter 2. Thus, it tackles the traditional problem of implementing TB admission controls: choosing the rate and depth parameters of the TB in order to satisfy performance requirements. As is mentioned in Chapter 2, the results of this analysis will be an input to the buyer's problem which we discuss in the next chapter.

This chapter has described the demand process considered in this dissertation and defined the two most common service levels: percentage of periods with loss and percentage of jobs processed (fill rate). It turns out that the analysis of the service level realized from TB admission controls is not simple. Thus, an important result of this paper is that this service level can be approximated using large deviations techniques.

We focus on the analysis of the percentage of periods with loss (backlog) as a service level for the TB (TBwRC) admission control. We show that the bucket level (without rate control) behaves as a two-sided regulated random walk. This random walk can be bounded (from below) with the one-sided regulated random walk that represents the bucket level with rate control. We then develop bounds and asymptotics to the service level, provided to buyers using either mechanism, based on known results from the theory of large deviations and sequential analysis.

The upper bound on the backlog probability, which is the probability that a two-sided regulated random walk is on its "rare" threshold, is given by Theorem 5. This bound is one of the main results of this dissertation. In order to make this result applicable, we described a procedure to approximate the expected subcycle length. We demonstrated this procedure for cases of normal or exponential demands. These two demand examples will be carried with us through the rest of this dissertation. We will display the use of the tools provided in this

dissertation using these demand examples and they will also allow us to examine the numerical accuracy of our analytical results.

Finally, section 3.5 discusses the fill rate as a service measure and provides a method to approximate it based on known results regarding the distribution of a two-sided regulated random walk.

In the next chapter, we will use the results developed here as surrogates for the actual service level provided by the TB admission controls. Using this result we will show that a constrained version of the buyer's problem is convex, which in turn allows us to find efficient bounds to the buyer's problem.

Chapter 4

The Buyer's Problem

4.1 Introduction

This chapter analyzes the optimal rate and depth choice of TB schemes in order to minimize the buyer's expenditures subject to service level constraints. This problem was described in Chapter 2. We analyze the optimal choice for cases of deterministic demand, where service level is defined as no losses, and of stochastic demand, where service level is such that the long-term loss probability in any period is lower than a threshold. For stochastic demand, we use the bounds and asymptotics developed in Chapter 3. Using these results, the buyer's optimal parameter choice under both TB pricing mechanisms can be estimated. We give a closed-form solution to cases where demand is normally or exponentially distributed and demonstrate our procedure by solving the buyer's problems under such assumptions. A numerical simulation to compare the results of our approximation to other parameter choices is given. It shows that, in the normal demand case, the cost of our proposed approximation is typically within 1-2% of the optimal solution (the worst results in this study are less than 8% of optimal) for the TB case, and within 1% of optimal for the TBwRC pricing scheme. In the latter case, the methodology presented here leads to the optimal solution when demand is exponential. We consider the results of this chapter as essential for the implementation of TB pricing schemes.

When the cost of a token is $R > 0$, and the cost of renting (for each period) a token storage place (in the bucket) is $D > 0$, the buyer's problem is:

$$\min_{d,r}(Dd + Rr)$$

$$s.t$$

$$\text{service level constraints}$$

$$r \geq 0$$

$$d \geq 0.$$

Note that since D is the cost of having a place to store a bucket for one period, whereas R is the cost of purchasing a new token, there is no point in purchasing any depth if $R \leq D$.

It is important to be aware that the buyer's problem presented above is equivalent to the problem faced by a manufacturer who needs to choose her production capacity and warehouse size when the cost of capacity is R per-period-per-unit of production capacity and the cost of a storage location for each unit is D per-period. Such a representation is realistic for environments with continuous production, such as in the oil or milk industry. (See (Graves, 1982) for such models.) In more general settings, one can consider the manufacturer's choice of production capacity at a cost of D per-unit and an average variable manufacturing cost of R per-unit-per-period.[1] A similar parallelism is with a production planning problem, with lost sales and a base stock level d inventory control (with holding costs of D dollars per-unit-per-period), and with r as the planned expected production rate (with a unit production cost of R).

In Chapter 3, we saw that the service level constraint can be expressed using a one-sided random walk (TBwRC) or a two-sided regulated random walk (TB). Moreover, the service level constraint cannot be written in closed-form and requires approximations. In addition, the constraint on a two-sided regulated random walk is more complicated than the constraint on a one-sided regulated one. We dealt with these major difficulties using the bounds and asymptotics developed in Chapter 3.

An additional source of complexity in solving the buyer's problem is that the buyer's controls

[1] Note that both representations assume that all of the capacity is used during any production period.

are both the rate and the depth. Alternatively, using the random walk terminology, the buyer's controls are both the drift and the threshold not to be crossed by the random walks. To the best of our knowledge, the work most similar to the present study is (Glasserman, 1997). However, he focuses on solving the problem using only the threshold as a control. Moreover, (Glasserman, 1997) confines his work to the backlog case (the TBwRC); thus, he does not treat the two-sided regulated random walk case (TB).

As a reminder, in the stochastic case we consider an infinite horizon problem. The demand is assumed to be iid with a moment generating function $G_u(s) = E_u(e^{su})$ that has a conjugate point, i.e., $G_u(s)\, e^{-sr} = 1$ has a positive solution, noted s^*. The sequence of events is as follows: at the beginning of each period the bucket level is increased by r (so the highest bucket level a period can begin with is $r + d$); then the usage is realized, and tokens are consumed to satisfy this demand, up to the total number of tokens in the bucket. If the usage is lower than the tokens rate, no more than d tokens are allowed to be carried to the next period. In a similar manner, if the usage is higher than the token rate, the additional demand will be lost in the TB case, but will be backlogged in the TBwRC case.

The service level we focus on is the percentage of periods with losses that is required to be smaller than a threshold. For the fill rate service level we replace the service level constraint with the approximation based on a Brownian motion (as discussed in Chapter 3). This exchange makes the buyer's problem convex and can be easily solved numerically. Simulation results, which are not reported here, show that for a normal demand (with the same parameters as in subsection 4.4.1) the resulting fill rates are close to, and higher than, the requested ones. (These results are better as the variance decreases.) Moreover, as Corollary 14 in the next chapter shows, when the buyer's demand is exponential both the fill rate and the percentage of periods with losses service levels are identical. Thus, the solutions presented here are also viable for the fill rate service level in this demand case.

Section 4.2 presents a solution to the buyer's problem in a deterministic case, gives the solution to the buyer's problem in two simple cases, and claims that if these cases happen the TB mechanism is not appropriate. Since these cases depend on the TB cost parameters, this analysis points to the range the different cost parameters of the TB scheme can take.

Section 4.3.1 gives a general solution to the buyer's problem when she faces a TB pricing

scheme. This solution can be used to numerically find the optimal parameters of TB pricing schemes. Finally Section 4.4 gives a few different closed-form bounds and approximations to the optimal solution of the buyer's problem when demand is normally or exponentially distributed. It also includes numerical results that show the efficiency of the approximation presented here for these demand cases.

4.2 Preliminary Analysis of the Buyer's Problem

When demand is deterministic the service level could be chosen to be no losses, and we present the analysis of this case in the next subsection. Note that in the practical[2] stochastic case TB pricing schemes have an infinite number of depth and rate pairs that can satisfy each service level. Thus, by choosing the right price sellers can induce the buyer to choose the best rate and depth parameters (from their perspective).

4.2.1 Deterministic Demand

As we mentioned in the presentation of the TB method, the long average usage rate allowed by a choice of token rate parameter r is r. Therefore, if the buyer's demand is deterministic with a fixed rate r she must purchase r as her token rate in order to avoid losses. Moreover, in such a case there is no need for purchasing bucket depth. A more interesting deterministic case is when there are different usage levels. Here we consider a usage rate level of x during α of the time, and a usage rate $y > x$ during the rest of the period, $(1 - \alpha)$ of the time. This demand pattern is shown in Figure 4-1. Assuming that at time zero the usage level just dropped from y to x and the bucket is empty, the buyer's optimization problem is

[2]The exact meaning of "practical" will be given later in Sections 4.2.2 and 4.3.

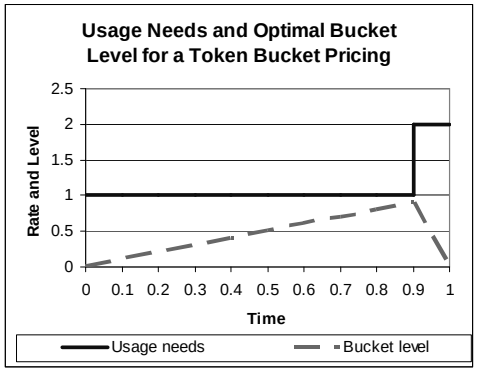


Figure 4-1: Required usage level and bucket level evolution during one deterministic usage cycle in which $x = 1$, $\alpha = 0.9$, $y = 2$. The optimal rate and depth parameters are: $r = 1.1$ and d=0.9.

$$\begin{aligned} & \min_{r,d} Rr + Dd \\ & s.t \\ r \quad & \geq \quad y(1-\alpha) + x\alpha \qquad (4.1) \\ d \quad & \geq \quad (y-r)(1-\alpha) \qquad (4.2) \\ d \quad & \geq \quad 0, \end{aligned}$$

where constraint (4.1) is the service level constraint and it dictates that the rate chosen is at least as high as the expected period's usage rate. Constraint (4.2) dictates that the buyer's choice of the bucket depth is such that the bucket is large enough to collect (during the period of low demand) enough tokens to cover the usage during the period of high demand. Note also that the non-negativity constraint on the rate parameter is redundant in view of the service level constraint.

The above is a simple linear programming problem for which the graphic solution is presented in Figure 4-2.

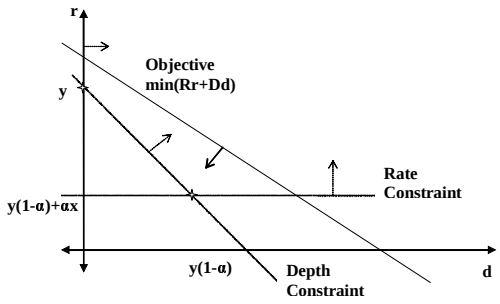


Figure 4-2: The rate and the depth should be chosen in accordance with the feasible set as described here. For $R > 0$, $D > 0$, the solution is in either one of the extreme points, noted with stars.

The solution to this linear programming problem is given by

$$\begin{aligned} r^* &= y(1-\alpha) + x\alpha \\ d^* &= (y-r)(1-\alpha) = (y-x)(1-\alpha)\alpha, \end{aligned}$$

when $D/R \leq (1-\alpha)$, or by

$$\begin{aligned} r^* &= y \\ d^* &= 0 \end{aligned}$$

when $D/R \geq (1-\alpha)$. In this case the depth is very expensive, compared with the rate, or the low usage level period is very short, and thus the buyer has no reason to purchase any bucket depth. In fact, for this case, the TB pricing mechanism is not appropriate for the buyer, and the seller should choose the ratio R/D such that this solution will not be optimal.

Note that only the ratio between the prices of the rate and depth decides which extreme point is optimal. We will see that the dependence of the solution on the ratio between the parameters' prices also characterizes the approximated solutions for the stochastic case, which is presented in Section 4.3.1.

Note that this analysis can be easily extended to cases of multiple usage levels and less restrictive service level requirements. Figure 4-1 shows the required usage level and bucket

level during one usage cycle with $x = 1$, $\alpha = 0.9$, $y = 2$, resulting in $r = 1.1$ and $d = 0.9$, a solution in accordance with the first case.

4.2.2 Stochastic Demand

To analyze the service level of percentages of periods with losses, we recall that if a period had a loss, the bucket level carried to the next period will be zero (this assumption holds, almost surely, when buyer's demand is continuous). Also, for simplicity, we assume that every time the bucket level is zero, losses happen. Thus, denoting the bucket level in steady state as $\widetilde{L}$, we can write the buyer's problem as

$$\begin{aligned}
&\min_{d,r} Dd + Rr \\
&s.t. \\
&P\left\{\widetilde{L} = 0\right\} \leq 1 - \alpha \\
&r \geq 0 \\
&d \geq 0,
\end{aligned}$$

where α, the service level, is typically of the order of 90%. Replacing the service level constraint with $P(L \leq 0)$, where we define the bucket level in steady state as L for the TBwRC, gives us the buyer's problem in this case. We recall from Chapter 3 that the distribution of the L and $\widetilde{L}$ processes is identical to that of one- and two-sided regulated random walks, respectively.

In the optimal solution at least one of the three constraints (service level constraint, and non-negativity constraints of the rate and the depth parameters) should be active.[3] Therefore, we solve the buyer's problem under these three different cases. It turns out that the non-negativity of the rate constraint cannot be active in the optimal solution, since the cost of the corresponding solution grows indefinitely for any positive service level requirement.

We focus here on solving the buyer's problem when facing the TB admission control and remark on the application of the results for the TBwRC case.

[3] This holds for any continous and differentiable cost function and can be easily seen from the Karush-Kuhn-Tucker conditions when the mixed second derivative of the service level constraint regarding r and d is not 0.

With the buyer's problem in mind, two trivial cases arise. In the first one, the ratio between the rate price and the depth price, D/R, is large (i.e., $D/R \to 1$). Thus, there is no point in purchasing any depth and the depth non-negativity constraint is active. In such a case, buyers will purchase a very high rate. Then, any period with demand higher than r results in a loss, and r should be chosen as the $1-\alpha$ percentile of the usage level distribution, leading to the following solution and cost:

$$\begin{aligned} r^* &= F_u^{-1}(\alpha) \\ d^* &= 0 \\ Z^* &= RF_u^{-1}(\alpha). \end{aligned}$$

The second case, when the ratio D/R is small, hints that in the optimal solution the rate non-negativity constraint might be active. However, for the TBwRC, the rate constraint is $E(u) < r$ in order for a steady state distribution to exist (else from some point of time all jobs will be backlogged). Even in the TB case the rate non-negativity constraint cannot be active. For an infinite horizon, when no tokens enter the bucket, all jobs will be lost. Moreover, we claim that for a "reasonable" pair of service level requirement and demand, the real constraint on r is, as in the case of a TBwRC, $E(u) < r$. However, $E(u) < r$ is not required in the TB, since loss of jobs is allowed. Thus, a skewed distribution, for which the demand is extremely high with a low probability, can satisfy the service level constraint without $E(u) < r$.

We claim that in these two extreme cases the use of a TB pricing scheme is inappropriate. In the first case ($D/R \to 1$), there is no point in operating a TB framework since it only constrains the rate r. The second case ($E(u) > r$), of huge spikes in demand that can be lost on a regular basis, is not practical. In these cases the seller's choice of the cost parameter is poorly gauged, and thus does not give buyers the right incentive to control their usage.

In the rest of the paper we assume that in order to satisfy the service level constraint the token rate must be higher than the expected usage. Thus, the Buyer's Problem (BP) in the

TB with stochastic demand case is:

$$\min_{d,r} Dd + Rr \qquad \text{(BP)}$$

$$s.t$$

$$P\left(\widetilde{L} = 0\right) \leq 1 - \alpha \qquad \text{(Service Level Constraint)}$$

$$r \geq E(u) \qquad \text{(Rate constraint)}$$

$$d \geq 0. \qquad \text{(Depth Constraint)}$$

When considering the buyer's problem in the TBwRC and stochastic demand case, the service level constraint above is replaced by $P\left(L \leq 0\right) \leq 1 - \alpha$.

4.3 Analysis of the Buyer's Problem When Token Bucket Pricing Schemes are Practical

In Chapter 3, we saw that L and $\widetilde{L}$ are distributed as a one-sided and a two-sided regulated random walk, respectively, and we developed bounds and asymptotics for this service level. Thus, the results of Chapter 3 deal with the challenge of expressing the service level constraint. In this section we deal with the second challenge of solving the buyer's problem, namely, the control of both rate and depth that is available for the buyer. We reduce the buyer's problem to a problem with only a rate control, and show the convexity of the latter.

This section gives upper and lower bounds on the optimal solution to the buyer's problem facing either the TB or the TBwRC pricing scheme. The lower bounds are based on relaxation of the buyer's problem, whereas the upper bounds are based on solving constrained versions of the buyer's problem, using the results developed in Chapter 3. We recommend the latter as the buyer's solution since they will satisfy her service level constraint.

In order to show that the constrained version of the buyer's problem is convex, we need the following result –interesting in itself– that shows that s^* is strictly increasing in r and $1/s^*$ is strictly convex in r.

Theorem 6 *Let S be the random walk associated with the random variable x, $E(x) < 0$, and*

let $G_x(s) = E_x\left(e^{sx}\right)$ *be the moment generating function of* x*, such that* $G_x\left(s\right) = 1$ *for some* $s^* > 0$*. Then:*

A. $s^*\left(r\right)$ *is strictly increasing with* r*.*

B. $1/s^*\left(r\right)$ *is strictly convex in* r*.*

C. The first two derivatives of s^* *(regarding to* r*) are given by:*

$$\frac{ds^*}{dr} = \frac{s^* e^{s^* r}}{G_u'\left(s^*\right) - r e^{s^* r}}. \tag{4.3}$$

$$\frac{d^2 s}{dr^2} = \frac{s e^{sr}\left(sG_u' + e^{sr}\right) + s' e^{sr}\left(G_u' - re^{sr} + srG_u' - sG_u''\right)}{\left(G_u' - re^{sr}\right)^2}. \tag{4.4}$$

The proof of Theorem 6 is given in Appendix A.4.

4.3.1 Upper Bound on the Buyer's Cost

This section focuses on solving an approximation for the buyer's problem for a general demand process. The approximation for the results from relaxing the service level constraint, according to:

$$P\left(\widetilde{L} = 0\right) \leq P(L \leq 0) \leq e^{-s^* d}, \tag{4.5}$$

results from the analysis in Chapter 3. We briefly recall the analysis that led to (4.5):

- Subsection 3.3 ends with (3.6):

$$P\left(\widetilde{L} = 0\right) \leq P(L \leq 0) = P\left(Y \geq d\right) = P(S_N \geq d).$$

  The interpretation of this equality is that the steady state probability for the bucket-level process to be zero is not higher than for the steady state probability of the shortfall process to cross the threshold d. This, in turn, is equal to the probability of a one-sided regulated random walk ever crossing d, according to (3.5) from Corollary 2.

- Part A of Lemma 3 bounds the steady state threshold crossing probability for a one-sided regulated random walk in (3.7):

$$P(S_N \geq d) \leq e^{-s^* d},$$

Thus, (4.5) allows us to replace the service level constraint with a tighter constraint in BP resulting in the following problem, which we name the Buyer's Constrained Problem (BCP):

$$\min_{d,r} Dd + Rr \tag{BCP}$$

$$s.t$$

$$\exp\left[-s^*\left(r\right)d\right] \leq 1-\alpha$$

$$r \geq E(u)$$

$$d \geq 0.$$

Clearly, the optimal solution to BCP will result in restrictive values for (d, r) due to our use of an upper bound on the loss probability.

Theorem 7 gives the optimal solution to BCP, and designates Z^* as the optimal cost.

Theorem 7 *The solution to problem BCP is given by:*

$$r^* = \frac{\text{LambertW}\left(s^*\left(r^*\right)G_u'\left(s^*\left(r^*\right)\right)e^C\right) - C}{s^*\left(r^*\right)} \tag{4.6}$$

$$d^* = -\frac{\ln\left(1-\alpha\right)}{s^*\left(r^*\right)} \tag{4.7}$$

$$Z^* = R\frac{\text{LambertW}\left(s^*\left(r^*\right)G_u'\left(s^*\left(r^*\right)\right)e^C\right)}{s^*\left(r^*\right)}, \tag{4.8}$$

where the LambertW (x) *is a function that satisfies:*

$$\text{LambertW}\left(x\right)e^{\text{LambertW}(x)} = x,$$

and is discussed in (Corless et al., 1996) and (Corless et al., 1997), as well as in Appendix A.6. The solution is on the principal (zero) branch of the LambertW (x)*, and*

$$C \equiv -D\ln\left(1-\alpha\right)/R. \tag{4.9}$$

The steps to the proof of Theorem 7 are as follows. We show that the two non-negativity constraints (on d and on $E\left(x\right)$) of BCP are never binding, in the presence of the bound on the

service level constraint. Thus, a less restrictive problem is being solved. Moreover, since this problem has only one constraint that is shown to be active, its solution can be calculated by substituting one of the variables for the other. Since r is present in the service level constraint in an implicit form, we choose to present the optimal d as a function of r. Using Theorem 6, the resulting problem is shown to be convex. Then we solve the first-order conditions of a non constrained optimization problem, and show that their solution is given by (4.6) and (4.7), as has been claimed.

Proof. of Theorem 7.

In the presence of the service level constraint, the rate constraint is never binding.[4] This holds true, since $s^*(r) \to 0$ as $r \to E(u)$. Thus, satisfying the service level constraint requires $d \to \infty$, which will never be optimal when d is costly. It follows that the constraint $r \geq E(u)$ is redundant.

Taking the log of both sides of the approximated service level constraint, and emphasizing the dependency of s^* on r, we get the next problem:

$$\min_{d,r} Dd + Rr$$

$$s.t$$

$$-s^*(r)\, d \leq \ln(1-\alpha).$$

Moreover, since both r and d are costly, the service level constraint will be active[5] and one can transform this problem to a non-constrained problem with one variable, using

$$d = -\ln(1-\alpha)/s^*(r). \tag{4.10}$$

To simplify the notation, we suppress the dependency of s^* on r and the superscript $*$. We also define $C \equiv -D\ln(1-\alpha)/R$ and note that $C > 0$. The buyer's problem is transformed

[4] Another way to show that the non-negativity constraints are redundant in BCP, is to ignore them during the solution of BCP and then to show that the solution always satisfies them. From (4.7), it is clear that $d^* > 0$ (since $\ln(1-\alpha)$ is negative). However, showing that $r^* > E(u)$ from (4.6) is more demanding.

[5] This follows straightforwardly from the fact that the objective function tries to decrease both r and d to $-\infty$, which is prevented since s^* is increasing in r and s^*d should stay positive in order for the service level constraint to hold.

to:

$$\begin{aligned} \min_r Z(r) &= \min_r Rr - D\ln(1-\alpha)/s \\ Z(r) &= Rr + RC/s. \end{aligned} \tag{4.11}$$

Before solving (4.11), we note that due to Part B of Theorem 6, this problem is convex. Thus, showing that (4.6) and (4.7) satisfy the first-order conditions is enough to prove Theorem 7.

Now, taking the derivative of Z and comparing it to zero, we get:

$$R - \frac{s^{'}CR}{s^2} = 0.$$

Dividing by R and substituting $s' = \frac{se^{sr}}{\left(G_u'(s) - re^{sr}\right)}$ using (4.3), from Theorem 6, we get:

$$\begin{aligned} 0 &= 1 - \frac{e^{sr}C}{s\left(G_u'(s) - re^{sr}\right)} \\ 0 &= sG_u'(s) - sre^{sr} - e^{sr}C. \end{aligned}$$

After rearranging terms, the optimal r should satisfy

$$r - e^{-sr}G_u'(s) + C/s = 0, \tag{4.12}$$

and solving for r we get:

$$r^* = \frac{\text{LambertW}\left(sG_u'(s)\,e^C\right) - C}{s}.$$

The principal (zero) branch of the $LambertW(x)$ is needed because the arguments within the function are all positive, and this branch is the only one to have real solutions for positive arguments of the function. (See page 17 of (Corless et al., 1996).) The $LambertW(x)$ function is briefly discussed in Appendix A.6. Substituting (4.6) into (4.12) and using the definition of the $LambertW$ function shows that (4.6) and (4.7) are the optimal solution for BCP. The cost of this solution is given in (4.8), concluding the proof. ■

We notice that the optimal solution of the Buyer's Constrained Problem (BCP) depends, as intuition suggests, on the ratio D/R and not on their actual values. Moreover, the optimal solution depends on the service level via a multiplication by the log of one minus the loss probability. Both these dependencies are captured in the constant C.

Let the reader note that the solution to the buyer's problem will be given by Theorem 7 as long as its resulting cost (4.8) is lower than the cost of the trivial solution discussed in subsection 4.2. We recall that solution here:

$$\begin{aligned} r^* &= F_U^{-1}(\alpha) \\ d^* &= 0 \\ Z^* &= RF_U^{-1}(\alpha). \end{aligned}$$

The cost of this solution is lower than the cost in (4.8) if

$$F_U^{-1}(\alpha) \leq \frac{\text{LambertW}\left(s^*(r^*)\, G_u'(s^*(r^*))\, e^C\right)}{s^*(r^*)}. \tag{4.13}$$

If this trivial solution is chosen it is the **optimal** solution to the buyer's problem, since it requires no approximations.

It is important that the reader be aware that (4.6) does not give a closed-form solution for r since s^* is a function of r. However, (4.6) gives a clear representation of the requirements that a solution of BCP be optimal, and can be solved numerically. Furthermore, we emphasize that the results of Theorem 7 bound the cost of the buyer's problem under both pricing schemes.

Finally, incorporating the original upper bound developed in Chapter 3 (for the loss probability of the two-sided regulated random walk) for the TB case requires additional demand information. Thus, we defer the use of this bound to the next section, where we demonstrate its use for both the normal and exponential demand cases.

4.3.2 Lower Bounds on the Buyer's Cost

Using the Loose Lower Bound

Using the loose lower bound for the service level, a lower bound on the buyer's expenditures can be found by solving the following relaxed problem:

$$\begin{gathered} \min_{d,r} Dd + Rr \\ s.t \\ \overline{F}_U(r+d) \leq 1-\alpha \\ r \geq E(u) \\ d \geq 0. \end{gathered} \tag{4.14}$$

The solution to this problem is simple, since the service level constraint is symmetric regarding the decision variables. That is, for each unit increase in r or d, the increase in the right-hand side of (4.14) constraint is the same. However, since $R > D$ (as discussed in the introduction to this chapter), it is better to increase d rather than r. Thus, the optimal solution to this problem is

$$\begin{aligned} r_D &= E(u) \\ d_D &= F_U^{-1}(\alpha) - E(u) \\ Z_D &= (R-D)E(u) + DF_U^{-1}(\alpha). \end{aligned}$$

Note that, since we assume that in order to satisfy the required service level $r \geq E(u)$, we expect (or confine the service levels to be such that) $F_U^{-1}(\alpha) > E(u)$ as well. Again, this solution is applicable to both the TB and the TBwRC cases.

Using the Lower Bound

Using the lower bound for the service level, a lower bound on the buyer's expenditures can be found by solving the following relaxed problem:

$$\min_{d,r} Dd + Rr$$

$$s.t$$

$$(1-\alpha)\bar{F}_U(r) + \alpha\bar{F}_U(r+d) \leq 1-\alpha$$

$$r \geq E(u)$$

$$d \geq 0.$$

There are three possible solutions to this problem that can be found using Lagrange multipliers:

1. A trivial solution $r = F_u^{-1}(\alpha)$, $d = 0$, and a cost $Z = RF_u^{-1}(\alpha)$. This solution is not useful, since when $d = 0$, the bound gives the loss probability in the TB case accurately. Thus, if this is the optimal solution to the relaxed problem, it is also the optimal solution for the original problem, as explained in subsection 4.2.2 (and then a bound is not necessary since we know the optimal solution).

2. A zero drift solution $r = E(u)$, $d = F_u^{-1}\left(1 - \frac{1-\alpha}{\alpha}F_u(E(u))\right)$, and a cost $Z = RE(u) + DF_u^{-1}\left(1 - \frac{1-\alpha}{\alpha}F_u(E(u))\right)$.

3. A tight service level constraint, in which r and d are calculated from:

$$\begin{aligned} \frac{D(1-\alpha)}{\alpha(R-D)} &= \frac{f_u(r+d)}{f_u(r)} \\ (1-\alpha)\bar{F}_u(r) + \alpha\bar{F}_u(r+d) &= 1-\alpha. \end{aligned} \tag{4.15}$$

   Unfortunately, we could not find a closed-form expression for the cost of this solution for a general distribution. However, it is, in fact, the best lower bound for the buyer's problem when the service level is high.

4.4 Examples

This section presents the analysis of the buyer's decision in the cases of normal and exponential demand. For both cases, we get closed-form solutions for the values of r and d when solving the buyer's problem in the case of a TBwRC. In the exponential demand case, this solution is the theoretical optimal due to remark 3 following Lemma 3 in Chapter 3. For the TB case, numerical simulation results are provided and show that the proposed approximations work well.

4.4.1 The Normal Demand Case

As mentioned in the discussion of the demand model in subsection 3.2.1 in Chapter 3, the normal usage is a good approximation to the aggregate usage of computer power. We consider cases for which the ratio between the mean demand and its standard deviation is such that the probability of negative usage is negligible.

In this subsection we give an analytic analysis of the buyer's problem and then present numerical results of a simulation, showing that the heuristic developed works well.

Upper Bounds when Using the Token Bucket with Rate Control

In the case of normal demand, closed-form approximations to the buyer's problem can be calculated based on the different parts of Lemma 3. The corresponding solutions are given here, and we denote them r_i^*, d_i^* where $i = B$ for the upper **B**ound based on part A of the lemma, and $i = A$ for the **A**pproximation based on part B of the lemma (given in the next subsection).

First, from the bound in part A of the lemma, we can replace the service level constraint by:

$$\exp\left[-s^*\left(r\right)d\right] \leq 1-\alpha.$$

Assuming this constraint is active, we can calculate $d(r)$. Replacing the service level constraint with this bound leads to the following solution to BCP, which proof is given in Appendix A.7:

Proposition 8 *If the buyer's demand in each sub-period is iid and normally distributed:* $u \sim$

Normal $\left(\mu, \sigma^2\right)$*, then the solution to BCP is given by:*

$$r_B^* = \mu + \frac{\sigma\sqrt{2C}}{2} \tag{4.16}$$

$$d_B^* = -\frac{\sigma \ln(1-\alpha)}{\sqrt{2C}} \tag{4.17}$$

$$Z_B^* = R\left(\mu + \sigma\sqrt{2C}\right) \tag{4.18}$$

where $C = -D\ln(1-\alpha)/R$*, as in (4.9).*

Based on Part B of Lemma 3: Based on the approximation in part B of Lemma 3 in Chapter 3, another approximate solution to the buyer's problem, in case of normal demand, can be obtained. The threshold crossing probability for a random walk with a negative drift and a normal distribution with expectation $\mu - r$ and standard deviation σ can be approximated using the technique of (Siegmund, 1985) (which is briefly discussed in Appendix 3.4.3):

$$P(S_N \geq d) = e^{-\frac{2(r-\mu)}{\sigma^2}d} V\left(\frac{2(r-\mu)}{\sigma}\right) \approx e^{-\frac{2(r-\mu)}{\sigma^2}d} e^{-0.583\left(\frac{2(r-\mu)}{\sigma}\right)}, \tag{4.19}$$

where the function $V(\mu) \equiv 2\mu^{-2}\exp\left(-2\sum_1^\infty n^{-1}\Phi\left(-0.5\left|\mu\right|\right)n^{0.5}\right)$, in which $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Moreover, $V(\mu)$ can be approximated with $\exp(-0.583\mu)$. This approximation is more accurate as $r-\mu$ is smaller (i.e. R/D is small). Replacing (4.19) for the service level constraint in the buyer's problem leads to the following approximation solution to BCP, which proof is given in Appendix A.7:

Proposition 9 *If the buyer's demand in each sub-period is iid and normally distributed* $u \sim$ *Normal* $\left(\mu, \sigma^2\right)$*, then an approximate solution to the buyer's problem is*

$$\begin{aligned} r_A^* &= \mu + \frac{\sigma\sqrt{2C}}{2} \\ d_A^* &= -\frac{\sigma\ln(1-\alpha)}{\sqrt{2C}} - 0.583\sigma \\ Z_A^* &= R\left(\mu + \sigma\sqrt{2C}\right) - 0.583D\sigma, \end{aligned}$$

where $C \equiv -D\ln(1-\alpha)/R$ *(as in 4.9).*

Note that the difference between this solution and the one given in Proposition 8 is the correction factor: -0.583σ in the depth parameter. Due to this correction factor, the depth value here is smaller than in Proposition 8; moreover, the resulting depth might be negative (when depth is much more expensive than rate), which is infeasible for the original problem. In such cases the approximate solution to the buyer's problem is the one for the case $d = 0$.

Note that since the above procedure uses an approximation for the loss probability of the shortfall process, the resulting service level might be lower than the requested one. However, the numerical results presented in the next subsection show that the above approximation is very accurate for the service level in the TBwRC case and, therefore, these results still provide the service level required for the bucket level process. Furthermore, this solution is (almost) indistinguishable from the optimal solution (the difference in cost is always less than one percent) to the buyer's problem when facing the TBwRC.

The insights provided from the two closed-form solutions in Propositions 8 and 9 are the same. First, when service level increases, the rate and depth parameters (and the cost) are increasing, with a rate that depends on $\ln(1-\alpha)$. Second, the optimal token rate (depth) increases (decreases) with the ratio D/R. Finally, and most importantly, the rate linearly increases with the mean demand and with the standard deviation (with a slope $\sqrt{2C}/2$), and the depth linearly increases with the standard deviation (with different slopes).

These properties also hold in the exponential demand example that we discuss in the next subsection. Moreover, the linear dependency on the demand parameters can be used to update the rate and depth parameters when demand increases (from one contract to another), and it can also be helpful when looking for a numerical solution according to Proposition 7, in cases of general demand.

Upper Bound when Using the Token Bucket

We use here the results developed in Chapter 3 to bound the loss probability for the two-sided regulated random walk. This bound states that when demand is normal, we can bound this loss probability by a constant times the loss probability in the TBwRC (one-sided regulated

random walk) as follows:

$$\begin{aligned} P\left(\tilde{L}=0\right) &\lesssim \frac{1+\bar{F}_u(r)\left[\frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}-1\right]}{\frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}-1}P\left(L\leq 0\right) \\ &= A\cdot P\left(L\leq 0\right), \end{aligned}$$

where $\bar{F}_u\left(r\right)=1-F_u\left(r\right)$ is the probability that the usage will be higher than r and where we introduce A to correspond to the expression before $P\left(L\leq 0\right)$, for notational convenience. Combining this with the bounds on the service level in the TBwRC, as given in (4.19), we get:

$$\begin{aligned} P\left(\tilde{L}=0\right) &\lesssim \frac{1+\bar{F}_u(r)\left[\frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}-1\right]}{\frac{\sigma}{(r-u)^2\sqrt{2\exp\left(-\frac{1.166(r-u)}{\sigma}\right)}}-1}e^{-\frac{2(r-\mu)}{\sigma^2}d}e^{-0.583\left(2\frac{r-\mu}{\sigma}\right)} \\ &= A\exp\left(-0.583\left(2\frac{r-\mu}{\sigma}\right)\right)e^{-\frac{2(r-\mu)}{\sigma^2}d}. \end{aligned} \tag{4.20}$$

However, when demand is normal the value of $F_u\left(r\right)$ can be found only numerically; thus, the buyer's problem cannot be solved in general using this enhancement. We therefore propose a **Better Bound Approximation**. First, find the optimal rate to the buyer's problem using Proposition 9 and calculate the constant A given this rate. Second, if A is not between zero and one,[6] take the solution of Proposition 9; otherwise, use the solution obtained from Proposition 9 for a problem with a relaxed service level constraint:

$$P\left(\tilde{L}=0\right)\leq\left(1-\alpha\right)/A.$$

[6]This can happen if our approximation for the expected subcycle length results in a number smaller than 1, which might happen when the approximations are inaccurate. In practice this happens for high service levels, where the expected subcycle length is very close to 1, due to the strong positive drift of the bucket process.

Loose Lower Bound

In the normal demand case the exact computation of the lower bound is not simple; thus, we use the loose lower bound:

$$\begin{aligned} r_L &= \mu \\ d_D &= \sigma\Phi_U^{-1}(\alpha) - \mu \\ Z_D &= (R-D)\,\mu + D\sigma\Phi_U^{-1}(\alpha), \end{aligned}$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal random variable.

Numerical Results

In order to check the performance of the methods developed in the subsections above, we ran an extensive simulation with a wide range of parameters and compared the losses of the solutions based on these heuristics to the optimal choice (found using the simulation). We used a normal demand with mean 10 and standard deviation of one, two, or three, and considered D/R cost ratios of 0.9, 0.5, 0.2, 0.1 (recall that if D>R, there is no point in purchasing depth). We ran a simulation of 500,000 periods with service level requirements of 80%, 90%, 95%, and 99%.

The results of our numerical study show that for the TB case the approximation based on the tight bound (developed in subsection 4.4.1) is very good and the costs are typically within 2% of the optimal ones. In order to find the optimal parameters choice, we ran an exhaustive search, looking for a combination of d and r that provided a loss probability that was higher, up to 0.2% from the requested one. Furthermore, using the loose lower bound, we can show analytically that the cost of the approximation based on the rate control bound (Proposition 9) is within 23% of the loose lower bound, and therefore of the optimal.

Table 4-3 is an example of the results we gathered for each service level, D/R cost ratio, and standard deviation pair. It presents the results for the case of a service level of 99%, $D/R = 0.1$, and standard deviation 1. The table compares the performances of the upper bound (Proposition 8), the rate control bound (Proposition 9), the approximation based on the tight bound (subsection 3.4.2, in Chapter 3), the optimal choice (based on our simulation), and the (loose) lower bound (subsection 4.4.1). For each such method, the table presents the

Performance comparison of bounds and optimal choice for a service of 99%, normal (10,1) demand, and D/R=0.1

D/R=0.1	Rate	Depth	Cost	PBacklog	PLoss	Error
Upper Bound	10.48	4.80	10.96	0.59%	0.32%	1.30%
Rate Control	10.48	4.22	10.90	0.95%	0.95%	0.76%
Tight Upper Bound	10.46	3.98	10.85	1.56%	1.56%	0.33%
Lower Bound	10.00	2.33	10.23	99.37%	20.31%	-5.42%
Optimal	10.38	4.44	10.82	1.01%	99.37%	0.00%

Figure 4-3: An example of the results for each service level, D/R cost ratio, and standard deviation pair for the case of service level 99%, $D/R = 0.1$, and standard deviation=1.

Rate (r), Depth (d), Cost (Z), loss probability in the TBwRC process (PLoss), loss probability in the bucket level process (PBacklog), and the error (Error) of the solution (defined as the difference between the optimal cost and the solution's cost, divided by the optimal cost). The table shows that the service level decreases and the cost increases as we go from the weakest upper bound to the loose lower bound.

These results also show that the loss probability in the TBwRC process predicted by the rate control bound is very accurate. Therefore, the differences in cost between the optimal choice for the TBwRC (that are not reported here) and the results of the rate control bound (Proposition 9) are negligible (less than one percent).

The cost errors of the tight bound for service levels of 99%, 95%, 90%, and 80% are lower than 2%, 3.5%, 5%, and 8%, respectively. In fact for a lower standard deviation, the typical difference was smaller, as seen in part (a) of Figure 4-4, which compares the efficiency of the tight bound solution for a service level of 80% along different D/R cost ratios. The results in this figure are typical of higher service levels as well. Thus, it is clear that the efficiency of the tight bound approximation does not depend on the ratio D/R. However, when this ratio is high, the tight bound does not improve the results from the rate control bound (since our approximation for A, as defined in subsection 3.4.2, is not between zero and one).

(a) **(b)**

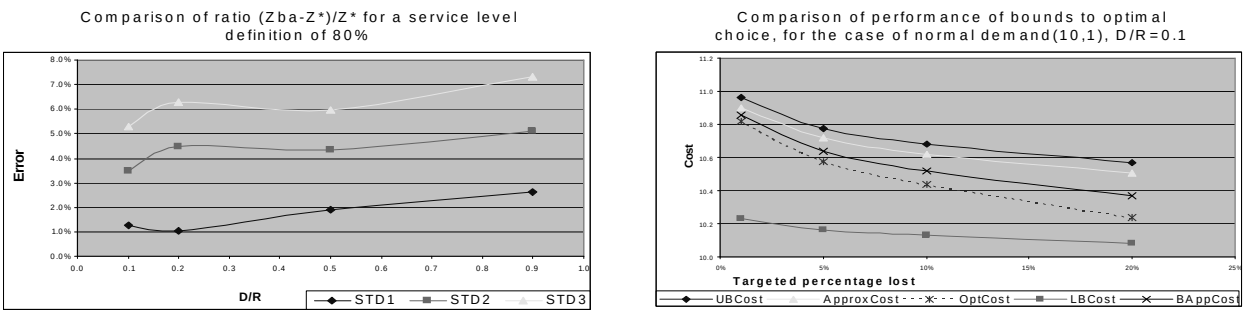
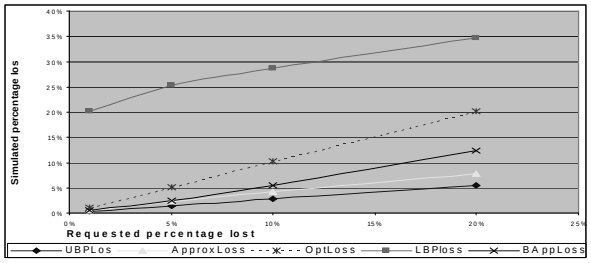


Figure 4-4: Part (a) –Upper bound's error (comparing to optimal results) when variance changes for a service level of 80%. Part (b) –Comparison of the perfomance of bounds to the optimal choice, for the case of normal demand (10,1) and D/R=0.1

Part (b) of Figure 4-4 shows the costs vs. targeted percentages lost for the five different approaches. All of the upper bounds improve as the service level increases. This improvement is a general property of the large deviation theory, which gives better bounds for lower loss probabilities. Moreover, since the difference between the two bounds for the TBwRC is in a fixed bucket level, the cost difference between them is fixed. As remarked earlier, if we add the graph of the optimal solution for the TBwRC, it would be (almost) indistinguishable from the graph of the bound from Proposition 9.

Part (a) of Figure 4-5 compares the simulated loss percentage with the requested ones for the case of normal demand with standard deviation 2 and $D/R = 0.1$. The most important observation here is that there is a linear relation between the simulated loss probabilities of the upper bounds and the requested ones. However, we do not have a theoretical justification for this observation. If this relation holds, buyers can use the following procedure to obtain the optimal solution to their problem: First, solve the approximated problem using one of the upper bounds. Second, run a single simulation with the resulting r and d, and plot the comparison between the simulated loss probabilities and the requested ones. Third, find the service level that should be allowed by this upper bound and which will yield the requested service level. Finally, solving the approximated buyer's problem with this new service level will result in the optimal solution to the buyer's problem.

(a)



(b)

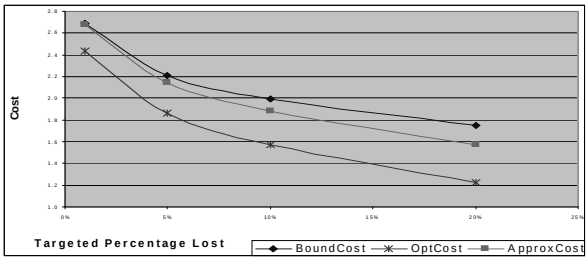


Figure 4-5: Part (a) - Comparison of targeted and simulated loss performance for different methods normal demand, standard deviation 2 and D/R=0.1. Part (b) - Comparison of the performance of the bound and the heuristic to the optimal choice, for the case of exponential demand with mean 1 and D/R ratio is 0.2

4.4.2 The Exponential Demand Case

The analysis of the exponential demand is given here since, in this case, a closed-form solution to the loss probability in the shortfall process can be calculated, based on Parts B or C of Lemma 3 (in Chapter 3). Thus, the bound of this lemma gives the actual loss probability. Therefore, a closed-form solution to the buyer's problem when facing the TBwRC is given in Proposition 10. In addition, based on Corollary 14 in Chapter 5, for the exponential case, the fill rate and the percentages of periods with losses are identical; thus, the following solutions also hold for the fill rate service level. Moreover, the solution method here shows how a closed-form solution (or approximations) to the buyer's problems can be found even when $s(r)$ cannot be expressed in closed form.

We combine the equations for $s(r)$ as the positive solution to $G_x(s) = 1$, and of the first-order conditions of the optimization problem. These two equations with unknowns d and s, can then be solved, leading to a closed-form solution to BCP in this demand case as well.

Optimal Solution When Using the Token Bucket with Rate Control

In the case of exponential demand, the exact solution to the coefficient of the probability of backlog (for the TBwRC process), presented in Lemma 3, is given by: $1 - s/\lambda$ (see remark 3 after Lemma 3). The reader should be aware that since this expression is accurate, in the

exponential demand case, it results in the optimal solution for the case of a TBwRC and a theoretically proven upper bound for the service level of the bucket process.

We rewrite the service level constraint, with the exact expression for the backlog probability when demand is exponential:

$$\frac{\lambda - s^*}{\lambda} \exp\left[-s^* d\right] \leq 1 - \alpha. \tag{4.21}$$

Assuming this constraint is active, we can substitute $d(r)$, based on the service level requirement. This substitution leads to a closed-form solution to BCP and to the buyer's problem, which proof is given in Appendix A.7:

Proposition 10 *If buyer's demand in each sub-period is iid and exponentially distributed:* $u \sim Exp\,(\lambda)$*, then the solution to the buyer's problem in the token bucket with rate control case is given by:*

$$\begin{aligned} r^* &= \frac{-\ln(1-\alpha))}{\lambda} \\ d^* &= 0 \\ Z^* &= R\frac{-\ln(1-\alpha))}{\lambda} \end{aligned}$$

if

$$\alpha \leq \frac{LambertW(\beta)+1}{LambertW(\beta)} \tag{4.22}$$

and by

$$\begin{aligned} r^* &= \frac{LambertW(\beta)}{\lambda}\left(\frac{\ln(1-\alpha)}{K\,(LambertW(\beta)+1)} - 1\right) \\ d^* &= \frac{LambertW(\beta)\left[LambertW(\beta) - \ln\left(-\beta(1-\alpha)\right)\right]}{\lambda\,(LambertW(\beta)+1)} \\ Z^* &= R\frac{LambertW(\beta)}{\lambda}\left(\frac{\ln(1-\alpha)}{K\,(LambertW(\beta)+1)} - 1\right) \\ &\quad + D\frac{LambertW(\beta)\left[LambertW(\beta) - \ln\left(-\beta(1-\alpha)\right)\right]}{\lambda\,(LambertW(\beta)+1)} \end{aligned}$$

otherwise. The -1 *branch of the* $\mathrm{LambertW}\,(x)$ *(discussed in Appendix A.6) should be used in the above expressions,*

$$K \equiv R/D - 1, \tag{4.23}$$

and

$$\beta \equiv \left(-\frac{(1-\alpha)^{1/K}}{e} \right). \tag{4.24}$$

The insights from this closed-form solution are the same, although with different slope coefficients, as the ones we got from the solutions for the normal demand case, i.e., the rate linearly increases with the mean demand and with the standard deviation, and the depth linearly increases with the standard deviation.

Upper Bound When Using the Token Bucket

We use here the results developed in Chapter 3 to bound the loss probability for the two-sided regulated random walk. This bound states that when demand is exponential, we can bound this loss probability by a constant times the loss probability in the TBwRC (one-sided regulated random walk), as follows:

$$\begin{aligned} P\left(\tilde{L}=0\right) &\lesssim \left(\frac{2\left(1-e^{-r\lambda}\right)\left[1+e^{r\lambda}\left(r\lambda-1\right)\right]\left(r\lambda-1\right)+r^2\lambda^2}{r^2\lambda^2 e^{r\lambda}-2\left[1+e^{r\lambda}\left(r\lambda-1\right)\right]\left(r\lambda-1\right)} \right) P\left(L\leq 0\right) \\ &= A\cdot P\left(L\leq 0\right), \end{aligned} \tag{4.25}$$

where we introduce A to correspond to the expression before $P\left(L\leq 0\right)$, for notational convenience. Combining (4.25) with the bounds on the service level in the TBwRC, as given in (4.21) we get:

$$\begin{aligned} P\left(\tilde{L}=0\right) &\lesssim \left(\frac{2\left(1-e^{-r\lambda}\right)\left[1+e^{r\lambda}\left(r\lambda-1\right)\right]\left(r\lambda-1\right)+r^2\lambda^2}{r^2\lambda^2 e^{r\lambda}-2\left[1+e^{r\lambda}\left(r\lambda-1\right)\right]\left(r\lambda-1\right)} \right) \frac{\lambda-s^*}{\lambda}\exp\left[-s^* d\right] \\ = \quad A\frac{\lambda-s^*}{\lambda}\exp\left[-s^* d\right]. \end{aligned}$$

Due to the complex A, the buyer's problem cannot be solved in general using this enhancement. We therefore propose a **Better Bound Approximation**: first, find the optimal rate to the buyer's problem using Proposition 10 and calculate A given this rate; now, if A is between zero and one, take the solution of Proposition 10; otherwise, use the solution obtained from

applying Proposition 10 to a problem with a relaxed service level constraint:

$$P\left(\tilde{L}=0\right) \leq (1-\alpha)/A.$$

Loose Lower Bound

When demand is exponentially distributed with mean λ, the solution is:

$$\begin{aligned} r_D &= 1/\lambda \\ d_D &= -\frac{\ln(1-\alpha)+1}{\lambda} \\ Z_D &= \frac{1}{\lambda}\left[(R-D)-D\ln(1-\alpha)\right], \end{aligned}$$

where $d>0$, when $\ln(1-\alpha)>-1$, for a service level higher than 0.6321, as we need.

Lower Bound

When demand is exponentially distributed with mean λ, the problem is:

$$\min_{d,r} Dd+Rr$$

$$s.t$$

$$(1-\alpha)e^{-\lambda r}+\alpha e^{-\lambda(r+d)} \leq 1-\alpha$$

$$r \geq 1/\lambda$$

$$d \geq 0.$$

First, assuming $d=0$, we get $r=-\frac{1}{\lambda}\ln(1-\alpha)$, and a cost $-\frac{R}{\lambda}\ln(1-\alpha)$, which is the optimal cost, as explained in Proposition 10. Then, assuming $r=1/\lambda$, we get $d=-\frac{1}{\lambda}\ln\left(\frac{(1-\alpha)(e-1)}{\alpha}\right)$, and a cost $\frac{1}{\lambda}\left[R-D\ln\left(\frac{(1-\alpha)(e-1)}{\alpha}\right)\right]$. Finally, we can find the third solution from (4.15):

$$\begin{aligned} \frac{D(1-\alpha)}{\alpha(R-D)} &= \frac{e^{-\lambda(r+d)}}{e^{-\lambda r}} \\ (1-\alpha)e^{-\lambda r}+\alpha e^{-\lambda(r+d)} &= 1-\alpha. \end{aligned}$$

The solution to these two equations leads to:

$$\begin{aligned} r_l &= -\frac{1}{\lambda}\ln\left(1-D/R\right) \\ d_l &= \frac{1}{\lambda}\ln\left[\frac{\alpha\left(R-D\right)}{D\left(1-\alpha\right)}\right] \\ Z_l &= -R\frac{\ln\left(1-D/R\right)}{\lambda}+D\frac{\ln\left[\frac{\alpha(R-D)}{D(1-\alpha)}\right]}{\lambda}. \end{aligned}$$

We noted this solution with l for the lower bound, since a comparison among the three solutions shows that for a high service level, it results in the lowest cost for that bound.[7]

Numerical Results

In order to check the performance of the bound and approximation developed in the subsections above, we ran an extensive simulation, using an exponential demand with mean one and the same parameters as in the normal demand cases. We compared the resulting losses in these two heuristics to the optimal choice (obtained using the simulation).

Part (b) of Figure 4-5 compares the optimal costs with the costs of the solution based on the upper bound, which is the optimal solution for the TBwRC (from Proposition 10) and the approximation resulting from the tight bound (in subsection 4.4.2), in the case of $D/R = 0.2$. These results show that the approximation suggested in subsection 4.4.2 improves the solution given by the optimal solution for the TBwRC.

4.5 Summary

This chapter analyzes the buyer's problem when facing the TB or the TBwRC pricing schemes. It uses the bounds and asymptotics developed in Chapter 3 to solve relaxed or constrained versions of the buyer's problem. The results of Chapter 3 helped in dealing with the first challenge of solving the buyer's problem: expressing the service level constraint.

An additional source of complexity in solving the buyer's problem follows from the buyer's controls, which are both the rate and the depth. Alternatively, using the random walk termi-

[7] For a D/R ratio of $\left(1-e^{-1}\right) = 0.632\,12$ the second solution reaches its lowest cost, and only at this point its cost is identical to the cost of the third solution (independent of the service level).

nology, the buyer's controls are both the drift and the threshold not to be crossed by the random walks. This chapter has dealt with this challenge. We used the lower and upper bounds on the service level provided to a buyer who faces one of TB pricing schemes, to provide bounds on the costs of the buyer's problem for a general demand process. In order to provide such bounds, we show that a constrained version of the buyer's problem is convex. This proof of convexity teaches us two important properties of a conjugate point: it increases with the drift of the random walk and its reciprocal is convex with the rate. These properties are summarized in Theorem 7 and are of interest in their own right. Furthermore, we demonstrated the above solutions and the proposed approximations for exponential (normal) demand process, leading to closed-form solutions (approximation) of the buyer's problem when facing the TBwRC. In both cases, the rate linearly increases with both the expectation and the standard deviation of the demand, and the depth linearly increases with demand's standard deviation. These results make TB schemes more attractive for practical purposes, when possibly only the first two moments of the demand are known and a heuristic search for the optimal rate and depth is performed. This linear relation should also guide the numerical search for a solution, as given in Proposition 7. We showed the efficiency of our approximations and bounds, both analytically and numerically. The computational experiment suggests that in the TB case for high service levels, the proposed approximation leads to a solution that is within 8% of the optimal one, and we showed analytically that the costs are within 23% of the costs of the loose bound (in the normal demand case). For the TBwRC and exponential demand, the solution proposed here is optimal, and for the normal demand case it is within 1% of optimal.

To the best of our knowledge, the most similar work to this one is (Glasserman, 1997). However, he focuses on solving the problem using only the threshold as a control. Moreover, (Glasserman, 1997) confined his work to the backlog case (the TBwRC); he did not treat the two-sided regulated random walk case (TB).

In the next chapter we give approximations for the output from TB admission controls. This output, termed "effective demand", contributes one element of demand to the aggregated demand faced by a seller who uses TB pricing schemes.

Chapter 5

The Output from Token Bucket Admission Controls

5.1 Introduction

We have characterized the output stream from token-bucket and token-bucket-with-rate-control admission control mechanisms. This is the buyer's effective demand, u^e. We recall that this demand contributes one element to the aggregated demand the seller faces, and therefore, it is a required input to the seller's problem when using the TB pricing scheme. We assume that the buyer's demand distribution is known, and develop the cumulative distribution function (CDF) and probability distribution function (PDF) for the output of TB (with or without rate control) admission controls.

Naturally, the effective demand is a function of the buyer's demand and the bucket level distributions. Fortunately, these two distributions are independent, since demand is iid and the bucket level at the beginning of period i is independent of the demand in this period. We use this fact extensively in the rest of this chapter, without repeating it. We approximate the bucket level distribution using a Brownian motion since it has been shown to behave as a two-sided or one-sided regulated random walk. We discussed the use of this approximation in Section 3.5. This allows us to approximate the density of the effective demand. However, the Brownian approximation ignores the probability of the random walk to be on its boundaries, therefore we also present an enhancement for the approximation based on Brownian motion.

This enhancement combines known results from large deviations theory and Brownian motion analysis. The enhancement is based on a relation between the fill rate –percentage of units of work (or tokens) that are being processed– to the percentage-of-periods-with-losses of units of work (or tokens). To relate the two different service level measures, we provide results that approximate the expected number of units of work (or tokens) lost in any period that shows a loss of work (or tokens).

We summarize our analysis in four different algorithms to approximate the effective demand in the case of a TB or TBwRC. We consider these algorithms as the main contributions of this chapter.

Section 5.2 presents the analysis of the effective demand resulting from the TB admission control. This analysis is much simpler than the one for the case of a TBwRC admission control that is presented in section 5.3. Both sections first present the general analysis of the PDF and the CDF as a function of the buyer's demand and the bucket level distributions. Then, both chapters present results based on the Brownian motion approximation and end with its enhancement. Finally, Section 5.4 presents numerical results for both admission control, in cases where the buyer's demand is normally or exponentially distributed. These results include comparison of the CDF based on the different approximations to the one resulting from a simulation, and a similar comparison between the first moment and standard deviation of the simulation to the ones calculated based on the approximations. The errors obtained from the approximation of the moments are less than 2% for the TB case and 5% for the TBwRC cases. Typical results are within the range of 1% from the actual moments.

5.1.1 Notation

We use the following notation:

Buyer's demand u, is drawn from a CDF $F_U(u)$, with a Moment Generating Function (MGF) $G_U(s) \equiv \int_{\Omega(u)} e^{-su} dF_U(u)$. Buyer's demand at the i^{th} period will be noted using the subscript i. We assume that the demand is nonnegative and continuous (with a possible positive probability at zero).

Effective demand u^e, the output stream from the admission control, is denoted by use of the superscript e. Thus, we are looking for $F_{U^e}(u^e)$, $f_{U^e}(u^e)$, or $G_{U^e}(s)$ in two different cases.

The first case is of the TB mechanism, and the second one is of the TBwRC mechanism. We treat these two cases in different sections; therefore, we use the same notation for the effective demand in both cases.

The token rate r, and the bucket depth d. Both are assumed to be known.

Bucket level process at steady state, without rate control $\widetilde{L}$, and with rate control L. Their distribution is given by $F_{\widetilde{L}}\left(\widetilde{l}\right)$, and $F_L(l)$, respectively.

The a to b partial MGF of a probability distribution with CDF $F_X(x)$ is defined as: $G_{a\leq X<b}(s) \equiv \int_a^{b^-} e^{-su} dF_U(u)$. Note also that it exists for any X that has a regular MGF.

The delta function δ is defined such that an integration of a density at a point with the delta function is equal to one.

Partial expectation, $E(x; x \in [a,b])$, is the expectation of the random variable x taken on the range $[a,b]$.

The CDF of a standard normal distribution evaluated at the point x, is noted as: $\Phi(x)$, and the PDF of a standard normal distribution evaluated at the point x,is noted: $\phi(x)$.

The value "just before x", i.e. $\lim_{\epsilon\to 0}(x-\epsilon) \equiv x^-$. This will be used when integrations are taken until just before x.

5.2 Token Bucket

We note again that in the TB case, the evolution of the bucket level process $\widetilde{L}$ is described by:

$$\widetilde{L}_{i+1} = \min\left\{d, \max\left[0, \widetilde{L}_i + r - u_i\right]\right\},$$

which is equivalent to the evolution of a two-sided regulated random walk (with regulators at zero and d).

5.2.1 The Probabilistic Description of the Effective Demand

In the case of a TB, we can find the PDF of the effective demand by conditioning on the bucket level. Given a bucket level $\widetilde{l}_i$ at period i, the effective demand in this period is:

$$u_i^e = \begin{cases} u_i & if\ u_i < \widetilde{l}_i + r \\ \widetilde{l}_i + r & if\ u_i \geq \widetilde{l}_i + r \end{cases} \tag{5.1}$$

The top part of (5.1) states that when the buyer's demand is lower than the number of available tokens (sum of the bucket level at the beginning of the period $\widetilde{l}_i$ and the token rate r) this demand will be fully satisfied. In this case, the effective demand will equal the buyer's demand. The bottom part of (5.1) states that if the buyer's demand is higher than the number of available tokens, then the effective demand will equal this number (i.e., the sum of the bucket level and the token rate). We use the steady state bucket distribution to write the probability density function (PDF) of the effective demand as:

$$f_{U^e}(u^e) = f_U(u^e)\left[1 - F_{\widetilde{L}}(u^e - r)\right] + f_{\widetilde{L}}(u^e - r)\left[1 - F_U(u^e)\right]. \tag{5.2}$$

The interpretation of (5.2) is similar to the interpretation of (5.1). The effective demand is u^e in one of two disjoint events. The first one is when the buyer's demand equals u^e and there are more than $u^e - r$ tokens in the bucket; then, all of the demand is admitted and the effective demand is equal to the requested one. The second case is when the demand is higher than or equal to u^e, but the bucket level is $(u^e - r)$; then, the effective demand is the bucket level plus the token rate (which is lower than u). It is important that the reader be aware that we use $f_{\widetilde{L}}(u^e - r)$, which includes two delta functions, one at $\widetilde{L} = 0$, and the other at $\widetilde{L} = d$. Finally, note that the effective demand is always in the range $[0, d + r]$, since buyer's demand is assumed to be non-negative, and $F_{\widetilde{L}}(u^e - r)$ equals one for $u \geq r + d$ and zero for $u < r$.

Equation (5.2) is central for our analysis, since it summarizes our novel approach. Traditionally, trials to characterize the output stream of one- or two-sided regulated random walks modeled the system using either a queuing or a Brownian motion analogy. An extensive study of this nature is (Berger and Whitt, 1992). The disadvantages of queueing analogies are that they lead to closed-form solutions only for a limited range of distributions (such as the exponen-

tial one). The Brownian motion approach is more tractable in general; however, its drawback is that it catches only the first two moments of the random walk. Thus, the results from such approximations are not sensitive to higher moments of the arrival or service processes. In contrast, (5.2) hints that we can use a Brownian motion approximation to approximate the bucket level process; moreover, we can still consider the actual demand and arrival processes in this analysis. Indeed, as we show in the next sections, such an approach combines the advantages of the Brownian motion analogy with the ones of the queueing analogy; it does not ignore the higher moments of the arrival and departure processes, but is still tractable. Furthermore, in the numerical results section we show that in the TB (with or without rate control) cases, when demand is exponential or normal, this approach leads to fairly accurate characterizations of the random walk behavior, and its first two moments.

Although we have demonstrate the use of this technique for the bucket level, which is by nature a simple version of a random walk (due to the fixed rate), the methodology developed here is applicable under more general settings as well.

The reader can verify that (5.2) can also be written in a more detailed form as:

$$f_{U^e}(u^e) = \begin{cases} f_U(u^e) & \text{for } u^e \in [0, r) \\ f_U(r)\left[1 - P\left(\widetilde{L} = 0\right)\right] + [1 - F_U(r)]\, P\left(\widetilde{L} = 0\right)\delta & \text{for } u^e = r \\ f_U(u^e)\left[1 - F_{\widetilde{L}}(u^e - r)\right] + [1 - F_U(u^e)]\, f_{\widetilde{L}}(u^e - r) & \text{for } u^e \in (r, r + d) \\ [1 - F_U(r + d)]\, P\left(\widetilde{L} = d\right)\delta & \text{for } u^e = r + d \\ 0 & \text{otherwise} \end{cases}$$

In Appendix A.8, we use (5.2) and provide general expressions for the first and second moments of the effective-demand process resulting from the TB admission control.

Similar logic to the one that justifies (5.2) can be applied for computing the CDF of the effective demand based on the sample path behavior of the TB admission control:

$$\{u_i^e \le x\} = \{u_i \le x\} \cup \left\{\widetilde{L}_{i-1} \le x - r\right\},$$

which shows that the event of the effective demand at period i being lower or equal to x, is the union of the events that the buyer's demand at period i is lower or equal to x and the event that

the bucket level at the beginning of the i^{th} period $\widetilde{L}_{i-1}$ is lower or equal to $x-r$. Considering this description in steady state, we can write the CDF of the effective demand:

$$F_{U^e}\left(u^e\right) = F_U\left(u^e\right) + F_{\widetilde{L}}\left(u^e - r\right) - F_U\left(u^e\right) F_{\widetilde{L}}\left(u^e - r\right) \tag{5.3}$$

Clearly, (5.2) and (5.3) lead to each other by carefully integrating (5.2) or taking a derivative of (5.3).

From (5.2) and (5.3) we see that if $F_{\widetilde{L}}\left(l\right)$ were known, we could express $F_{U^e}\left(u^e\right)$ analytically. However, the distribution of the bucket level is *not* known. Therefore, in what follows we approximates this distribution.

The next subsections present two approximations for the CDF and PDF of the effective demand. The first one is based upon a Brownian motion approximation for the bucket's density and the second one is an enhancement of this approximation. This enhancement compensates for the differences between the continuous nature of the Brownian motion and the discrete nature of the random walk that represents the bucket level. To simplify the implementation of these methods, we summarize each method with an algorithm that implements its results. We demonstrate the use of these algorithms in the normal and exponential demand cases.

5.2.2 Approximating the Effective Demand Using Brownian Motion Techniques

In this section we approximate the bucket level's PDF, described by a two-sided regulated random walk, using the PDF of a Brownian motion. Based on this approximation we could later approximate the PDF and CDF of the effective demand. Algorithm 11, whose inputs are buyer's demand CDF $F_U\left(u\right)$ (with a mean $E\left(u\right)$ and standard deviation σ), and the buyer's choice of the TB parameters (bucket's depth d, and token rate r), summarizes the results of this section. We then demonstrate this algorithm in the normal and exponential demand cases.

Approximating the Probabilistic Description of the Bucket Level

As remarked earlier, the density of the bucket level process is not known. However, the density of a two-sided regulated Brownian motion, with a mean $r - E\left(u\right)$ and a standard deviation σ, is

known to be (see Proposition 5 in page 90 of (Harrison, 1985), and (Berger and Whitt, 1992)):

$$f_{\widetilde{L}}\left(\widetilde{l}\right)=\begin{cases}\frac{-\theta e^{\theta\widetilde{l}}}{1-e^{\theta d}} & \widetilde{l}\in[0,d]\\ 0 & \text{otherwise,}\end{cases} \tag{5.4}$$

where $\theta=2\left(r-E\left(u\right)\right)/\sigma^{2}>0$. The bucket level's CDF is:

$$F_{\widetilde{L}}\left(\widetilde{l}\right)=\begin{cases}0 & \widetilde{l}<0\\ \frac{1-e^{\theta\widetilde{l}}}{1-e^{\theta d}} & \widetilde{l}\in[0,d]\\ 1 & d<\widetilde{l}.\end{cases} \tag{5.5}$$

Approximating the Probabilistic Description of the Effective Demand

In the following, we suppress the superscript e from the value of the effective demand within formulas whenever no confusion arises. Then, using (5.4) and (5.5) as an approximation for the bucket level's PDF and CDF, we approximate the effective demand's PDF in the range $u^{e}\in[r,d+r]$, based on (5.2):

$$\begin{aligned}f_{U^{e}}\left(u\right) &\approx f_{U}\left(u\right)\left[1-\frac{1-e^{\theta(u-r)}}{1-e^{\theta d}}\right]+\left[1-F_{U}\left(u\right)\right]\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)}\\ &= f_{U}\left(u\right)\left[\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}}\right]+\left[1-F_{U}\left(u\right)\right]\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)},\end{aligned}$$

resulting in the detailed approximation for the PDF of the effective demand:

$$f_{U^{e}}\left(u\right)\approx\begin{cases}0 & \text{for } u<0\\ f_{U}\left(u\right) & \text{for } 0\leq u<r\\ f_{U}\left(u\right)\left[\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}}\right]+\left[1-F_{U}\left(u\right)\right]\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)} & \text{for } r\leq u\leq r+d\\ 0 & \text{for } d+r<u.\end{cases} \tag{5.6}$$

Similarly, we approximate the CDF, in the range $u^e \in [r, d+r]$ based on (5.3):

$$\begin{aligned} F_{U^e}(u) &\approx F_U(u) + \frac{1-e^{\theta(u-r)}}{1-e^{\theta d}} - F_U(u)\frac{1-e^{\theta(u-r)}}{1-e^{\theta d}} \\ &= \frac{1-e^{\theta(u-r)}}{1-e^{\theta d}} + F_U(u)\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}} \end{aligned}$$

resulting in the following approximation for the CDF:

$$F_{U^e}(u) \approx \begin{cases} 0 & \text{for } u < 0 \\ F_U(u) & \text{for } 0 \le u < r \\ \frac{1-e^{\theta(u-r)}}{1-e^{\theta d}} + F_U(u)\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}} & \text{for } r \le u < r+d \\ 1 & \text{for } d+r \le u \end{cases}. \tag{5.7}$$

Using (5.6) we can approximate the effective demand's MGF as:

$$\begin{aligned} G_{U^e}(s) &= \int_0^r e^{-su} f_U(u)\,du \\ &\quad + \int_r^{d+r} e^{-su}\left[f_U(u)\left[\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}}\right] + [1-F_U(u)]\frac{-\theta e^{\theta(u-r)}}{(1-e^{\theta d})}\right]du \\ &= G_{0\le U<r}(s) + \frac{1}{1-e^{\theta d}}\left[e^{-\theta r}G_{r\le U<d+r}(s-\theta) - e^{\theta d}G_{r\le U<d+r}(s)\right] \\ &\quad -\frac{\theta e^{-\theta r}}{(1-e^{\theta d})}\left[\frac{e^{-sr}\left(1-e^{-(s-\theta)d}\right)}{s-\theta} - \int_r^{d+r} e^{-(s-\theta)u}F_U(u)\,du\right]. \end{aligned}$$

It is clear that the approach we take here should not lead to a very good approximation of the effective demand since the approximation for the bucket level ignores the two delta functions of the bucket level distribution.

Moreover, since the bucket level has a positive drift, $P\left(\widetilde{L}=0\right) < P\left(\widetilde{L}=d\right)$. This means that the approximation based on the Brownian motion ignores more events of high demand (full bucket) than of low demand (empty bucket). Therefore, one might assume that the effective demand's first moment predicted using the Brownian motion approximation will be substantially lower than the actual mean (of the effective demand). However, ignoring the probability of a full bucket means ignoring the probability of effective demand to be equal to $d+r$. This event happens with probability $P\left(\widetilde{L}=d\right)[1-F_U(r+d)]$. In a similar manner,

ignoring the probability of an empty bucket means ignoring the probability of effective demand to be equal to r. This event happens with probability $P\left(\widetilde{L}=0\right)[1-F_U(r)]$. Thus, the positive drift effect from above is being balanced by the fact that $[1-F_u(r+d)]<[1-F_u(r)]$. Thus, the higher probability $P(L=d)$, which ignores high effective demand, ignores an event with an a priori lower probability $F_u(r+d)$, and the other way around. In the numerical results section we see that, in general, neither one of these effects is dominant.

However, in the next subsection we present an enhancement that approximates the delta functions and, indeed, the numerical results (in Section 5.4), show that the enhancement leads to better results.

Approximated Algorithm for the Effective Demand

The following algorithm summarizes the results in this subsection:

Algorithm 11 *Step 0 (Inputs and initialization): Inputs are: Buyer's demand CDF ($F_U(u)$, with a mean $E(u)$ and standard deviation σ), and the buyer's choice of bucket's depth and token rate (d,r, respectively). Let $\theta=\frac{r-E(u)}{\sigma}$.*
Step 1 (Approximate the bucket level probabilistic behavior): Approximate the bucket level PDF and CDF according to (5.4), and (5.5), respectively.
Step2 (Approximate the effective demand probabilistic behavior): Approximate the effective demand PDF and CDF according to (5.6) and (5.7), respectively. Moments can be calculated using the PDF.

Examples

Due to the simplicity of Algorithm 11, we skip step 1 and give the results based on step 2. To simplify the notation we do not use the tilde and the superscript e for the two-sided regulated bucket level and the effective demand, respectively.

Normal Demand When demand is normally distributed, with mean μ and standard deviation σ, we assume that the probability of demand lower than 0 is negligible, and we can combine the densities of the normal distribution and the approximation to the bucket level PDF (5.4)

and CDF (5.5) to approximate the CDF of the effective demand based on (5.5):

$$F_{U^e}\left(u^e\right) \approx \begin{cases} 0 & \text{for } u^e < 0 \\ F_U\left(u^e\right) & \text{for } 0 \leq u^e < r \\ \frac{1-e^{\theta\left(u^e-r\right)}}{1-e^{\theta d}} + \frac{e^{\theta\left(u^e-r\right)}-e^{\theta d}}{1-e^{\theta d}} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{u^e} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx & \text{for } [r \leq u^e \leq d+r] \\ 1 & \text{for } d+r < u^e. \end{cases}$$

We approximate the first moment (see computation in Appendix A.8):

$$\begin{aligned} E\left(u^e\right) &= \int_0^r 1 - F_U\left(u\right) du + \frac{\left(1-F_U\left(r\right)\right)}{1-e^{\theta d}} \left[\left(\frac{e^{\theta d}-1}{\theta}\right) - de^{\theta d}\right] \\ &\quad - \frac{e^{\theta d}\left(1/\theta - d - r\right)\left(F_U\left(d+r\right) - F_U\left(r\right)\right)}{1-e^{\theta d}} - e^{\theta d} E_u\left(x; x \in [0, d+r]\right) \\ &\quad + \frac{e^{-\theta r} e^{\left(\theta^2\sigma^2/2 + \mu\theta\right)}}{\theta} \left[\Phi\left(\frac{d+r-\mu-\theta\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\theta\sigma^2}{\sigma}\right)\right] \\ &= \int_0^r 1 - \Phi\left(\frac{u-\mu}{\sigma}\right) du + \frac{\left(1-\Phi\left(\frac{r-\mu}{\sigma}\right)\right)}{1-e^{\theta d}} \left[\left(\frac{e^{\theta d}-1}{\theta}\right) - de^{\theta d}\right] \\ &\quad - \frac{e^{\theta d}\left(1/\theta - d - r\right)\left(\Phi\left(\frac{d+r-\mu}{\sigma}\right) - \Phi\left(\frac{r-\mu}{\sigma}\right)\right)}{1-e^{\theta d}} - e^{\theta d} E_u\left(x; x \in [0, d+r]\right) \\ &\quad + \frac{e^{-\theta r} e^{\left(\theta^2\sigma^2/2 + \mu\theta\right)}}{\theta} \left[\Phi\left(\frac{d+r-\mu-\theta\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\theta\sigma^2}{\sigma}\right)\right] \end{aligned}$$

where $\theta = 2\left(r-\mu\right)/\sigma^2 > 0$.

Computation of additional moments can be done based on the PDF, which can be expressed using (5.6).

Exponential Demand Using (5.6) the density of the effective demand can be written as:

$$f_{U^e}\left(u\right) \approx \begin{cases} 0 & \text{for } u < 0 \\ \lambda e^{-\lambda u} & \text{for } 0 \leq u < r \\ \lambda e^{-\lambda u}\left[\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}}\right] + \left[e^{-\lambda u}\right]\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)} & \text{for } r \leq u \leq r+d \\ 0 & \text{for } d+r < u. \end{cases}$$

This leads to the following expressions for the first and second moments (see calculations in Appendix A.8):

$$\begin{aligned} E\left(u^{e}\right) &= \frac{1-e^{-\lambda r}\left(\lambda r+1\right)}{\lambda} \\ &-\frac{e^{-\lambda r}}{1-e^{\theta d}}\frac{1}{\lambda-\theta}\left[e^{-(\lambda-\theta)d}\left(\lambda-\theta\right)d-\left(1-e^{-(\lambda-\theta)d}\right)\left(\left(\lambda-\theta\right)r+1\right)\right] \\ &+\frac{e^{\theta d-\lambda r}}{\lambda\left(1-e^{\theta d}\right)}\left[\left(e^{-\lambda d}-1\right)\left(\lambda r+1\right)-e^{-\lambda d}\lambda d\right] \end{aligned} \tag{5.8}$$

$$\begin{aligned} E\left(\left(u^{e}\right)^{2}\right) &= \frac{2-e^{-\lambda r}\left[\left(\lambda r+1\right)^{2}+1\right]}{\lambda^{2}} \\ &+\frac{e^{\theta d-\lambda r}}{\lambda^{2}\left(1-e^{\theta d}\right)}\left[e^{-\lambda d}\left[\left(\lambda\left(r+d\right)+1\right)^{2}+1\right]-\left[\left(\lambda r+1\right)^{2}+1\right]\right] \\ &-\frac{1}{1-e^{\theta d}}\frac{e^{-\lambda r}}{\left(\lambda-\theta\right)^{2}}\left[e^{-(\lambda-\theta)d}\left[\left(\left(\lambda-\theta\right)\left(r+d\right)+1\right)^{2}+1\right]-\left[\left(\left(\lambda-\theta\right)r+1\right)^{2}+1\right]\right], \end{aligned} \tag{5.9}$$

where $\theta=2\lambda^{2}\left(r-1/\lambda\right)>0$.

5.2.3 Enhanced Approximation for the Effective Demand

Approximating the bucket level distribution with the distribution of a corresponding Brownian motion ignores the discrete nature of the bucket level (the random walk). One example of this is that, as a result of the continuous nature of the Brownian motion approximation, the two delta functions of the bucket level, at $\widetilde{l}=0$, and $\widetilde{l}=d$, are ignored. Therefore, in this subsection we present an original approach to approximate these delta functions and enhance the approximation presented in the previous subsection.

We use interchangeably the terms percentage-of-periods-with-losses, loss probabilities, and probability of loss. These probabilities can be related to either units-of-work (or, for simplicity, work) when it is noted as $P\left(\widetilde{L}=0\right)$, or to tokens, and when it is noted as $P\left(\widetilde{L}=d\right)$.

At the end of this subsection Algorithm 13, whose output is the probabilistic description of the effective demand, summarizes the results to that point. The inputs to the algorithm are buyer's demand CDF ($F_{U}\left(u\right)$, with a mean $E\left(u\right)$ and standard deviation σ) and the buyer's choice of the TB parameters (bucket's depth d and token rate r). We then demonstrate this

algorithm on the exponential and normal demand cases.

This section continues with describing primary and secondary methods to estimate the percentage-of-periods-with-losses of tokens or work. The secondary method considers the buyer's service level requirement (percentage-of-periods-with-losses is smaller than $1-\alpha$) as a surrogate for $P\left(\widetilde{L}=0\right)$. However, since buyers cannot solve their problems accurately, this gives only an upper bound on $P\left(\widetilde{L}=0\right)$. Furthermore, this method does not allow for a straightforward approximation of the delta function at $P\left(\widetilde{L}=d\right)$. Therefore, we suggest as the primary method to use the relations between the fill rate, which can be calculated from the Brownian motion approximation, to derive the percentage-of-periods-with-losses. Unfortunately, the primary method does not always work; in such cases we retreat to the secondary method.

Primary Method to Evaluate the Probabilities of Losses

Here we evaluate the percentage-of-periods-with-losses of tokens or of units-of-work. We first relate these service measures to fill rates, and next give these fill rates based on a Brownian motion approximation for the bucket level. Then, we explain how to approximate the expected-losses-per-periods, based on results drawn from the theory of large deviations and sequential analysis. Finally, we show how to use these relations in order to approximate the required tokens and units-of-work loss probabilities.

Relating the Fill Rate to the Percentage of Periods with Losses We define x_i as the number of units of work lost in period i and recall the definitions of the fill rate, noted FR, and the percentage-of-periods-with-losses (of work), noted α:

$$\begin{aligned} FR &\equiv 1-\frac{\lim_{N\to\infty}\sum_{i=1}^{N}x_i}{\lim_{N\to\infty}\sum_{i=1}^{N}u_i}=1-\frac{E\left(x\right)}{E\left(u\right)} \\ \alpha &\equiv 1-\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}I\left\{x_i>0\right\}=1-P\left(x>0\right), \end{aligned}$$

where $I\{x_i > 0\}$, is the indicator function for the event $x_i > 0$ (losses happened), i.e., it gets the value one if $x_i > 0$, and zero otherwise (if $x_i = 0$). The expectation of x_i can be written as:

$$\begin{aligned} E(x) &= E(x\,|x>0)\,P(x>0) + E(x\,|x=0)\,P(x=0) \\ &= E(x\,|x>0)\,[1-\alpha]\,. \end{aligned}$$

Thus, we can write:

$$FR = 1 - \frac{E(x\,|x>0)\,[1-\alpha]}{E(u)}$$

and after some algebra we can express the percentage-of-periods-with-losses as a function of the fill rate, the expected work lost in a period with losses, and the mean of buyer's demand. This relation is summarized in Lemma 12:

Lemma 12 *The relation between the percentage-of-periods-with-losses, noted α, to the work fill rate, noted FR, is given by:*

$$\alpha = 1 - \frac{(1-FR)\,E(u)}{E(x\,|x>0)}, \tag{5.10}$$

where x is the number of units of work lost in a period, and $E(u)$ is the mean of buyer's demand. In a similar manner, the relation between the percentage-of-periods-with-losses of tokens, noted α_T, to the tokens fill rate, noted FR_T, is given by:

$$\alpha_T = 1 - \frac{(1-FR_T)\,r}{E(x_T\,|x_T>0)}, \tag{5.11}$$

where $E(x_T\,|x_T>0)$ is the expected number of tokens lost in a period with losses of tokens, and r is the number of tokens that are added to the bucket in a period.

Approximating the Fill Rates In 3.23 we show that based on Proposition 5 on page 90 of (Harrison, 1985), the work fill rate realized from a two-sided regulated Brownian motion, with a drift $E(u) - r$ and a standard deviation σ, is:

$$FR = 1 - \frac{r - E(u)}{E(u)\,(e^{\theta d} - 1)}, \tag{5.12}$$

where $\theta = 2\,(r - E(u))\,/\sigma^2 > 0$. Similarly, the tokens' fill rate can be written as:

$$FR_T = 1 - \frac{r - E(u)}{r(1 - e^{-\theta d})}. \tag{5.13}$$

Remark : Approximating the expectation of the effective demand can be based on the demand's approximated fill rate:

$$\begin{aligned} E(u^e) &\approx r * FR_T = E(u) FR \\ &= E(u) - \frac{r - E(u)}{e^{\theta d} - 1} \end{aligned} \tag{5.14}$$

We will use (5.14) as an additional approximation for the expectation of the effective demand, as a part of the approximation using the Brownian motion. Later in this subsection we use the same idea to come out with an approximation for the expectation of the effective demand based on the fill rate resulting from the enhancement.

Approximating the Expected Losses in a Period with Losses We use (5.12) and (5.13) to approximate the fill rates realized by the corresponding random walk. Thus, in order to estimate the percentage-of-periods-with-losses we need to approximate the expressions for the expected losses of work/tokens in a period with losses of work/tokens.

$$\begin{aligned} E(x \,|x > 0) &= ? \\ E(x_T \,|x_T > 0) &= ? \end{aligned}$$

We use a similar procedure to approximate both expressions. There are two cases in which losses of work can happen. Clearly, in both the period's demand is higher than r. In the first case, the bucket level at the beginning of the period is positive. Then the bucket level goes from a positive value to zero, and the expected work lost is minus the expected level of the bucket, assuming there was no regulator at zero (i.e., the expected number of backlogs, if they were allowed). This value is also known as the undershoot of zero (by the non-regulated random walk). In the second case, the bucket level at the beginning of the period was zero. Then the expected work lost is given by:

$$E(x \,|x > 0) \approx E(u \,|u > r) - r. \tag{5.15}$$

Whereas, the expected work lost in the first case can be approximated by the expected residual life of the buyer's demand, given that it is higher than r, as explained in (Siegmund, 1985) and is given in (3.17), which we rewrite here:

$$E\left(x\,|x>0\right) \approx \frac{E\left((u-r\,|u>r\,)^2\right)}{2E(u-r\,|u>r\,)}.$$

Recall that this approximation works well when the bucket level's total down-movement from its maximum (after the last visit at zero) to zero goes to infinity. This happens when the above maximum is d, i.e., at the first arrival of the bucket level to zero (that is the first period in the first subcycle of a cycle, using our definitions from the discussion of the buyer's problem). However, this does not hold for other periods in which the bucket arrives at zero, such as periods within a subcycle, or at the beginning of any other (not the first) subcycle in a cycle.

Thus, using a combination of (5.15) and (3.17), $E\left(x\,|x>0\right)$ can be approximated.[1] It is important to note that in the exponential case both (5.15) and (3.17) are equal to $1/\lambda$. Thus, we know the exact value for $E\left(x\,|x>0\right)$. Unfortunately, in other demand cases both expressions lead to different values for $E\left(x\,|x>0\right)$; in general, we cannot even say if one is larger than the other.[2] As a guideline, when the service level is high, it is reasonable that most of the losses will happen from at least partially full bucket (i.e., in the beginning of a subcycle with losses) and, therefore, we tend to approximate the expected number of units lost in a period (given that loss has occurred) using (3.17).

We can write similar expressions for $E\left(x_T\,|x_T>0\right)$:

$$E\left(x_T\,|x_T>0\right) \approx \frac{E\left((r-u\,|u<r\,)^2\right)}{2E(r-u\,|u<r\,)} \tag{5.16}$$

$$E\left(x_T\,|x_T>0\right) \approx E\left(r-u\,|u<r\,\right) \tag{5.17}$$

Clearly, $E\left(x_T\,|x_T>0\right)<r$, and for high service level, which results in a strong positive drift, losses of tokens are more likely from a full bucket d. Thus, we tend to represent the expected

[1] Recall that bounds for the expected undershoot are cited from (Glasserman, 1997) and (Ross, 1974) in our work on the buyer's problem.

[2] More elegant results for the bounds are available regarding distributions that satisfy conditions of New Better (Worse) than Used, in both (Glasserman, 1997) and (Ross, 1974).

losses of tokens per period (given that loss of tokens has occurred) using (5.17).

An additional method to approximate the expected units lost in a period with losses of work, is to relate the work fill rate and the percentage-of-periods-with-losses based on (Glasserman, 1997), where it is shown that:

$$\begin{aligned} 1-FR &\approx C/\left[E\left(u\right)s^{*}\right]\left(1-e^{-s^{*}r}\right)e^{-s^{*}d} \\ 1-\alpha &\approx Ce^{-s^{*}d}, \end{aligned}$$

where s^{*} is the conjugate point of the demand distribution defined by the positive solution of $G_{U}\left(s\right)=e^{sr}$. In general, s^{*} can be numerically found; moreover, the solution we gave for the buyer's problem includes closed-form expressions for s^{*} for the cases of normal or exponential distributions of the buyer's demand. Thus, we can approximate

$$1-\alpha \approx \left(1-FR\right)\frac{E\left(u\right)s^{*}}{1-e^{-s^{*}r}}.$$

Using this relation, we see that the expected units lost in a period with losses can be estimated as:

$$E\left(x\left|x>0\right.\right)\approx\frac{1-e^{-s^{*}r}}{s^{*}}. \tag{5.18}$$

Putting It All Together Substituting the approximation for the fill rate (5.12), and for the expected units of work lost in a period with losses (3.17) (5.15) or (5.18) into (5.10), and using a similar procedure for tokens, using (5.13) to approximate the fill rate and (5.16) or (5.17) for the expected tokens lost in a period with losses, we can approximate the required delta

functions. Here we demonstrate this procedure for $P\left(\widetilde{L}=0\right)$ using (3.17):

$$\begin{aligned} P\left(\widetilde{L}=0\right) &= 1-\alpha \\ &\approx \frac{(1-FR)\,E\,(u)}{E\,(x\,|x>0\,)} \\ &= \frac{2E(u-r\,|u>r\,)\,(1-FR)\,E\,(u)}{E\,((u-r\,|u>r\,)^2)} \\ &= \frac{2E(u-r\,|u>r\,)\left(\frac{r-E(u)}{E(u)\left(e^{\theta d}-1\right)}\right)E\,(u)}{E\,((u-r\,|u>r\,)^2)} \\ &= \frac{2E(u-r\,|u>r\,)\,(r-E\,(u))}{E\,((u-r\,|u>r\,)^2)\,(e^{\theta d}-1)}. \end{aligned}$$

Secondary Method to Evaluate the Probabilities of Losses

You recall that an alternative method to estimate $P\left(\widetilde{L}=0\right)$ was mentioned at the beginning of this subsection. The buyer's service level requirement (percentage-of-periods-with-losses is smaller than $1-\alpha$) can be considered as a surrogate for $P\left(\widetilde{L}=0\right)$. Also, this service level is an upper bound on the service level provided by the admission control mechanisms, since the buyer's solution to its problem typically results in a higher service level than the required one. We will use this method when the fill rate predicted by the Brownian approximation is useless, which happens when $d\rightarrow 0$, or when the percentage-of-periods-with-losses predicted based on this fill rate is lower than the requested service level.

When we approximate $P\left(\widetilde{L}=0\right)=1-\alpha$, we still need to express the percentage-of-periods-with-losses of tokens. This can be done as follows. First, calculate the fill rate resulting from this service level, based on (5.10). This fill rate is given by:

$$FR=1-\frac{(1-\alpha)\,E\,(x\,|x>0\,)}{E\,(u)}. \tag{5.19}$$

We approximate $E\,(x\,|x>0\,)$ with the same methods mentioned in subsection 5.2.3. Note that in the exponential demand case there is no need to approximate $E\,(x\,|x>0\,)$, since the fill rate is equal to the percentage-of-periods-with-losses, as we show in Corollary 14.

Finally, having an estimate for the work fill rate, we calculate the token's fill rate from $E\,(u)\,FR=rFR_T$, which in turn allows us to estimate $P\,(L=d)$, again, this is similar to

(5.10).

The Enhanced Probabilistic Description of the Bucket Level and the Effective Demand

We define $c = 1 - P\left(\widetilde{L} = 0\right) - P\left(\widetilde{L} = d\right)$ and use the superscript E to note the approximation based on the enhancement. The following are the bucket's level PDF and CDF based on the enhancement:

$$f_{\widetilde{L}}^{E}\left(\widetilde{l}\right) \approx \begin{cases} P\left(\widetilde{L} = 0\right)\delta & \text{for } d \leq \widetilde{l} \\ \frac{c\left(1-e^{\theta\widetilde{l}}\right)}{1-e^{\theta d}} & \text{for } 0 < \widetilde{l} < d \\ P\left(\widetilde{L} = d\right)\delta & \text{for } d \leq \widetilde{l} \end{cases} . \tag{5.20}$$

$$F_{\widetilde{L}}^{E}\left(\widetilde{l}\right) \approx \begin{cases} 0 & \text{for } \widetilde{l} < 0 \\ P\left(\widetilde{L} = 0\right) + \frac{\left(1-e^{\theta\widetilde{l}}\right)\left[1-P\left(\widetilde{L}=0\right)-P\left(\widetilde{L}=d\right)\right]}{1-e^{\theta d}} & \text{for } 0 \leq \widetilde{l} < d \\ 1 & \text{for } d \leq \widetilde{l} \end{cases} . \tag{5.21}$$

Finally, we can write the enhanced approximation for the PDF of the effective demand:

$$f_{U^e}\left(u^e\right) = \begin{cases} f_U\left(u^e\right) & \text{for } 0 \leq u^e < r \\ f_U\left(r\right)\left[1 - P\left(\widetilde{L} = 0\right)\right] + \delta P\left(\widetilde{L} = 0\right)\left[1 - F_U\left(r\right)\right] & \text{for } u^e = r \\ f_U\left(u^e\right)\left[1 - F_{\widetilde{L}}^{E}\left(u^e - r\right)\right] + f_{\widetilde{L}}^{E}\left(u^e - r\right)\left[1 - F_U\left(u^e\right)\right] & \text{for } r < u^e < r + d \\ \delta P\left(\widetilde{L} = d\right)\left[1 - F_U\left(r + d\right)\right] & \text{for } u^e = d + r \\ 0 & \text{otherwise,} \end{cases}$$

which can be related to the PDF of the approximation as

$$f_{U^e}(u^e) = \begin{cases} f_U(u^e) & \text{for } 0 \le u^e < r \\ f_U(r)\left[1 - P\left(\widetilde{L}=0\right)\right] + \delta P\left(\widetilde{L}=0\right)[1 - F_U(r)] & \text{for } u^e = r \\ f_U(u^e)\left[1 - P\left(\widetilde{L}=0\right) - cF_{\widetilde{L}}(u^e - r)\right] + cf_{\widetilde{L}}(u^e - r)[1 - F_U(u^e)] & \\ & \text{for } r < u^e < r + d \\ \delta P\left(\widetilde{L}=d\right)[1 - F_U(r+d)] & \text{for } u^e = d + r \\ 0 & \text{otherwise,} \end{cases} \tag{5.22}$$

and the enhanced approximation for the CDF of the effective demand:

$$F^E_{U^e}(u^e) = \begin{cases} F_U(u^e) & \text{for } -\infty < u^e < r \\ F_U(r) + P\left(\widetilde{L}=0\right)[1 - F_U(r)] & \text{for } u^e = r \\ F_U(u^e) + F^E_{\widetilde{L}}(u^e - r)[1 - F_U(u^e)] & \text{for } r < u^e < r + d \\ 1 & \text{for } d + r \le u^e, \end{cases}$$

which can be related to the CDF of the approximation as

$$F^E_{U^e}(u^e) = \begin{cases} F_U(u^e) & \text{for } -\infty < u^e < r \\ F_U(r) + P\left(\widetilde{L}=0\right)[1 - F_U(r)] & \text{for } u^e = r \\ F_U(u^e) + \left(P\left(\widetilde{L}=0\right) + cF_{\widetilde{L}}(u^e - r)\right)[1 - F_U(u^e)] & \\ & \text{for } r < u^e < r + d \\ 1 & \text{for } d + r \le u^e. \end{cases} \tag{5.23}$$

Remark: Note that the enhancement presented in this subsection also changes the approximated fill rate of units of work, which in turn enhances the expectation of the effective demand from its approximation based on the Brownian motion given in (5.14). The enhancement also

considers the fill rate when the delta functions are active:

$$\begin{aligned} E^E(u^e) &\approx r * FR_T + rP\left(\widetilde{L}=0\right)[1-F_U(r)] + (r+d)\,P\left(\widetilde{L}=d\right)[1-F_U(r+d)] \\ &= E(u)\,FR + rP\left(\widetilde{L}=0\right)[1-F_U(r)] + (r+d)\,P\left(\widetilde{L}=d\right)[1-F_U(r+d)] \\ &= (1-P(L=0)[1-F_u(r)] - P(L=d)[1-F_u(r+d)])\left(E(u) - \frac{r-E(u)}{e^{\theta d}-1}\right) \\ &\quad + rP(L=0)[1-F_u(r)] + (r+d)\,P(L=d)[1-F_u(r+d)] \end{aligned} \tag{5.24}$$

The comparison of the expressions for the moments of the effective demand with and without the enhancement, where the enhancement is noted with the superscript E, shows that:

$$\begin{aligned} E^E\left((u^e)^i\right) - E\left((u^e)^i\right) &= \int_0^{r+d} u^e\left(f^E_{U^e}(u^e) - f_{U^e}(u^e)\right)du^e \\ &= rP\left(\widetilde{L}=0\right)[1-F_U(r)] \\ &\quad -P\left(\widetilde{L}=0\right)\int_r^{r+d} u f_U(u)\,du \\ &\quad +\frac{(1-c)}{1-e^{\theta d}}\int_r^{r+d} u\left[f_U(u)\left(1-e^{\theta(u-r)}\right) - \theta e^{\theta(u-r)}[1-F_U(u)]\right]du \\ &\quad +(r+d)\,P\left(\widetilde{L}=d\right)[1-F_U(d+r)] \end{aligned} \tag{5.25}$$

Enhanced Algorithm for Approximating the Effective Demand

We suggest using the following algorithm, which for ease of usage includes references to the equations that are explained earlier in this subsection.

Algorithm 13 *Step 0 (Input and Initialization): Inputs are: Buyer's demand CDF ($F_U(u)$, with a mean $E(u)$, and standard deviation σ), the buyer's choice of bucket's depth and token rate (d and r respectively), and buyer's required service level (α). Let $\theta = \frac{r-E(u)}{\sigma} > 0$.*
Step 1 (Estimate $E(x\,|x>0)$, and $E(x_T\,|x_T>0)$): Approximate the expected units of work lost in a period with losses of work $E(x\,|x>0)$, based on (3.17), (5.15), or (5.18). Approximate the expected tokens lost in a period with losses of tokens as $E(x_T\,|x_T>0)$ based on (5.16) or (5.17).
Step 2 (Approximate the fill rates): Approximate the work fill rate, FR, based on (5.12), and the token fill rate, FR_T, based on (5.13). If the fill rates are in the range $(0,1)$, go to step 4.

Step 3 (Alternative derivation of work fill rate): Approximate the percentage-of-periods-with-losses of jobs as $P\left(\widetilde{L}=0\right)=1-\alpha$. *Approximate work fill rate* FR *based on (5.19), and* $E\left(x\left|x>0\right.\right)$. *Approximate the token fill rate* FR_T *from* $rFR_T=E\left(u\right)FR$. *Go to step 5.*
Step 4 (Derivation of work loss probability): For units of work, translate the fill rate FR *to percentage-of-periods-with-losses of work* $P\left(\widetilde{L}=0\right)$ *(our approximation for* $1-\alpha$*), using* $E\left(x\left|x>0\right.\right)$ *and (5.10). If the approximation leads to* $P\left(\widetilde{L}=0\right)>1-\alpha$, *go to step 3.*
Step 5 (Derivation of tokens loss probability): For tokens, translate the fill rate FR_T *to percentage-of-periods-with-losses of tokens* $P\left(\widetilde{L}=d\right)$ *(our approximation for* $1-\alpha_T$*), using* $E\left(x_T\left|x_T>0\right.\right)$ *and (5.11).*
Step 6 (Approximate the bucket level probabilistic behavior): Approximate the enhancement for the bucket level PDF $f_{\widetilde{L}}^{E}\left(\widehat{l}\right)$ *and CDF* $F_{\widetilde{L}}^{E}\left(\widehat{l}\right)$ *based on (5.20) and (5.21), respectively.*
Step 7: (Approximate the effective demand probabilistic behavior): Approximate the effective demand PDF and CDF according to (5.22) and (5.23), respectively. Moments can be calculated using the PDF.

Examples

Normal Demand In the case of normal demand, we use the approximation given in (Siegmund, 1985) for the expected undershoot of a normal random variable with mean $r-\mu$ and standard deviation σ:

$$E\left(x\left|x>0\right.\right)\approx\frac{\sigma}{2\left(r-\mu\right)\sqrt{2\exp\left(-\frac{1.166(r-\mu)}{\sigma}\right)}},$$

and, therefore, we approximate the work loss probability, $P\left(\widetilde{L}=0\right)$:

$$P\left(\widetilde{L}=0\right)\approx\frac{r-\mu}{\left(e^{\theta d}-1\right)}\frac{2\left(r-\mu\right)\sqrt{2\exp\left(-\frac{1.166(r-\mu)}{\sigma}\right)}}{\sigma}$$

$\theta=2\left(r-\mu\right)/\sigma^2>0$.

Assuming that the overshoot starts at a full bucket, we estimate token loss probability based

on (5.17):

$$P\left(\widetilde{L}=d\right)\approx\frac{r-\mu}{\left(1-e^{-\theta d}\right)}/\left(r-\Phi\left(\frac{r-\mu}{\sigma}\right)\int_{0}^{r}\frac{u}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(u-\mu\right)^{2}}{2\sigma^{2}}\right)du\right),$$

We summarize it and give the CDF for the bucket level:

$$F_{\widetilde{L}}\left(\widetilde{l}\right)\approx\begin{cases}0 & \text{for } \widetilde{l}<0\\ P\left(\widetilde{L}=0\right)+\frac{\left(1-e^{\theta\widetilde{l}}\right)\left(1-P\left(\widetilde{L}=0\right)-P\left(\widetilde{L}=d\right)\right)}{1-e^{\theta d}} & \text{for } 0\leq\widetilde{l}<d\\ 1 & \text{for } d\leq\widetilde{l}\end{cases}$$

and the CDF for the effective demand

$$F_{U^{e}}\left(u\right)\approx\begin{cases}0 & \text{for } u<0\\ \Phi\left(\frac{u-\mu}{\sigma}\right) & \text{for } 0\leq u<r\\ \Phi\left(\frac{r-\mu}{\sigma}\right)+P\left(\widetilde{L}=0\right)\left(1-\Phi\left(\frac{r-\mu}{\sigma}\right)\right) & \text{for } u=r\\ \left(P\left(\widetilde{L}=0\right)\left(1-\Phi\left(\frac{u-\mu}{\sigma}\right)\right)+\frac{\left(1-e^{\theta(u-r)}\right)(1-c)}{\left(1-e^{\theta d}\right)}\right)\left(1-\Phi\left(\frac{u-\mu}{\sigma}\right)\right) & \\ +\Phi\left(\frac{u-\mu}{\sigma}\right) & \text{for } r<u<r+d\\ 1 & \text{for } d+r\leq u\end{cases},$$

where $\theta=2\left(r-\mu\right)/\sigma^{2}>0$.

Exponential Demand In the case of exponential demand, as remarked after (5.15), $E\left(x\left|x>0\right.\right)=1/\lambda$. Therefore, both service level measures are equivalent. We summarize this fact in Corollary 14:

Corollary 14 *In the exponential demand case the percentage-of-periods-with-losses of work is equal to the work fill rate.*

Proof. Due to the memoryless property of the exponential demand case, $E\left(x\left|x>0\right.\right)=E\left(u-r\left|u>r\right.\right)=1/\lambda$. Plugging this into Lemma 12, $\alpha=1-\frac{(1-FR)E(u)}{E(x|x>0)}=1-\frac{(1-FR)1/\lambda}{1/\lambda}=1-FR$. ■

Thus, using Corollary 14 we approximate the delta function for $P\left(\widetilde{L}=0\right)=1-FR$:

$$P\left(\widetilde{L}=0\right)=\frac{1/\lambda-r}{1/\lambda\left(1-e^{\theta d}\right)},$$

where $\theta=2\left(r-1/\lambda\right)/\left(1/\lambda\right)^{2}=2\lambda\left(\lambda r-1\right)>0$.

Similarly, for the tokens:

$$\begin{aligned} E\left(r-u\,|u<r\right) &= r-E\left(u\,|u<r\right) \\ &= r-\int_{0}^{r}\frac{u\lambda e^{-\lambda u}}{1-e^{-\lambda r}}du \\ &= r+\frac{e^{-\lambda r}\left(\lambda r+1\right)-1}{\left(1-e^{-\lambda r}\right)\lambda} \\ &= \frac{\lambda r+e^{-\lambda r}-1}{\left(1-e^{-\lambda r}\right)\lambda}. \end{aligned} \tag{5.26}$$

Alternatively, we recall (3.19) that show $\frac{E\left((r-u|u<r)^{2}\right)}{2E(r-u|u<r)}=\frac{e^{r\lambda}\left(2-2r\lambda+r^{2}\lambda^{2}\right)-2}{2\lambda\left[1+e^{r\lambda}(r\lambda-1)\right]}$:

Using (5.26) and $FR_{T}=1-\frac{r-E(u)}{r\left(1-e^{-\theta d}\right)}$, we approximate:

$$\begin{aligned} P\left(\widetilde{L}=d\right) &= \frac{\left(1-FR_{T}\right)r}{E\left(x_{T}\,|x_{T}>0\right)} \\ &\approx \frac{r-1/\lambda}{e^{-\theta d}-1}\left(\frac{\left(1-e^{-\lambda r}\right)\lambda}{\lambda r+e^{-\lambda r}-1}\right). \end{aligned}$$

In sum, the bucket level CDF can be approximated:

$$F_{\widetilde{L}}\left(\widetilde{l}\right)\approx\begin{cases} 0 & \text{for } \widetilde{l}<0 \\ P\left(\widetilde{L}=0\right)+\frac{\left(1-e^{\theta\widetilde{l}}\right)\left(1-P\left(\widetilde{L}=0\right)-P\left(\widetilde{L}=d\right)\right)}{1-e^{\theta d}} & \text{for } 0\leq\widetilde{l}<d \\ 1 & \text{for } d\leq\widetilde{l} \end{cases}.$$

Then the effective demand CDF is approximated as:

$$F_{U^e}(u) \approx \begin{cases} 0 & \text{for } u < 0 \\ 1 - e^{-\lambda u} & \text{for } 0 \leq u < r \\ \left(\frac{1 - r\lambda + \left(1 - e^{\theta(u-r)}\right)\left(1 - \frac{1/\lambda - r}{1/\lambda\left(1-e^{\theta d}\right)} - \frac{r - 1/\lambda}{e^{-\theta d}-1}\frac{\left(1-e^{-\lambda r}\right)\lambda}{\lambda r + e^{-\lambda r} - 1}\right)}{\left(1 - e^{\theta d}\right)} \right) e^{-\lambda u} & \\ + \frac{(1/\lambda - r)e^{-\lambda r}}{1/\lambda\left(1-e^{\theta d}\right)} + 1 - e^{-\lambda r} & \text{for } r \leq u \leq r + d \\ 1 & \text{for } d + r < u, \end{cases}$$

where $\theta = 2\left(r - 1/\lambda\right) / \left(1/\lambda\right)^2 = 2\lambda\left(\lambda r - 1\right) > 0$.

Finally, the first and second moments can be written using (5.8), (5.9), and (5.25) (see computation at Appendix A.8):

$$\begin{aligned} E^E(u^e) &= \frac{1 - e^{-\lambda r}(\lambda r + 1)}{\lambda} \\ &\quad - \frac{ce^{-\lambda r}}{1 - e^{\theta d}} \frac{1}{\lambda - \theta} \left[e^{-(\lambda-\theta)d}(\lambda - \theta)\, d - \left(1 - e^{-(\lambda-\theta)d}\right)\left((\lambda - \theta)\, r + 1\right) \right] \\ &\quad + \frac{ce^{\theta d - \lambda r}}{\lambda\left(1 - e^{\theta d}\right)} \left[\left(e^{-\lambda d} - 1\right)(\lambda r + 1) - e^{-\lambda d}\lambda d \right] \\ &\quad + P\left(\widetilde{L} = 0\right) e^{-\lambda r} r + P\left(\widetilde{L} = d\right) e^{\lambda(d+r)}(r + d) \end{aligned}$$

$$\begin{aligned} E\left((u^e)^2\right) &= \frac{2 - e^{-\lambda r}\left[(\lambda r + 1)^2 + 1\right]}{\lambda^2} \\ &\quad + \frac{ce^{\theta d - \lambda r}}{\lambda^2\left(1 - e^{\theta d}\right)} \left[e^{-\lambda d}\left[(\lambda(r + d) + 1)^2 + 1\right] - \left[(\lambda r + 1)^2 + 1\right] \right] \\ &\quad - \frac{1}{1 - e^{\theta d}} \frac{ce^{-\lambda r}}{(\lambda - \theta)^2} \left[e^{-(\lambda-\theta)d}\left[((\lambda - \theta)(r + d) + 1)^2 + 1\right] - \left[((\lambda - \theta)\, r + 1)^2 + 1\right] \right] \\ &\quad + P\left(\widetilde{L} = 0\right) e^{-\lambda r} r^2 + P\left(\widetilde{L} = d\right) e^{-\lambda(r+d)}(r + d)^2 . \end{aligned}$$

Ranges for Bucket level Ranges for effective demand	$L \leq -r$	$-r < L < 0$	$L = 0$	$0 < L$
$0 \leq u^e < r$	NA	$u - L = u^e$	$u = u^e$	$u = u^e$
$r = u^e$	$w.p\ 1$	$u - L \geq r$	$u \geq r$	$u = r$
$r < u^e < r + d$	NA	NA	NA	$\{u = u^e; L > u^e - r\}$ or $\{u > u^e; L = u^e - r\}$
$r + d = u^e$	NA	NA	NA	$\{u > r + d; L = d\}$

Table 5.1: This is the key table for the realization of the effective demand in the token bucket with rate control case. The left column includes the possible ranges for the effective demand, and the top row includes the possible ranges for the bucket level. The table's cells list the conditions that the buyer's demand needs to satisfy in order for the requested effective demand (in the left column) to be realized, given the bucket level (in the top row).

5.3 Token Bucket with Rate Control

5.3.1 The Probabilistic Description of the Effective Demand

In the case of a TBwRC, the bucket level behaves as a one-sided regulated random walk in the range minus infinity to d. Due to the negative values possible for the bucket level, the characterization of the effective demand's PDF is more complex than in the TB case.

First, we note that effective demand in the range $u^e = (r, r + d]$ cannot occur when the bucket level is negative, and therefore, in this range the PDF of the effective demand in the TBwRC is the same as the one of the effective demand in the case of a TB. However, in the range $u^e = [0, r]$ there is a possibility that the bucket level was below zero, and thus more attention is needed. Therefore, Table 5.1 presents the conditions under which different ranges of effective demand are realized. The left column includes the possible ranges for the effective demand, and the top row includes the possible ranges for the bucket level. The table's cells list the conditions that the buyer's demand needs to satisfy in order for the requested effective demand (in the left column) to be realized, given the bucket level (in the top row).

The interpretation of Table 5.1 is as follows. If the bucket level is lower than $-r$, the backlog is so large that all of the next period's tokens supply will be used. Therefore, in this case, the effective demand will always be r. In cases where the bucket level is in the range $[-r < L < 0]$, and the sum of buyer's demand and the backlog $(-L)$ is higher than r, the effective demand will be r (since the token supply will constrain the effective demand). Note that, in such

cases, the bucket level at the end of this period will stay non-positive. However, even when the bucket level is in this range ($[-r < L < 0]$), lower effective demand can be realized. This will happen when the sum of buyer's demand and the backlog ($-L$) is lower than r; in such cases the effective demand will equal this sum. Note that in these cases the bucket level at the end of this period will be positive. As mentioned earlier, cases of non-negative bucket level are identical to the ones discussed for the TB admission control. Thus we do not review them here; however, we will recall them in the next paragraph, when explaining the effective demand's PDF.

Based on Table 5.1, the PDF of the effective demand can be written as:

$$f_{U^e}(u^e) = \begin{cases} f_U(u^e) P(L \geq 0) + P(u - L = u^e; -r < L < 0; 0 \leq u < u^e) & \text{for } u^e \in [0, r) \\ P(L \leq -r)\delta + P(u - L \geq r; -r < L < 0)\delta + & \\ f_U(r) P(L \geq 0) + [1 - F_U(r)] P(L = 0)\delta & \text{for } u^e = r \\ f_U(u^e) P(L > u^e - r) + P(u > u^e) f_L(u^e - r) & \text{for } u^e = (r, r + d) \\ P(u \geq r + d) P(L = d)\delta & \text{for } u^e = r + d \\ 0 & \text{otherwise} \end{cases} \tag{5.27}$$

Like (5.2), equation (5.29) is central to our development of the probabilistic description of the effective demand. The interpretation of (5.29) is easier to follow using Table 5.1.

For effective demand lower than r, two disjoint cases can happen: *Either* the effective demand is equal to the buyer's demand when the bucket level is nonnegative. This gives us the first term, corresponding to columns three and four in the table. *Or* the effective demand is equal to $u - L$ (the sum of buyer's demand and backlogs) when the bucket level satisfies $-r < L < 0$. In this case the buyer's usage satisfies $0 \leq u < u^e$; thus, to later simplify the integration of the PDF, we include this range restriction in the second element that corresponds to column two in the table.

For effective demand equal to r: The first element corresponds to the first column of the table; whenever the bucket level is at or below $-r$ the effective demand will be r. The second element corresponds to the second column of the table, when the bucket level satisfies $-r < L < 0$. Then the effective demand would be r if the backlog plus buyer's demand in this period is higher than (or equal to) r, i.e., $u - L \geq r$. The third and fourth elements

correspond to columns three and four. The third element holds when the buyer's demand is r and the bucket level is nonnegative. The fourth element happens whenever the buyer's demand is higher than r but the bucket is empty. Finally, we remind the reader of our use of the delta functions, whenever the above expressions led to a probability rather than density.

For effective demand in the range $(r, d+r)$: We recall that if there is any backlog, this event cannot happen (since at most the effective demand will be the tokens supply r). Therefore, the density in this case is the same as in the TB case: it is the union of two disjoint events. The first one is when the buyer's demand equals u^e and there are more than $u^e - r$ tokens in the bucket; then, all of the demand is admitted and the effective demand is equal to the requested one. The second case is when the demand is higher than or equal to u^e, but the bucket level is $(u^e - r)$; then, the effective demand is the bucket level plus the token rate (which is lower than u). Both events appear in column four of the table.

For effective demand equal to $d+r$: This will happen when buyer's demand is higher than (or equal to) $r + d$ and the bucket is full (corresponding to column four in the table).

With additional algebra we translate (5.27) to:

$$f_{U^e}(u^e) = \begin{cases} f_U(u^e)\left[1 - F_L(0^-)\right] + \int_0^{u^{e-}} f_U(y) f_L(y - u^e)\, dy & \text{for } u^e \in [0, r) \\ \delta\left[F_L(-r) + P(-r < L < 0)\left[1 - F_U(r)\right] + \int_0^r \int_{-r^+}^{x^- - r} f_U(x) f_L(y) dy dx\right] & \\ + f_U(r) P(L \geq 0) + \left[1 - F_U(r)\right] P(L = 0)\, \delta & \text{for } u^e = r \\ f_U(u^e)\left[1 - F_L(u^e - r)\right] + \left[1 - F_U(u^e)\right] f_L(u^e - r) & \text{for } u^e = (r, r+d) \\ \left[1 - F_U\left((r+d)^-\right)\right] P(L = d)\, \delta & \text{for } u^e = r + d \\ 0 & \text{otherwise} \end{cases} \tag{5.28}$$

The straight forward steps are to introduce the CDF of the bucket level and buyer's demand where we can. We focus on explaining the integrals, and the second element at $u^e = r$ in (5.28). Both result from the expressions where the bucket level is in the range $(-r < L < 0)$, corresponding to the second column of Table 5.1.

For an effective demand lower than r, we express $P(u - L = u^e; -r < L < 0; 0 \leq u < u^e)$ by conditioning on buyer's demand $u = y$, then the bucket level should be $y - u^e$, in order for the effective demand to equal u^e. We integrate over the values for buyer's demand that can lead to

an effective demand u^e, i.e., demand values in the range zero to u^{e-}, since in (5.27) the bucket level is strictly negative (the upper limit of the integral includes a minus sign). Clearly, this integral could be expressed also as $\int_{x=-r^+}^{-0^-} f_U\left(u^e+x\right) f_L\left(x\right) dx$, by conditioning on the bucket level. In fact, by changing the parameters of integration $x = y - u^e$ it can be seen that both integrals are the same.

To get the second and third elements, when the effective demand is r, we divide the element corresponding to a bucket level in the range $(-r < L < 0)$ of (5.27) into two. When the buyer's demand is higher than r it is clear that $u^e - L \geq r$, which we summarize as $P\left(-r < L < 0\right)\left[1 - F_U\left(r\right)\right]$. The double integral in (5.28) corresponds to buyer's usage which is lower or equal to r. Then, a bucket level between $-r^+$ to $u^{e-} - r < 0$ will satisfy the requirement $u^e - L \geq r$, which is summarized as $\int_0^r \int_{-r^+}^{x^- - r} f_U(x) f_L(y) dy dx$.

We can further simplify (5.28) to:

$$f_{U^e}\left(u^e\right) = \begin{cases} f_U\left(u^e\right)\left[1 - F_L(0^-)\right] + \int_{y=0}^{u^{e-}} f_U\left(y\right) f_L\left(y - u^e\right) dy & \text{for } u^e \in [0, r) \\ \left[F_L(-r) F_U\left(r\right) + \int_0^r \int_{-r^+}^{x^- - r} f_U(x) f_L(y) dy dx + \left[1 - F_U\left(r\right)\right] F_L\left(0\right)\right] \delta & \text{for } u^e = r \\ f_U\left(u^e\right)\left[1 - F_L\left(u^e - r\right)\right] + \left[1 - F_U\left(u^e\right)\right] f_L\left(u^e - r\right) & \text{for } u^e = (r, r+d) \\ \left[1 - F_U\left(r + d\right)\right] P\left(L = d\right) \delta & \text{for } u^e = r + d \\ 0 & \text{otherwise} \end{cases} \tag{5.29}$$

For translating (5.28) to (5.29) we use our assumption that buyer's demand is continuous, and therefore, we can replace $(r+d)^-$, with $r + d$ for the buyer's demand distribution, without changing the PDF (almost surely). We also collected the arguments multiplied by $1 - F_U\left(r\right)$.

Taking the integral of (5.29) we can express the CDF of the effective demand as (see com-

putation in Appendix A.9):

$$F_{U^e}(u^e) = \begin{cases} 0 & \text{for } u^e < 0 \\ \int_0^{u^e} \int_0^{x^-} f_U(y) f_L(y-x)\, dydx + F_U(u^e)\left[1 - F_L(0^-)\right] & \text{for } u^e \in [0, r) \\ F_U(r)\left[1 - F_L(0)\right] + F_L(0) & \text{for } u^e = r \\ F_U(u^e) + F_L(u^e - r)\left[1 - F_U(u^e)\right] & \text{for } u^e = (r, r+d) \\ 1 & \text{for } u^e \geq r + d \end{cases} \tag{5.30}$$

It seems that if we could know $F_L(l)$ we could express $F_{U^e}(u^e)$ analytically. However, the distribution of the bucket level is not known. Therefore, in what follows, we approximate this distribution.

Remark : We can express the different moments of the effective demand using its PDF. However, since in the token bucket with rate control all of the buyer's demand has been processed (finally), the first moment of the effective demand is the buyer's expected demand $E(u^e) = E(u)$.

Similarly to our analysis of the TB case, we present here two approximations for the CDF of the effective demand. The first one is based upon a diffusion approximation for the bucket density and the second one is an enhancement of this approximation that compensates for the discrete nature of the TB model. Again, to simplify the implementation of these methods, we summarize each method with an algorithm that implements its results. The inputs to the algorithms are buyer's demand CDF ($F_U(u)$, with a mean $E(u)$ and standard deviation σ), the buyer's choice of the TB parameters (bucket's depth d and token rate r), and the requested service level α (which is not always required). We demonstrate the use of these algorithms in the normal and exponential demand cases.

5.3.2 Approximating the Effective Demand Using Brownian Motion Techniques

In this section we approximate the bucket level's PDF, described by a one-sided regulated random walk, using the PDF of a Brownian motion. Based on this approximation, we later approximate the PDF and CDF of the effective demand. Algorithm 15, whose inputs are buyer's demand CDF ($F_U(u)$, with a mean $E(u)$ and standard deviation σ) and the buyer's

choice of the TB parameters (bucket's depth d and token rate r), summarizes the results of this section.

Approximating the Probabilistic Description of the Bucket Level

As remarked earlier, the density of the bucket level process is not known. However, the density of a Brownian motion that is regulated at zero and has a mean $r - E(u)$ and a standard deviation σ, is known to be exponential with parameter $\theta = 2(r - E(u))/\sigma^2 > 0$ (Harrison, 1985). Thus, in our case the shortfall process $Y \equiv d - L$ is exponentially distributed, and the density of the bucket level process L can be approximated as:

$$f_L(l) \approx \begin{cases} \theta e^{-\theta(d-l)} & l \in [0, \infty) \\ 0 & \text{otherwise.} \end{cases} \tag{5.31}$$

Integrating the above on the range $(-\infty, l]$ we approximate the CDF:

$$F_L(l) \approx \begin{cases} e^{-\theta(d-l)} & l \leq d \\ 1 & d < l. \end{cases} \tag{5.32}$$

Approximating the Probabilistic Description of the Effective Demand

In the following, we suppress the superscript e from the value of the effective demand within formulas whenever no confusion arises. Then, using (5.31) and (5.32), we approximate the effective demand's PDF, using (5.29) (detailed calculations are given at Appendix A.9):

$$f_{U^e}(u) \approx \begin{cases} \int_0^{u^e} f_U(x)\,\theta e^{-\theta(d-(x-u^e))}dx + f_U(u^e)\left[1 - e^{-\theta d}\right] & \text{for } u^e \in [0, r) \\ e^{-\theta(d+r)} \int_0^r f_U(x) e^{\theta x} dx\delta + [1 - F_U(r)]e^{-\theta d}\delta & \text{for } u^e = r \\ f_U(u^e) - e^{-\theta(d-(u^e-r))}\left[f_U(u^e) + F_U(u^e)\,\theta - \theta\right] & \text{for } u^e = [r, r+d] \\ 0 & \text{otherwise} \end{cases} \tag{5.33}$$

Similarly, based on (5.30), the CDF in the range $u^e \in [0, r)$ is given by:

$$\begin{aligned} F_{U^e}(u^e) &\approx \theta e^{-\theta d} \int_0^{u^e} \int_0^x f_U(y)\, e^{\theta(y-x)} dy dx + F_U(u^e)\left[1 - e^{-\theta d}\right] \\ &= e^{-\theta d} F_U(u^e) - e^{-\theta(d+u^e)} \int_0^{u^e} f_U(y)\, e^{\theta y} dy + F_U(u^e)\left[1 - e^{-\theta d}\right] \\ &= F_U(u^e) - e^{-\theta(d+u^e)} \int_0^{u^e} f_U(y)\, e^{\theta y} dy, \end{aligned}$$

where the first equality follows from change of order of integration and the second one from cancellation. Thus, the CDF can be written as:

$$F_{U^e}(u^e) \approx \begin{cases} 0 & \text{for } u^e < 0 \\ F_U(u^e) - e^{-\theta(d+u^e)} \int_0^{u^e} f_U(y)\, e^{\theta y} dy & \text{for } u^e \in [0, r) \\ F_U(r)\left[1 - e^{-\theta d}\right] + e^{-\theta d} & \text{for } u^e = r \\ F_U(u^e)\left[1 - e^{-\theta(d-(u^e-r))}\right] + e^{-\theta(d-(u^e-r))} & \text{for } u^e = (r, r+d) \\ 1 & \text{for } u^e \geq r + d \end{cases} \tag{5.34}$$

Getting the MGF from (5.33) is trivial and adds no insight; thus it is omitted.

Approximated Algorithm for the Effective Demand

The following algorithm summarizes the results in this subsection:

Algorithm 15 *Step 0 (Inputs and initialization): Inputs are: buyer's demand CDF ($F_U(u)$, with a mean $E(u)$, and standard deviation σ) and the buyer's choice of bucket's depth and token rate (d and r respectively). Let $\theta = \frac{r - E(u)}{\sigma}$.*
Step 1 (Approximate the bucket level probabilistic behavior): Approximate the bucket level PDF and CDF according to (5.31), and (5.32), respectively.
Step2 (Approximate the effective demand probabilistic behavior): Approximate the effective demand PDF and CDF according to (5.33) and (5.34), respectively. Moments can be calculated using the PDF.

5.3.3 Enhanced Approximation for the Effective Demand

In the case of the bucket level with rate control, the bucket level behaves as a one-sided regulated random walk; therefore, the bucket level distribution has a delta function at $l = d$. Here we approximate this delta function, using a similar procedure to the one described in subsection 5.2.3.

Evaluating Tokens' Loss Probability

Evaluating tokens' fill rate in the TBwRC case is easy since no units of work are being lost. Therefore, the tokens' fill rate is given by $FR_T = E(u)/r$ (this would be ρ of the corresponding queue). We recall here equation (5.11):

$$\alpha_T = 1 - \frac{(1 - FR_T)\, r}{E(x_T \,|x_T > 0)}.$$

Substituting the tokens' fill rate into (5.11) gives:

$$\alpha_T = 1 - \frac{r - E(u)}{E(x_T \,|x_T > 0)}$$

The expressions given in (5.16) can be used to approximate $E(x_T \,|x_T > 0)$. These approximations allow us to approximate $P(L = d) = 1 - \alpha_T$:

$$\begin{aligned} P(L = d) &\approx \frac{2E(r - u \,|u < r\,)\,(r - E(u))}{E((r - u \,|u < r\,)^2)} \\ P(L = d) &\approx \frac{(r - E(u))}{E(r - u \,|u < r\,)} \end{aligned}$$

For similar reasons to the ones mentioned in subsection 5.2.3, we tend to choose the second expression to represent the tokens' loss probability.

The Enhanced Probabilistic Description of the Bucket Level and the Effective Demand

We further define $c = 1 - P(L = d)$ and use the superscript E to note the approximations based on the enhancement, so as to get the bucket's level PDF and CDF based on the enhancement:

$$f_L(l) \approx \begin{cases} ce^{-\theta(d-l)} & \text{for } l < d \\ 0 & \text{otherwise} \end{cases} \tag{5.35}$$

$$F_L^E(l) \approx \begin{cases} ce^{-\theta(d-l)} & \text{for } l < d \\ 1 & \text{for } d \leq \widetilde{l} \end{cases} . \tag{5.36}$$

Finally, we can write the resulting approximation for the PDF of the effective demand based on (5.29), (5.35), and (5.36), as is shown in Appendix A.9:

$$f_{U^e}^E(u) \approx \begin{cases} c\int_0^{u^e} f_U(x)\,\theta e^{-\theta(d-(x-u^e))}dx + f_U(u^e)\left[1 - ce^{-\theta d}\right] & \text{for } u^e \in [0, r) \\ ce^{-\theta(d+r)}\int_0^r f_U(x)e^{\theta x}dx\delta + c[1 - F_U(r)]e^{-\theta d}\delta & \text{for } u^e = r \\ f_U(u^e) - ce^{-\theta(d-(u^e-r))}\left[f_U(u^e) + F_U(u^e)\,\theta - \theta\right] & \text{for } u^e = (r, r+d) \\ \left[1 - F_U(r+d)\right]P(L = d)\,\delta & \text{for } u^e = r + d \\ 0 & \text{otherwise} \end{cases} \tag{5.37}$$

Integrating (5.37), we get the approximated CDF (see Appendix A.9):

$$F_{U^e}^E(u^e) \approx \begin{cases} 0 & \text{for } u^e < 0 \\ F_U(u^e) - ce^{-\theta(d+u^e)}\int_0^{u^e} f_U(y)\,e^{\theta y}dy & \text{for } u^e \in [0, r) \\ F_U(r)\left[1 - ce^{-\theta d}\right] + ce^{-\theta d} & \text{for } u^e = r \\ F_U(u^e)\left[1 - ce^{-\theta(d-(u^e-r))}\right] + ce^{-\theta(d-(u^e-r))} & \text{for } u^e = (r, r+d) \\ 1 & \text{for } u^e \geq r + d \end{cases} \tag{5.38}$$

Enhanced Algorithm for Approximating the Effective Demand

We suggest using the following algorithm, which for ease of usage includes references to the equations that are explained earlier in this subsection.

Algorithm 16 *Step 0 (Input and Initialization): Inputs are: buyer's demand CDF ($F_U(u)$,*

with a mean $E(u)$, and standard deviation σ), the buyer's choice of bucket's depth and token rate (d and r respectively) and buyer's required service level α. Let $\theta = \frac{r-E(u)}{\sigma} > 0$.

Step 1 (Estimate $E(x_T | x_T > 0)$): Approximate the expected tokens lost in a period with losses of tokens $E(x_T | x_T > 0)$ based on (5.16) or (5.17).

Step 2 (The tokens fill rate): Let $FR_T = \frac{E(u)}{r}$.

Step 3 (Derivation of token loss probability): Translate the fill rate FR_T to percentage-of-periods-with-losses of tokens $P\left(\widetilde{L} = d\right)$ (our approximation for $1 - \alpha_T$), using $E(x_T | x_T > 0)$ and (5.11).

Step 4 (Approximate the bucket level probabilistic behavior): Approximate the enhancement to the bucket level PDF $f^E_{\widetilde{L}}\left(\widetilde{l}\right)$ and CDF $F^E_{\widetilde{L}}\left(\widetilde{l}\right)$ based on (5.35) and (5.36) respectively.

Step 5: (Approximate the effective demand probabilistic behavior): Approximate the effective demand PDF and CDF according to (5.29) and (5.30), respectively. Recall that the first moment is $E(u)$; higher moments can be calculated using the PDF.

Examples

We give here our approximations for $P(L = d)$, for the cases of normal or exponential demands. Plugging these results into (5.37) and (5.38) is trivial.

Normal Demand In the case of normal demand, we again use the approximation given in (Siegmund, 1985) for the expected undershoot of a normal random variable with mean $r - \mu$ and standard deviation σ:

$$E(x | x > 0) \approx \frac{\sigma}{2(r-\mu)\sqrt{2\exp\left(-\frac{1.166(r-\mu)}{\sigma}\right)}}$$

therefore, we approximate the delta function for $P(L = d)$:

$$P(L = d) \approx \frac{2\sqrt{2\exp\left(-\frac{1.166(r-\mu)}{\sigma}\right)}}{\sigma}$$

or we can use the assumptions that most losses of tokens start at a full bucket:

$$P\left(L=d\right) \approx \left(r-\mu\right) / \int_0^r \frac{u}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(u-\mu\right)^2}{2\sigma^2}\right) du.$$

Exponential Demand In the case of exponential demand, we use (5.16) to approximate:

$$\begin{aligned} P\left(L=d\right) &= \frac{\left(r-1/\lambda\right)}{E\left(x_T \left| x_T>0\right.\right)} \\ &\approx \left(r-1/\lambda\right)\left(\frac{\left(1-e^{-\lambda r}\right)\lambda}{\lambda r+e^{-\lambda r}-1}\right). \end{aligned}$$

5.4 Numerical Results

In order to check the performance of the methods developed in this chapter, we ran an extensive simulation with a wide range of parameters and compared the CDF's and moments solutions based on the Brownian approximation and its enhancement to the ones from the simulation. Due to the large memory requirements of plotting the simulated and calculated CDF's, the simulation length we chose is of only 10,000 periods. We made this choice after running a few cases for 500,000 periods and finding that the results were comparable to the ones of the shorter simulation. In addition, since the results we got are consistent among all of the different cases (mentioned below), we consider this length as sufficient. Finally, we provide the 99% confidence intervals for the mean estimations of the simulation. In the normal case the confidence intervals were less than 0.8% off the estimated mean, and in the exponential case the confidence intervals were less than 2.8% off the estimated mean.

We used (5.18) to approximate the percentage-of-periods-with-losses of work, $P\left(L=0\right)$, and (5.17) to approximate the percentage-of-periods-with-losses of tokens, $P\left(L=d\right)$.

5.4.1 Normal Demand

We used a normal demand with mean 10; standard deviation of one, two, or three; service level requirements of 80%, 90%, 95%, and 99%; and D/R cost ratios of 0.9, 0.5, 0.2, 0.1 (recall that if D>R, there is no point in purchasing depth).

Figure 5-1 gives a comparison between the CDF of buyer's demand, when the standard

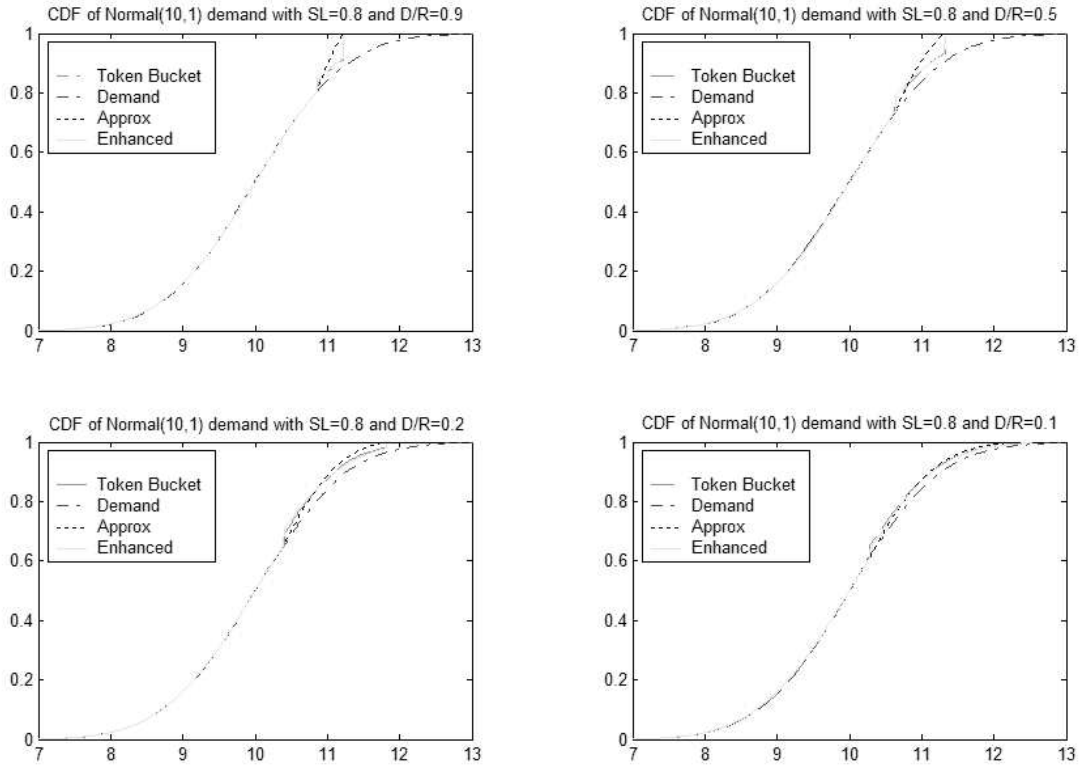


Figure 5-1: A comparison between the CDF of buyer's demand, drawn from a normal(10,1), to the output of the token bucket, the prediction based upon the Brownian motion approximation, and the prediction based on the enhancement. The requested service level is of percentage of periods with loss smaller than 80%. The cost ratios D/R of depth-to-rate are 0.1, 0.2, 0.5, and 0.9.

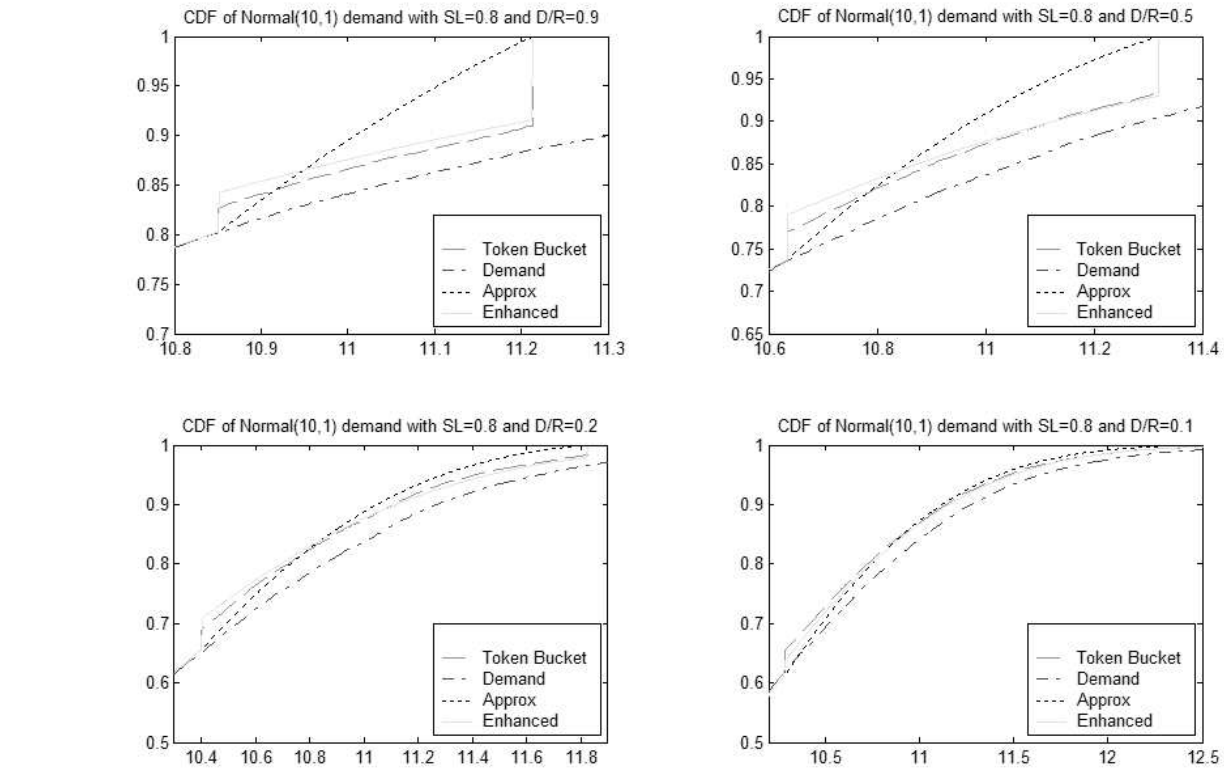


Figure 5-2: Zooming on the range $(r, d + r)$ of Figure 5-1: the comparison between the CDF of buyer's demand, drawn from a normal(10,1), to the output of the token bucket, the prediction based upon the Brownian motion approximation, and the prediction based on the enhancement. The requested service level is of percentage of periods with loss smaller than 80%. The cost ratios D/R of depth-to-rate are 0.1, 0.2, 0.5, and 0.9.

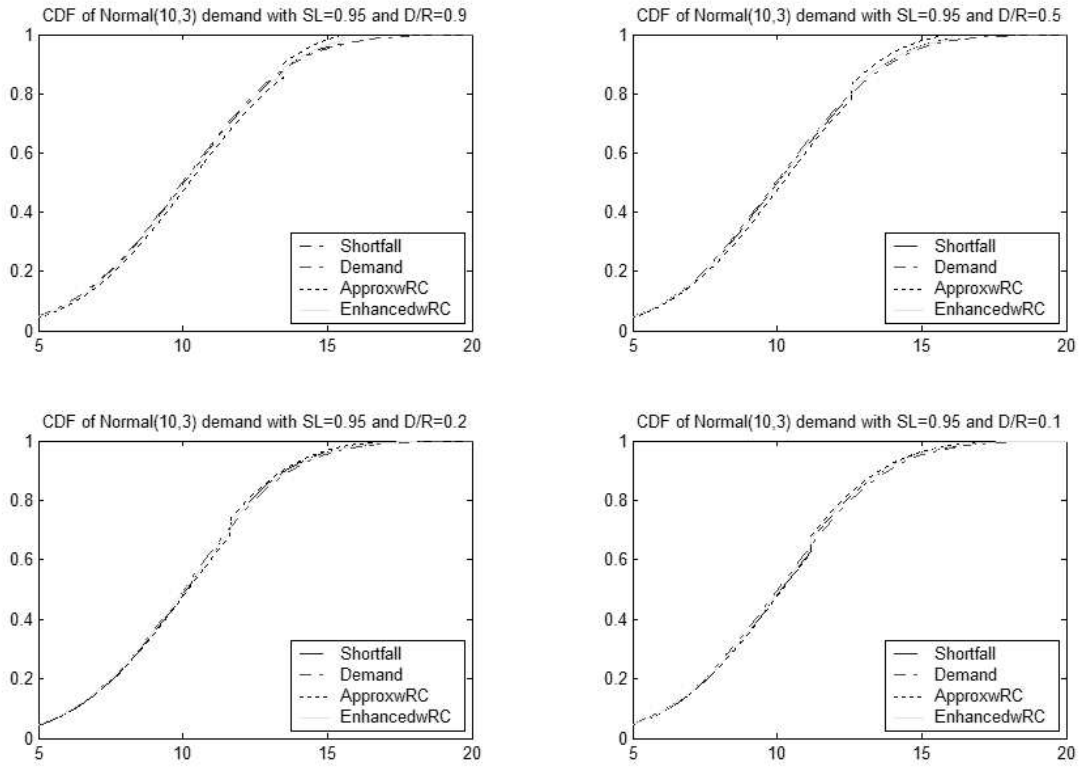


Figure 5-3: A comparison between the CDF of buyer's demand, drawn from a normal(10,3), to the output of the token bucket with rate control, the prediction based upon the Brownian motion approximation, and the prediction based on the enhancement. The requested service level is of percentage of periods with loss smaller than 95%. The cost ratios D/R of depth-to-rate are 0.1, 0.2, 0.5, and 0.9.

deviation is one, to the output of the TB, the prediction based upon the Brownian motion approximation, and the prediction based on the enhancement. The requested service level is of percentage-of-periods-with-losses smaller than 80%. The cost ratios D/R of the depth to rate are 0.1, 0.2, 0.5, and 0.9. As the theoretical calculations show, the relevant range for the TB case is between r and d; therefore, we present, in Figure 5-2, the same graphs zoomed around this range. It is seen that the CDF based on the Brownian motion approximation is closer to the one resulting from the simulation when depth is more expensive, i.e., ignoring $P(L = d)$ is less critical. Indeed, in all cases the enhancement does better than the approximation. Similar results hold for higher service level requirements as well.

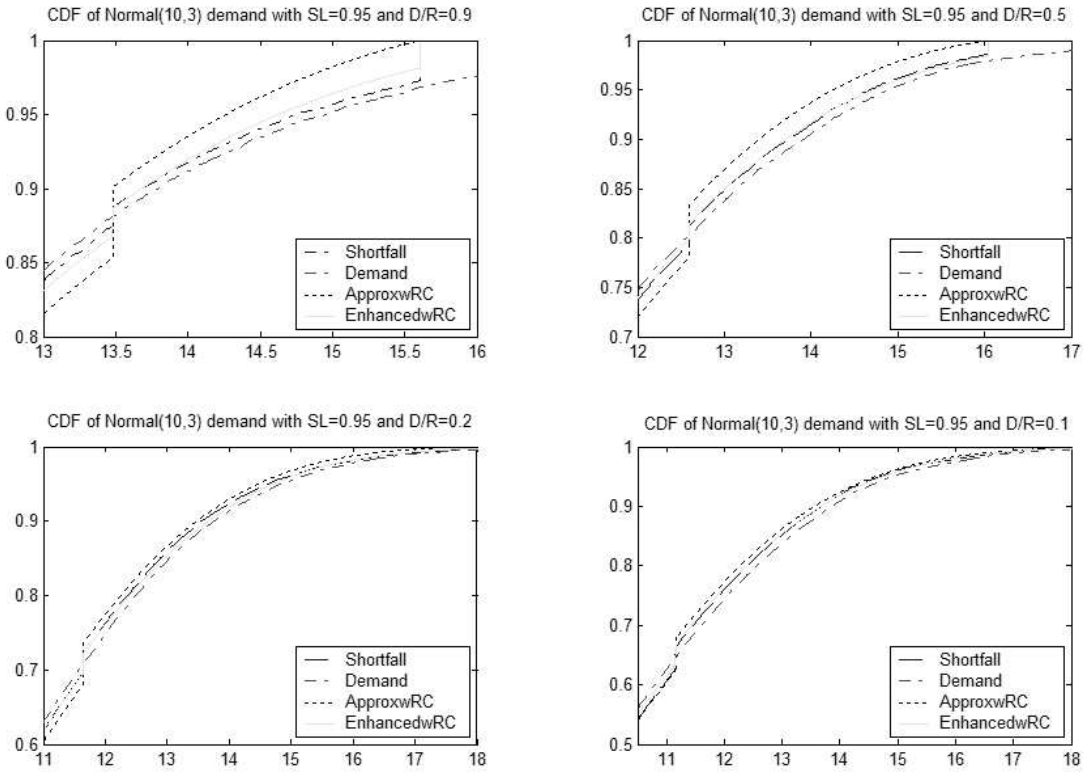


Figure 5-4: Zooming on the range $(r, d + r)$ of Figure 5-3: the comparison between the CDF of buyer's demand, drawn from a normal(10,3), to the output of the token bucket with rate control, the prediction based upon the Brownian motion approximation, and the prediction based on the enhancement. The requested service level is of percentage of periods with loss smaller than 95%. The cost ratios D/R of depth-to-rate are 0.1, 0.2, 0.5, and 0.9.

First Moment and Stdev Comparison for Normal(10,2) Demand, SL =95% with D/R=0.9, 0.5, and SL=99%with D/R=0.2,0.1

			Token Bucket					Token Bucket with Rate Control			
SL	D/R	Measure	Simu	ApproxFR	Approx	EnhanceFR	Enhance	RealMean	Simu	Approx	Enhance
95	0.9	Mean	9.94	9.44	9.95	9.56	9.97	10.00	9.98	10.06	10.01
		Stdev/2Mom	1.94	X	102.62	X	103.04	X	1.94	104.77	103.88
		STDev	1.94	X	1.90	X	1.93	X	1.94	1.88	1.92
		ErrorMean	0.50%	-5.05%	0.05%	-3.87%	0.21%	0.50%	-0.24%	0.62%	0.10%
		ErrorSTDev	X	X	-2.04%	X	-0.60%	X	X	-3.23%	-1.07%
	0.5	Mean	9.94	9.72	9.94	9.82	9.97	10.00	9.97	10.02	9.99
		Stdev/2Mom	1.94	X	102.49	X	103.15	X	1.93	103.84	103.61
		STDev	1.94	X	1.90	X	1.94	X	1.93	1.86	1.93
		ErrorMean	0.50%	-2.20%	0.01%	-1.21%	0.26%	0.50%	-0.25%	0.20%	-0.06%
		ErrorSTDev	X	X	-1.99%	X	0.28%	X	X	-3.72%	0.26%
99	0.2	Mean	10.00	9.97	9.99	9.97	9.99	10.00	10.01	10.00	10.00
		Stdev/2Mom	2.01	X	103.73	X	103.83	X	2.01	103.88	103.95
		STDev	2.01	X	1.98	X	1.99	X	2.01	1.97	1.99
		ErrorMean	0.52%	-0.35%	-0.15%	-0.31%	-0.11%	0.52%	0.11%	0.00%	0.00%
		ErrorSTDev	X	X	-1.63%	X	-1.25%	X	X	-2.06%	-1.14%
	0.1	Mean	10.01	9.98	9.99	9.99	9.99	10.00	10.02	10.00	10.00
		Stdev/2Mom	1.98	X	103.83	X	103.83	X	1.98	103.91	103.95
		STDev	1.98	X	1.99	X	1.99	X	1.98	1.98	1.99
		ErrorMean	0.51%	-0.28%	-0.17%	-0.26%	-0.18%	0.51%	0.16%	0.00%	0.00%
		ErrorSTDev	X	X	0.35%	X	0.44%	X	X	-0.07%	0.46%

Figure 5-5: Results and their errors for approximating the expectation and standard deviation of the effective demand output from a token bucket and a token bucket with rate control, for Normal(10,2) demand with service level requirements of percentage of periods with losses smaller than 95% and depth to rate cost ratio D/R of 0.9, 0.5, and for a service level requirement of 99% with D/R ratio of 0.2, 0.1.

Figures 5-3 and 5-4 are similar to Figures 5-1 and 5-2 for the TBwRC cases, for normal (10,3) demand with service level of 95%. The effective demand is noted as shortfall (since the shortfall describes the one-sided regulated random walk). Again, the enhancement did better than the approximation.

Figure 5-5 compares the simulations' results and the analytical ones (developed in this chapter) for different service levels and depth-to-rate cost ratio D/R. The interpretation of the different cells is as follows:

- "Mean" is the expectation of the effective demand. "STDev/2Mom" indicates the standard deviation or the second moment of the effective demand.

"ST Dev" is the standard deviation of the effective demand.

- TB case: "Simu" indicates results based on the effective output of the simulation. The "Error Mean" is the 99% confidence interval for the simulation. "ApproxFR" is the approximation of the expected effective demand based on the fill rate predicted by the Brownian motion approximation in equation (5.14). "Approx" indicates results based on the Brownian motion approximation. "EnhanceFR" is the approximation of the expected effective demand based on the fill rate predicted by the enhancement in equation (5.24). "Enhance" indicates results based on the enhancement.

- TBwRC case: "Real Mean" is the theoretic mean of the effective demand that is known since this is a backlog model. The "Error Mean" in this column is the 99% confidence interval for the mean calculated from the simulation. The "Simu," "Approx," and "Enhance " are as in the TB case. The "Error Mean" field in the "Simu" column is the error from the theoretical mean.

In the table, the error is defined as the measure minus the simulation (or true mean in the TBwRC case) normalized by the simulation result (or true result, when available). From the table it is seen that all of the approximations are rather close to the actual moments and, in general, the predictions of the enhancement are better. Similar results hold for the additional experiments we ran, and are not shown here.

5.4.2 Exponential Demand

We use an exponential demand with mean one and the same service level requirements and depth-to-rate cost ratio parameters as in the normal demand cases. We compared the resulting approximations, of the CDF, expectation, and standard deviation, to the simulation ones.

Earlier we showed that the optimal solution for the buyer's problem, when her demand is exponential and the cost ratio $D/R = 0.9$, is to choose $d = 0$, for service level requirement lower than 99%. Therefore, in these cases the fill rate approximated on the basis of the Brownian motion is not valid. Thus, when using Algorithm 13, we need to go through all its steps. Fortunately, in these cases buyer's solutions provide exactly the requested service level (in the TBwRC case). In addition, Corollary 14 shows that, for exponential demand, the percentage-of-periods-with-losses equals the fill rate. These two factors make our approximation for the

token fill rate better, despite the Brownian motion fill rate not being valid. However, cases in which the fill rate predicted by the Brownian motion (and its enhancement) is negative or larger than one can happen, when d is very small. In such cases, the approximated token fill rate resulting from steps three and four of Algorithm 13 might be less accurate.

Despite this difficulty, the numerical results, for the exponential case, show that the predictions made by the enhancement are within 1.6% of the simulation results for the mean, and within 2% for the standard deviation. Typically, the results are much better. But we consider these results excellent, and in fact, they are within the accuracy of the simulation. The approximations for the standard deviation for the TBwRC case are of the same quality in the cases where the bucket depth is not zero.

Figure 5-6 gives plots of the CDF based on the simulation and our approximations, for the TBwRC case, for exponential(1) demand with a service level of 90%. The effective demand is noted as shortfall (since the shortfall describes the one-sided regulated random walk). Again, the enhancement is looking better than the approximation. The figure with results for the TB case is similar and therefore is omitted.

Figure 5-7 compares the simulations' results and the analytical ones (developed in this chapter) for a service level of 90% and depth-to-rate cost ratio D/R of 0.9, 0.5, 0.2, 0.1. The names of the fields are as in the normal case, and the qualitative interpretation of the results is again that the enhancement is better than the Brownian motion approximation in predicting the standard deviation of the effective demand.

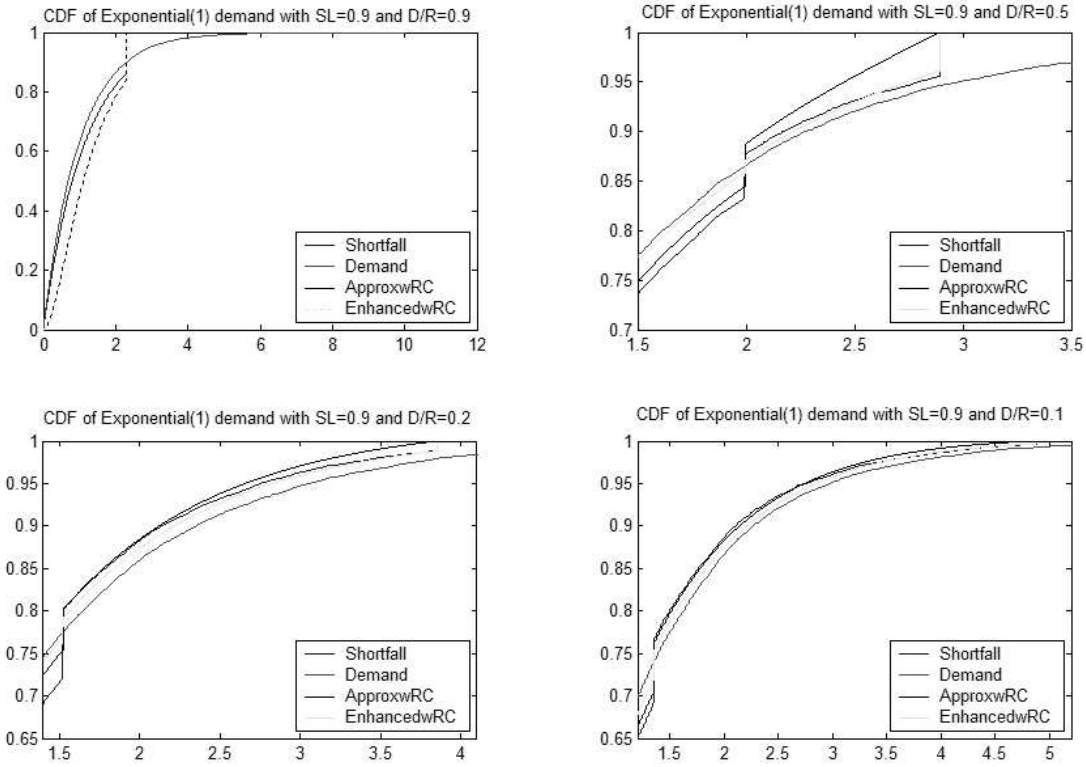


Figure 5-6: Zooming on the range $(r, d+r)$ of the comparison between the CDF of buyer's demand, drawn from an exponential(1), to the output of the token bucket with rate control, the prediction based upon the Brownian motion approximation, and the prediction based on the enhancement. The requested service level is of percentage of periods with loss smaller than 90%. The cost ratios D/R of depth-to-rate are 0.1, 0.2, 0.5, and 0.9.

First Moment and Stdev Comparison for Exponential(1) Demand, SL =90% with D/R=0.9, 0.5, 0.2, and 0.1

			Token Bucket				Token Bucket with Rate Control				
SL	D/R	Measure	Simu	ApproxFR	Approx	EnhanceFR	Enhance	RealMean	Simu	Approx	Enhance
90	0.9	Mean	0.90	0.90	0.90	0.90	0.90	1.00	1.00	1.22	1.22
		Stdev/2Mom	0.73	X	1.34	X	1.34	X	0.77	1.99	1.99
		STDev	0.73	X	0.73	X	0.73	X	0.77	0.71	0.71
		ErrorMean	2.09%	0.04%	0.04%	0.04%	0.04%	1.97%	0.13%	22.21%	22.20%
		ErrorSTDev	X	X	-0.21%	X	-0.21%	X	0.00%	-7.88%	-7.88%
	0.5	Mean	0.94	0.80	0.92	0.91	0.94	1.00	1.00	0.97	0.95
		Stdev/2Mom	0.81	X	1.45	X	1.52	X	0.81	1.53	1.56
		STDev	0.81	X	0.77	X	0.81	X	0.81	0.76	0.80
		ErrorMean	2.21%	-14.83%	-1.99%	-2.80%	-0.52%	2.09%	0.13%	-2.57%	-4.63%
		ErrorSTDev	X	X	-4.17%	X	-0.25%	X	0.00%	-6.60%	-1.14%
	0.2	Mean	0.97	0.95	0.95	0.99	0.96	1.00	1.02	0.97	0.98
		Stdev/2Mom	0.88	X	1.60	X	1.68	X	0.85	1.61	1.71
		STDev	0.88	X	0.84	X	0.87	X	0.85	0.82	0.87
		ErrorMean	2.33%	-2.41%	-2.78%	2.06%	-1.61%	2.15%	2.01%	-2.96%	-2.50%
		ErrorSTDev	X	X	-4.71%	X	-0.77%	X	0.00%	-3.97%	2.59%
	0.1	Mean	0.97	0.97	0.96	0.99	0.97	1.00	1.00	0.98	0.99
		Stdev/2Mom	0.91	X	1.72	X	1.78	X	0.88	1.70	1.79
		STDev	0.91	X	0.89	X	0.92	X	0.88	0.86	0.91
		ErrorMean	2.43%	-0.08%	-0.53%	1.68%	0.12%	2.27%	-0.21%	-1.80%	-1.36%
		ErrorSTDev	X	X	-2.36%	X	0.38%	X	0.00%	-2.37%	2.74%

Figure 5-7: Results and their errors for approximating the expectation and standard deviation of the effective demand output from a token bucket and a token bucket with rate control, for Exponential(1) demand with service level requirements of percentage of periods with losses smaller than 90% and depth to rate cost ratio D/R of 0.9, 0.5, 0.2, 0.1.

Chapter 6

Summary and Future Research Directions

6.1 Introduction

Today, many businesses outsource their computing needs (web-hosting, record keeping, databases) to external service providers that usually install dedicated computer resources and charge for hardware, software, and support for these resources. In the future, such service providers will share resources between customers to realize economies of scale. In such a case, pricing computing services might need to be usage-based.

An additional factor that raises the interest in managing of shared resources is the increasing interest in grid computing, or a utility data center. Grid computing is defined (Gridcomputing, 2003) as "a type of parallel and distributed system that enables the sharing, selection, and aggregation of geographically distributed 'autonomous' resources dynamically at runtime depending on their availability, capability, performance, cost and users' quality-of-service requirements." The notion of grid computing has been an active research area in leading universities for over a decade. In the last few years this notion has also enjoyed the attention of leading computer companies, such as HP, IBM, and Sun Microsystems. For the interested reader many articles and white papers on the subject of grid computing can be found on the Internet by searching for the term "grid computing" at the sites of the above companies, at (Gridforum, 2003), (Sharcnet, 2001), and (Gridcomputing, 2003), and at references therein.

This dissertation has discussed the pricing problem faced by the buyers and sellers of shared services. We has presented the TB and TBwRC admission controls and introduced their use as pricing schemes. We have investigated the application of TB mechanisms as pricing schemes by: looking at the resources availability provided by TB admission controls, solving the buyer's problem when facing such pricing schemes, and approximating the output process from TB admission controls. Yet the implementation of TB pricing schemes requires additional work.

In this chapter, we provide a qualitative comparison of pricing schemes mentioned in the work with TB pricing schemes, outline questions related to pricing of shared services and to the implementation of TB mechanisms as pricing schemes, and give an overview of the main results of this dissertation.

6.2 Comparison of Pricing Schemes

This section offers a qualitative comparison among the different pricing schemes mentioned in terms of the important attributes discussed in section 1.3. It is an extended version of the comparison provided there, which includes TB pricing schemes in addition to the pricing schemes in the original comparison. The results of this comparison are summarized in Table 6-1. Since this is only a qualitative comparison we rank the different pricing schemes as best, very good, good, medium, and worse. These ranks should be recognized as informs judgements rather than actual performances on each attribute. Finally, it is important to emphasize that these rankings are from the author's point of view and are based on a discussion with a small group of practitioners and academics. We think that this ranking is reasonable; however, we strongly recommend more detailed study of the subject.

The conclusion of the comparison given here is that TB mechanisms appear to be good candidates to be used as pricing schemes. Still, implementing such schemes will require additional work. Therefore, in the next section, we summarize research questions that remain to be addressed in order to make TB pricing schemes more practical, .

As can be seen from the table, none of the pricing schemes is perfect (or even just the best) in each of the required attributes. However, TB schemes performs fairly well in all of them, proving that it might be a good pricing scheme for shared resources. The remainder of this

Comparison of Pricing Schemes

	Fixed cost	Smart market	Flexible ser.plan	95/5	Token bucket
Supports demand variability	Good	Best	Very Good	Good	Very Good
Supports service level guarantees	Worse	Best	Good	Medium	Very Good
Supports resource planning	Worse	Very Good	Very Good	Medium	Best
Gives incentives	Worse	Best (*)	Good	Medium	Very Good
Easy to understand and operate	Best	Worse	Best	Best	Best
Information requirements	Best	Worse	Good	Very Good	Medium
Known costs / revenues	Best	Worse	Good	Medium	Very Good
Cost of demand's profile changes	Best	Worse	Best	Best	Best

Figure 6-1: A qualitative comparison between the different pricing schemes, in terms of the important attributes for a shared resources scheme. (*) is not in seller's control.

section briefly discusses the rankings in the table.

For the attribute of allowing high demand variability, the smart market scheme performs best, since the price that buyers pay is independent of their demand and its variability. We consider both the flexible service plan and TB schemes as very good for this attribute, since, in both, buyers can easily purchase a higher level of resources, according to their needs. In fact, an integral part of both schemes is the consideration of demand variability. We rank both the fixed cost and the 95/5 schemes as having good performances as well, since in both the price for the additional demand does not vary greatly.

For the attribute of supporting service level guarantees, the best scheme is the smart market, because it guarantees the right service levels to all buyers, according to their evaluation of the services provided. Token bucket schemes are very good in supporting service levels guarantees, since buyers who choose the rate and depth parameters correctly are assured of getting their requested service levels. However, the choice of these parameters is not trivial, and the service level is only locally optimal; thus, TB schemes are not ranked first. The flexible plan can also provide good service level guarantees, but since buyers can change their demand profile without early notice, we think it is not as good as TB schemes. For this attribute the 95/5 and fixed price schemes are medium and worse correspondingly, since in both there are no constraints

on buyers' demand variability, yet the first gives a cost incentive to buyers not to substantially vary their demand too often.

For the attribute of supporting resource planning TB schemes are ranked first, since it limits the largest peak in demand that sellers should be ready for. Both the smart market and the flexible service plan schemes supports resource planning in a reasonable manner. The smart market gives buyers the right value of resource expansion; however, it does so only after the demand realization. The flexible service plan confines buyers' average demand and its peak. The 95/5 gives only a partial support for resource planning, since the 95^{th} percentile of buyers' predicted demand should be known by sellers, but the peak demand is ignored. The fixed cost scheme is the weakest scheme in this measure.

For the attribute of giving buyers the right incentives, the smart market mechanism does best; however, these incentives are not controlled by the seller. The TB schemes are ranked second, since the buyer's choice of the rate and depth parameters is influenced by the seller's choice of their price. Thus, it is in the buyers' best interests to choose and smoothen their demand, as well as to accurately report their predicted demand. The flexible service plan gives incentives to buyers to truly report their expected (predicted) demand as well as to smoothen it, yet it is hard to relate the usages of different resources using this scheme. Again, the 95/5 gives buyers some incentives, which is better than what is done using the fixed price scheme.

For the attribute of simplicity to understand and operate, we rank all schemes except the smart market as best.

For the attribute of information requirements, the fixed cost requires only expected demand to decide on price, and therefore it has the lowest requirements, and is ranked first. The information requirement is increasing when we go to the 95/5 scheme, which also requires a knowledge of the 95^{th} percentile. A further increase is required for a good implementation of the flexible service plan (buyers should know when they have higher demand). More information is required by buyers in order to use TB schemes; however, given buyers' choice of the TB parameters, the knowledge required from sellers is low. Clearly, the demand information required by the smart market scheme is the highest.

For the attribute of fixed costs and revenues, our ranking is clear.

For the attribute of cost of demand profiles changes, again all schemes but the smart market

got a good score (low cost of changes). In this case the smart market mechanism is deficient since the cost of such a change is unknown by both sellers and buyers; hence it is ranked last.

6.3 Future Research Problems

There are many directions for research on the subject of computer services and their pricing on either a dedicated or shared basis. This section lists some of these problems, some of which have a qualitative nature and others a quantitative one. This section has two subsections. The first includes research questions related to the provisioning of shared computer services, and the second includes problems related to the implementation of TB pricing schemes.

6.3.1 Provisioning of Shared Computer Services

One of the main challenges in the provisioning of shared computer services is that the needs and value of these services are correlated among the resources (computer power, bandwidth, and data base). Thus, this question complicates most of the problems below and in their description we do not mention it.

Mapping the Computer Services Supply Chain

In Section 1.2, we mapped the computer services supply chain. In this mapping we ignored the vertical and horizontal integrations of these companies and the business relations among them. Such a mapping is vital in order to understand the problems faced by the industry of computer services, including the pricing problem we focused on in this dissertation. For an example of the importance of such a mapping, one can look at HP. Today, HP is primarily a member of the supply chain's first layer of hardware and software providers. As such, it is interested in increasing the sales of this layer. On the other hand, HP is also interested in serving customers in more economical ways, such as by provisioning shared computer services. However, the main premise of providing shared (rather than dedicated) services is resources savings. Thus provisioning of shared services will decrease the sales of the first-layer participants. Thus, examining the place of HP in the supply chain raises questions of their business objectives.

The mapping of the computer services supply chain needs to consider the different services

that are included in a specific supply chain and to understand further the resources involved in providing these services. This complicated task involves talking to different players in the chain. An example of questions that might help to start up such interviews is provided in Appendix A.2. In addition, the work of (O'donnell, 2001) that focuses on the money flow in the Internet can be used as a starting point for such a mapping.

Technological Requirements for Providing of Shared Services

This work ignores the technological aspects of providing shared computer services, although such a barrier is of high importance. The technological provisioning of shared computer services includes three challenges. First, a technology that can manage the operation of large-scale computer centers must be developed. Such a technology will dynamically allocate resources to different operations. (A dynamic resources allocation mechanism that takes into account different service levels for different tasks is a Non Polynomial Complete problem). A second part of the technological question is the operation of the above "managing system" such that privacy and security of different users will be guaranteed. The third part of this problem is the implementation of the above software without using too many resources, avoiding the risk of significantly discounting the resource saving from providing shared services.

Understanding the Buyer's Needs

In the selling of any product, and even more in the selling of services, quantifying customers' needs is a challenge. When providing shared services, this problem is even more significant, because we need to understand customers' needs from a different method of service provisioning than the customary method today. Moreover, customers' needs depend on their core business; thus the variety of these needs is large.

Another important aspect of understanding buyers of computer services is the variation, in the timing of their needs. One of the aspects of this variation is the demand variation and its modeling is the next research problem we present.

Modelling the Demand Process

The motivation for providing computer services on a shared basis stems from low average usage levels of dedicated computer resources. Thus, aggregating the demand of a few customers may result in a smoother demand function. A smooth demand means a lower ratio between average demand and peak demand, and therefore sellers would be able to satisfy the same customer needs using a lower quantity of computer resources.

In order to quantify the savings from shared resource services, one needs to know the demand patterns of each customer and the correlation between the demand patterns of different users. Thus, a mathematical model of customer demand can help in estimating such cost savings. This problem should have a "field research" component of monitoring resource usage levels and a theoretical part of suggesting realistic demand models.

As a starting point, the customers' demand process is a discrete time stochastic process that can be modelled by a known discrete process or approximated by a continuous one.

The main properties of the actual demand process, X_t are:

1. $0 \leq X_t \leq M$, where M is some upper limit for the demand value.
2. X_t is a discrete stochastic process.
3. X_t gets only real values.
4. There is a dependence between X_t and $X_{t+\delta}$, where δ can be in the order of hours, days, weeks, etc.
5. There might be other dependence across different intervals.

Mapping Buyers' Needs to Resource Requirements

In order to examine the benefits of providing computer services in a shared manner one must take into account buyers' needs in terms of demand variation and service levels. Thus, we suggest here a simple, vector-based representation of these needs and their mapping into resource requirements.

We ignored this problem in the dissertation, by assuming that the buyer's needs are described only in terms of resources availability. Such an assumption is helpful when modeling

buyers' behavior. However, it is too simplistic for practical purposes. We consider the process of mapping buyers' needs to resource requirements as a necessary step in transferring a pricing scheme from theory to practice. Moreover, such a mapping is necessary to verify that provisioning shared services will lead to resource savings.

Customer needs can be partitioned into qualitative and quantitative needs. The qualitative ones have a broad range of importance but are hard to grasp mathematically. One of the most important needs of computer services users is the privacy and security of their data. This security means, for example, that buyers can use their customers' credit card data without being concerned that this information will be used to the customers' detriment by other parties. Thus, sellers must provide data security and privacy under shared resources, or else the shared service will not succeed. The quantitative needs are defined as a combination of the demand and service levels, and we recommend modelling them using a vector-based representation. In the rest of this subsection we discuss the mapping of these needs to resource requirements.

In order for shared computer services to be feasible, the basic agreement between the seller and the buyer should include both the demand and service level distributions.[1] Both demand and service levels should be considered as random variables rather than deterministic ones. Examples of parameters that can be used to measure the distributions of the demand and service levels are:[2] mean / median / mode / quantile / variance / etc. The buyer's needs have quantitative as well as qualitative aspects.

We propose to capture the distributions of demand and service levels using a vector of pre-specified **needs-fields**. We refer to this vector as the **needs-vector**. Thus, sellers define a vector with fields of possible customer needs that they can satisfy (e.g., the daily expected number of hits on the web page) and buyers specify the **need-level** for each such field using one or more of the statistical parameters described above. This proposed modelling of customer needs is the same for all the services provided by a seller.

As a first step to achieve such a representation of customer needs, one needs to define all of the relevant needs-fields. This subject requires input from both sellers and buyers (ASPs

[1] Here we use the term "distribution" to specify different levels of demand or service level, despite the fact that in some cases full information on the distribution is not required.

[2] These and other parameters were mentioned by Angela Hung in her primary document, from July 2001, that was sent to MIT from HP.

and companies). We suggest a simple structure of partition to guarantee a high service level whenever customer demand are within a given range (as defined in the needs vector), and a best-effort service whenever customer demands are above this specified range. However, from a mathematical perspective, the meaning of each field is not important; moreover, this model is flexible enough to allow for an increase in the number of needs-fields. Thus, it allows the definition of different levels of resource consumption and their corresponding requested service level.

One of the advantages of a vector representation for customer needs is that it helps customers define their needs according to this framework. However, its weakness is that such a representation cannot handle qualitative requirements that cannot be measured quantitatively, such as the professional level of the sellers' maintenance team. Despite the importance of such qualitative needs, their character makes it difficult to reasonably model them in a quantitative framework as suggested above; thus this weakness is something one should expect.

The main advantage of providing shared computer services is the statistical aggregation of demand. Thus, sellers must know the statistical behavior of buyers' demand in order to benefit from this aggregation. The literature we examined has no discussion of demand patterns for computer power and database usage. Yet it is claimed (Leland et al., 1994) that bandwidth demand pattern has a self-similar nature, which means that there is a high dependency between its usage in different periods.

The second aspect of customer needs is service level. As with the demand issue, the literature we examined discusses this subject only for the bandwidth service. In the network arena extensive work has been done on service level agreements, especially regarding frame relay network.

Given a needs-vector, a seller should be able to compute the required resources for satisfying these needs. This task of estimating the required computer resources and their distribution based upon customers' needs-vector is a complicated one and a viable research topic. Yet we assume that there is a good way to get the final resource requirements using a mapping function.

The mapping function can be represented mathematically: $f : N \rightarrow R$, where N is the set of needs, and R is the set of resources. In order to capture the benefits of resource sharing we need the function f to satisfy the strict increasing return to scale property. That is, for any

$n_1, n_2, n_3 \in N$, such that $n_3 = n_2 + n_1$, $f(\cdot)$ will satisfy: $f(n_1) + f(n_2) > f(n_3)$. Note that since f is a function of a vector space, the inequality sign represents at least one resource for which the requirements for n_3 (a shared resource allocation) are strictly smaller than the sum of requirements to satisfy each one of the needs n_1, n_2 (dedicated resource allocation).[3]

One of the possible methodologies for mapping customers' needs to resource requirements can be using the Quality Function Deployment method, also called the House of Quality, in its services version.

There are two major weaknesses in using a mapping function. The first one is that a mapping function cannot catch qualitative requirements even if the last one can be measured in a quantitative manner. For example, time-to-market (i.e., how much time it would take before the service is available) can be measured by days; yet the relations between different time-to-market results cannot be translated to a level of computer resources.[4] The second weakness is that we ignore both uncertainties of the computer resource levels and the correlation of resource usage levels among different customers.

Modeling Buyers' Utility

The Game Theory approach models decisions using a utility function for each participant in a "game." Given these utilities, we can compute buyers' best response for any pricing scheme and find what is the preferred pricing scheme from the seller's point of view. The modeling of buyers' utility from different combinations of demand and service levels is a complex task. One possible representation of such a utility is described in (Schilkrut, 2000). His analysis focuses on individuals buyers' best response in terms of choosing their usage levels (number of uses and the length of each one), whereas companies' usage levels are less controllable. Moreover, in our case, the price buyers are faced with depends on a negotiation between them and the seller. Yet it might be possible to extend the analysis in (Schilkrut, 2000) and get some intuition regarding the expected preferences of buyers. One of the ways to cope with the negotiation problem is comparing the variances of sellers' profits when the expected values of the pricing schemes are identical. If, for the same expected profit, the variance of a certain pricing scheme is lower

[3] We might need the strict increasing return to scale to hold for each one of the resources.

[4] In fact, we do not think that such qualitative properties should be included in the pricing model.

than the variance of another one, the first pricing scheme will be preferred by the seller.

However, we suspect that a different representation of buyers' utilities is appropriate. One of the inputs for finding such a representation is the solution to the problem discussed in section 6.3.1. After having such a representation in hand, we can consider n identical customers with known stochastic demands, quality of service requirements, and utilities for the multi-attributes service. For simplicity, we assume that the demand distribution, customers' utility function, resource levels required to satisfy customer needs (the demand and service level pairs), and services provisioning costs, are all known by the seller. Given these settings, we compare the profits generated by different pricing schemes and choose the best one. Even then, there are two main approaches that can be taken by the Internet Service Provider. The first one considers only grouping present customers and the second one also considers grouping future customers.

Grouping only present customers, who at present receive a dedicated service, means that the seller has full knowledge of the computer resource levels required by each customer and can compute the required computer resources for all customers together. Then, sellers can find the cost savings they could gain from the decrease in the required computer resources levels, and the price discount they can offer to their customers. This approach has two weaknesses. The first one is that the discount offered to customers takes into account their willingness to change their service from a dedicated to a shared service, but ignores the fact that some customers will not want to do so despite the discount. The second weakness is that the pricing decision has been made based upon the savings from changing a dedicated service to a shared one, a cost which is hard to estimate. Moreover, such a pricing strategy does not allow sellers to use value pricing techniques. The strengths of the "grouping only present customers" approach are several. The customers are going to see cost savings compared with the cost they pay at present. Furthermore, the uncertainty regarding the usage level of present customers is lower, and the correlation between their usage levels is known.

Grouping present and future customers needs to focus on the decision a seller should make when a new customer (with a given profile of demand, service level needs, and utility) asks to join the service. It is similar to the problem faced by insurance companies. Insurance companies manage their risk pool in accordance with strict regulations regarding securities they should hold. Thus, for each new potential customer, the insurance company considers the

costs of the additional securities they need in order to insure this customer and compares them to the revenues flows expected from the customer. The insurance company will agree to sell insurance if the total benefit is higher than the threshold. Similarly, a seller can consider its acceptance\rejection decisions based on the additional resource cost in comparison with that customer's willingness-to-pay.

6.3.2 The Implementation of Token Bucket Pricing Schemes

In this dissertation we began analyzing the use of TB pricing schemes. However, there are still many open questions that need to be addressed in order to make TB pricing schemes practical. We list a few of these problems here.

Extensions of the Analysis in this Dissertation

We consider a simple version of TB pricing schemes. One factor that simplifies our analysis is our assumption that buyers' demands are drawn from a "nice" distribution and that they are independent over time. However, these assumptions are not likely to hold in practice. (See subsection 6.3.1 for a discussion that justifies this claim.) Thus, extending our analysis to more realistic demand patterns is crucial for the implementation of TB pricing schemes.

Additional extensions for the two basic models presented here (the TB with or without rate control) might also be considered. An example of such an extension is adding a maximum length of storing tokens, i.e., making tokens perishable. Such an extension will give buyers an additional incentive to choose a small depth parameter (since old tokens might be lost), and therefore will help to smooth demand even more. However, this would substantially complicate the analysis presented in this dissertation. A second example would be the coordination of demand streams that have different importance using different buckets. That is, how does a buyer choose a few buckets and rates to control a few services with different priorities, all of which use the same resource? A solution for this problem will probably suggest a nested system of buckets.

The last extension we list here is the periodic update of rate and depth parameter by buyers. Such changes can be considered either on a long or short horizon. In the latter, buyers can change the rate and depth parameters from one period to the other. Thus, the buyers' problem

will be dynamic in nature, where they look for an optimal policy (d, r) as a function of their bucket level and period's demand distribution. Such a formulation will allow a relaxation of the stationary demand assumption in our original buyers' choice problem formulation. However, the resulting dynamic programming problem will have a gigantic state space, which decreases our chances of getting any intuition from solving it. Moreover, one of the important attributes of a pricing scheme for shared resources is that the cost (revenue) will be known ahead of time; thus a complex buyer's choice problem formulation is prone to not satisfy this attribute. For changes over a long horizon and their importance, think of buyers who want to purchase extra computing capacity for the period before Christmas. Such a model will still do well regarding the importance of the fixed cost; however, it requires a further understanding of the demand for computer resources.

Pricing of Multiple Resources

One of the main limitations of the analysis of TB pricing schemes in this dissertation was that it only focused on pricing and admission controls for one resource. However, the provisioning of shared computer services is based upon a combination of a few dependent services (computer power, bandwidth, and data base); therefore it requires admission control and pricing schemes that are suitable for managing multiple resources in parallel. In this subsection we list a few of the challenges regarding the use of TB pricing schemes for multiple resources and a few possible paths towards solving this problem.

Giving Incentives

- During the period in which sellers can change prices (currently this period is a quarter up to a year), is there a resource which is more expensive/congested than others?
- Assuming there is a resource which is always more expensive/congested than the others, will the more congested resource be known by the seller at the beginning of the period?

We suspect that the answer to both questions above is positive because engineered systems are usually planned and operated with a bottleneck. Thus, we probably want to shift demand from the congested resource to the cheap ones.

- In such a case, will a fixed ratio or a fixed quantity of exchange be a good model for such a translation (of demand)? Or is there another appropriate model?
- What is the translation ratio, i.e., can one unit of A be translated into a requirement of 1.3 units of B (or maybe 0.7 such units)?

Demand as a Vector of Resources Requirements We consider two resources, A and B. Each job arrives with a usage vector of requirements (U_A, U_B).

- We need a better definition of a job. We cannot implement an admission control that throws out a huge job (say ERP) because a small fraction of its resource requirements is not free. This problem is of less importance when considering the TBwRC admission control.
- Is there a positive (or any other) correlation between the U_A and U_B? A positive correlation make sense, since if a job needs a longer processing time, it would probably require more storage and communication. This question is closely related to the buyer's demand model question mentioned in subsection 6.3.1.
- Could we group jobs according to their sizes (i.e., small, medium, etc.) such that, within each group, the demand for the different resources is not correlated? In this case, each group can be treated with its own TB scheme. However, there might be many groups, which will make the implementation of TB schemes more demanding. (For example, how do we differentiate between groups, both in prediction of demand and at their arrival to the resources pool?) A large number of groups also makes it hard to give buyers incentives to use one resource more than the others for all of the groups in parallel.

Length of Periods

- What length of period in which tokens enter the bucket makes sense?

This question is important for the one resource case; however in the multiple resources case, it is of higher importance. We might assume that the intervals between tokens arrivals' are so short that only one job request can arrive during an interval, but such an assumption may be too restrictive in practice.

If more than one job arrives to the system at one period, things get complicated. A short example can explain this: Assume that two jobs with requirements: $U^1 = (1, 2)$ and $U^2 = (2, 1)$ arrive at time 0 to the resources pool. Then, the total demand within this period is $(3, 3)$. Now, if there are enough tokens to process both jobs, everything is fine; however, if there are not enough tokens in one bucket, we have a choice either to accept only one job which can be completed (randomly chosen?) or to accept both jobs, one of which would stay in a job queue. In the TB case when we throw one job out (which looks simpler), we are left with the problem of computing the number of tokens left in the full resource.

In our example, if there are only two tokens in one bucket and three in the other one: $D_0 = (2, 3)$, we might choose to process either one of the jobs. If we choose the first one ($U^1 = (1, 2)$), we will have a loss this period (a loss of a job), but neither one of the buckets will be empty. In this case $D_1 = (1, 1) + (r^1, r^2)$. If we choose to process the second job ($(2, 1)$), we will be left with $D_1 = (0, 2) + (r^1, r^2)$, and the computation of the level of tokens in the full bucket depends on the number of tokens from this element that the processed work required. We suspect that, despite the independent assumption between the resource requirements of each job, the computation of the level of tokens in the second bucket (in our example) is not trivial. Here it looks as if the TBwRC mechanism is much simpler to analyze.

Aggregating Demand to Only One Resource A totally different view can be adopted by considering only one token bucket for all the resources. With this assumption we are back in a single resource case, for which the demand process is aggregated among a few resources. This solves most of the above problems: if no tokens are in place, the job is lost. In this case, the service level definition of percentage of periods with loss is in fact percentage of jobs lost (the real fill rate).

For this view it is necessary to have further discussion of the units of measure for each resource and how they can be scaled into one TB.

Aggregated Demand

The next step towards solving the ASP pricing problem requires investigating the demand process resulting from the demand aggregation of many users who use the TB scheme. In the

network literature, there are many simulation studies that look at this aggregated demand. For an example see (Berger, 1991). Unfortunately the results reported are usually for one demand stream and then the aggregation of hundreds and thousands of identical demand streams. In contrast, ASPs are looking at the aggregated demand from a small number of different customers (typically 20 to 40).

Using the results we developed in Chapter 5, we can characterize the demand faced by the seller as the aggregation of the effective demand from each buyer. Using these results and assuming that buyers' demand is independent (among them), we can approximate the requested demand. However, it might be beneficial to approximate this as a normal demand (because it is aggregated from many sources); thus, we could use only the first two moments of a buyer's effective demand to approximate the seller's one.

Before taking any approach to model the effective demand faced by a seller, we also recommend running simulations to show that this demand is smoother when using TB admission controls than when not using them. Our simulation results for the effective demand of one buyer hint that the coefficient of variation of a buyer's effective demand is lower than the one of a buyer's demand. However, a more detailed study is required and should consider combinations of different relations between buyers (dependencies, number of buyers).

The Seller's Problem

The discussion in Chapter 2 presented the seller's problem as a revenue maximization problem. It also discussed the seller's profit-maximizing problem and its dependency on the seller's resource planning problem. We see a solution to either of these problems as an essential step towards the implementation of TB pricing schemes.

One of the issues that needs to be considered, in relation to the seller's problem, is this: since the number of buyers is small (hence, each buyer is a significant customer), the price charged for computer services is negotiated with each buyer independently. Despite this fact, we think that a theoretical solution of the seller's problem can model buyers as if they all face the same prices. We hope that such a model could catch the qualitative elements that would influence the seller's pricing choice. For example, we can stipulate that the seller's solution will only depend on the ratio of rate to depth prices.

Secondary Token Market

We consider here a token market operated by the seller. In such a case, buyers could purchase tokens (to fill their bucket) at a spot price, whenever their bucket level is low or when they predict a higher usage level. As we remarked in Chapter 2, operation of such a token market is necessary. However, such a spot market will complicate the buyer's choice problem, by adding to it a portfolio management aspect. Furthermore, we see little sense in solving the seller's problem without considering the buyer's choice while a token market is in place.

The most straightforward implementation for a token spot market is by adding a dummy bucket that will accumulate tokens that are not used in each period. Recall that the seller has a finite capacity and therefore, a finite rate of tokens in each period. Thus, at the beginning of each period the seller sends the tokens to each client according to their fixed rate choice, and sends the extra tokens to the dummy bucket. These tokens can now be purchased by buyers through an auction or any other selling mechanism. At the end of each period the unused tokens in this extra bucket are destroyed. Thus, buyers cannot purchase tokens that will overcommit the seller (or one can think of some sort of overbooking with these tokens).

Collaboration among Buyers Here, instead of a "legal" market operated by the ASP, two or more buyers can cooperate and share their rate and depth. These customers then can be considered by the ASP as one customer with their aggregated demand. Intuitively, it may look as if such a collaboration will be beneficial for the buyers since they can, at least partly, enjoy the statistical aggregation of demand. However, assuming no legal prevention of such collaboration, in equilibrium, the costs of managing such agreement should be high enough to prevent them. Otherwise, all buyers would collaborate and the ASP would face only one stream of demand. This is unlikely to happen if the TB scheme is indeed an efficient way to manage aggregated demand of multiple sources.

6.4 Overview of Main Results

This dissertation presents the application of TB admission controls to pricing of shared resources. In Chapter 1 we presented the pricing problem of Application Service Providers

(ASPs) when they consider the provisioning of services based on shared computer resources. One of the characteristics of this problem is its focus on pricing for a relatively small number of customers, which allows for a fine market segmentation by provisioning a highly tailored service. Thus, pricing computer services differs from traditional pricing problems, which focus on pricing of standard "tangible" products to a large number of customers. Section 1.3 provided a five-step framework for analyzing pricing of shared resources. It includes a short survey of the literature and practice of pricing of computer services, and a qualitative comparison of existing pricing schemes. The chapter's conclusion was that none of the pricing schemes surveyed is perfect for pricing of shared computer resources. Moreover, from the qualitative comparison of pricing schemes, we saw that the combination of admission control and pricing is beneficial. Chapter 2 presented TB admission controls and their potential uses as pricing schemes for shared services. We consider the idea of using TB admission controls as pricing schemes to be one of the main contributions of this work. Chapter 2 also considered the implementation of TB pricing schemes, and outlined the research questions that are addressed in the rest of the dissertation, all of which focus on using TBs pricing schemes for the pricing of a single shared resource.

Chapter 3 analyzed the performance of TB admission controls, where performance is considered as resource availability. It modeled the bucket level as a regulated random walk and used results from the theory of large deviations to provide bounds and asymptotics on the resources availability, when TB admission controls are used. One of the highlights of this work, which was given in this chapter, is a new bound developed for the loss probability of a two-sided regulated random walk. Chapter 4 used the results developed in Chapter 3 to solve the problem faced by buyers of a service, when it is priced using TB pricing schemes. The main tool used in order to find bounds on the solution to the buyer's optimization problem was a proof of convexity of a constrained version of the buyer's problem. The chapter demonstrated the use of the techniques developed in Chapters 3 and 4 on the exponential and normal demand cases, where closed-form solutions and approximations for the buyer's solutions were given. A numerical study reported that the results obtained for these demand cases were in the range of 1% away from optimal in the TBwRC case and of 8% away from optimal in the TB case. Chapter 5 characterized the output stream from TB admission controls. The aggregation of

this stream among many buyers is the effective demand faced by the seller, when using a TB pricing scheme. This chapter's results, which are based on a new approach towards characterizing the output process from TB admission controls, are another major contribution of this work.

Finally, in the present chapter we compared TB pricing schemes to the pricing schemes discussed in Chapter 1.3. This qualitative comparison, from the author's point of view, showed that TB pricing schemes do well on many of the important attributes for a pricing scheme for shared resources, and therefore we conclude that TB pricing schemes might be suitable for such pricing. However, the implementation of TB pricing schemes for shared resources requires answers to many questions that are outside the scope of this dissertation. We have outline some of these questions in this chapter, and we intend to pursue them in the near future.

Appendix A

Appendixes

A.1 Service Methods

This section presents possible methods for companies and organizations in the use of computer resources.

There are different methods by which computer services can be provided. The first important distinction is between dedicated computer services and shared ones. This distinction sounds clear; however, in reality, most of the cases are found somewhere between these two extreme points. That is, some of the computer resources are dedicated, e.g., the CPU, and some of them are shared, such as bandwidths.

The second distinction that exists in today's computer services is between Collocation and Managed Hosting. According to (Hurwitz, 2001), "Collocation refers to an approach whereby a customer 'collocates' its servers along with that of other customers in an environment in which they can take advantage of the provider's floor space as well as access to network bandwidth." Consequently, Managed Hosting describes customers that have no physical access to servers' infrastructure. Yet, we prefer to use the taxonomy shown in Table A.1: Different Service Methods. A first partition is between dedicated and shared resource services; the second partition of the dedicated service is formed according to the following:

At home –buyers own and operate the computer resource at their own location.

Collocated resources –sellers provide only the floor space and infrastructure; buyer owns and operates the equipment.

Self-managed –seller owns the resources, but buyer is responsible for their operation.

Full-managed –seller owns the resources and is responsible for their operation.

	Dedicated				Shared
	At Home	Collocation	Self-Managed	Full-Managed	Full-Managed
Location	Company	Provider	Provider	Provider	Provider
Ownership	Company	Company	Provider	Provider	Provider
Operation	Company	Company	Company	Provider	Provider

Table A.1: Different Service Methods

Today most computer services (except bandwidth) are provided on a dedicated basis.

However, the available capacity in the market is high at this times. Estimations indicates that the average utilization of computer resources is only in the range of 5-20%; such low usage rates cause high service cost. It is important to remember that there is high variability in demand for computer resources, so it might happen that even a system with an averages usage rate as low as 5% would be congested during periods of peak demand. The argument for extra capacity is that buyers must be assured of good service levels in periods of peak demand. Considering this reasoning, it might be advantageous to provide companies with computer services on a shared basis, with the intention of decreasing the cost of service by increasing usage levels.

A.2 Questions for Industry Representatives

The following are the first set of questions we would like to ask industry representatives. We divide them into three groups: surveying current practices, learning the demand, and verifying the legitimacy of suggested pricing schemes. We believe that before asking questions about the legitimacy of suggested pricing schemes, we need to process the answers for the first two subjects. Thus, there will be required at least two iterations with industry representatives.

Surveying Current Practices

1. What are the different segments of customers of computer services that can be recognized?

2. What are the needs (demand, and service level) of different customers' segments, and how do these needs differ among customer segments?

3. What are the pricing schemes used today for each resource (processing, memory size, memory access, and communication)?

4. How do today's pricing agreements define and enforce service levels?

5. Do the pricing schemes used today work well? How might they be improved?

6. How many computers and customers are in an IDC, and how many of them are in each segment?

Learning the Demand

1. Are demand profiles gathered by the providers and if so - how has it been done?

2. What is the pattern of companies' demand?

3. What is the probabilistic description (distribution, average, peak, other fractiles, standard deviation, etc.') of the usage of each resource?

4. What is the percentage usage of each resource on average and in peak times? (and does this depend on users' types?)

5. How is the above distribution correlated (along time and between different users), i.e. when there is (typically) a peak in demand (Monday morning?).

6. Can you map a vector of demand and service level needs of customers to resources level required to satisfy these needs?

7. Can you estimate what is companies (customers) utility from a given demand and service level, and how does this utility change with them?

Verifying the Legitimacy of Suggested Pricing Schemes

1. What is the technological feasibility of shared resource service's?

2. Will the time to market be longer for shared resource service than for dedicated ones? What are other problems in implementing shared resource service?

3. What other pricing scheme might be relevant for pricing of shared computer services?

4. Is a contract in the spirit of "token bucket" reasonable for each resource? why?

5. Would you as a buyer / seller agree to a pricing scheme of fixed price only, fixed price plus usage based, fixed charge with usage based fee and service level guarantees, and token bucket? Why (easy to implement etc.)?

A.3 Traditional Pricing Schemes

The pricing of any service is a complicated task. The high clock speed of the computer services industry adds a few difficulties to this task. The cost of computer resources is extremely hard to estimate.[1] This inability results in a difficulty to estimate customers' willingness-to-pay. Therefore, there is no wonder that today's pricing decisions are based neither on the cost of services provisioning nor on the utility to the customers. Thus, there is a lot of room for improving the pricing schemes used.

Pricing of shared computer resources further complicates the pricing decisions. On the one hand, under shared resources service, contracts should be restrictive enough to prevent buyers from overusing the resource pool, and prevent sellers from not allocating enough resources to the resource pool. On the other hand, such contracts should be flexible enough to allow for reasonable variations in both demand pattern and provided service level. Thus, two questions we are faced with, when considering pricing of shared computer resources, are what to consider as "reasonable" variation (by sellers and buyers for demand and service level) and how these definitions can be quantified and measured. The definition of service level specifications (latency, data availability, etc.) could be done once for all the prospective customers, but the

[1] Everyone knows that there is a big fixed cost element and a considerably lower marginal cost; but nobody knows the exact value of these costs. Thus, adding the difficulty of covering resources' marginal costs and contributing to the recovery of fixed costs (since these costs are hard to figure). Another challenge is measuring the benefits of customers from one usage level and Quality of Service (QoS) in comparison to another usage level and QoS.

service level should be negotiated with each one according to its needs and demand patterns, following the principles of value-pricing.[2]

A.3.1 Taxonomy of Traditional Pricing Schemes

In order to come up with a good pricing scheme we discuss the basic pricing schemes that we consider relevant for pricing of shared computer resources and the parameters it requires. For most schemes we assume that all the measures that are mentioned can be completely tracked on a per buyer basis (without harming buyers' privacy and security).

The first pricing scheme we consider is the **Fixed Price Only** (FPO), which is the simplest pricing scheme and it is the one most common today, when pricing dedicated computer resources. It is characterized by one parameter, f, the the membership fee, which is the fixed price for the whole package of services. Note that even this pricing scheme can be implemented only after the seller and the buyer agree on service level and demand behavior. Thus, the seller can estimate the level of resources required (to satisfy the service level distribution for the demand distribution agreed on), and then quote a fixed price for the service.

A more advanced pricing scheme is the **Fixed plus Usage Based** (FUB), which includes a cost for different levels of usage. In addition to f, the membership fee, this pricing scheme requires the definition of two parameters vectors, a needs (usage-levels) vector and a price-per-unit charged (in these levels) vector. This pricing scheme charges a price of P_1 per unit of the first X_1 units used, a price of P_2 per unit of the next X_2 units used, etc. The choice of levels and their prices should be such that for higher levels than the expectable ones, the prices will be significantly high, and will discourage buyers from overusing the common resources pool. It is similar to the pricing used by cell-phone and Energy companies. If the agreement includes service level of the form "the usage is measured only when the service level is higher than a threshold," then the different thresholds parameters should be defined as well. Note, it is possible that f, the fixed price, will cover the usage of X_1, the first units, by choosing P_1 to be zero. Furthermore, this pricing scheme is an image of different quantity discount pricing as is common in purchases of products and raw materials and, similar to the inventory theory, it

[2] Value pricing is the notion of fitting a package of products and services to a specific customer, in a way that he or she will not have a prior price estimation for the package. Such a package should then be priced according to the value it gives the customer.

can also be generalized to include any non decreasing convex price function per different usage levels.

A parallel pricing scheme is the **Fixed plus Service Level Guarantees,** which includes a discount vector (suggested by sellers to buyers), for cases in which the service level provided is lower than the agreed one as is represented in a service level vector. Again, this method can be generalized to include any non decreasing convex discount-function per different deviations from service levels. This pricing scheme is unrealistic, for the pricing of computer resources, since sellers cannot be "punished" for not providing the requested service level if buyers have no incentive to limit their usage. Thus, we ignore this pricing scheme in the rest of this document.

A combination of the last two pricing schemes is the **Fixed plus Usage Based and Service Level Guarantees** (FUSL). It includes a price, quantity, discount, and service level vectors. It is a symmetric pricing scheme. On the one hand, it gives buyers incentive to declare their right demand, in order to avoid high charges, thus allowing sellers to guarantee the service level provided. On the other hand, the FUSL Pricing Scheme gives sellers incentive to operate enough resources (for providing the demand and service level of all their customers), in order to avoid discounts.

For simplicity we name the four pricing schemes above: **Traditional Pricing Schemes**. The logic behind these schemes is as follows: Since the Fixed Price Only pricing scheme cannot enforce usage levels on the buyer, we enhanced it to the Fixed plus Usage Based pricing mechanism. In this pricing scheme, customers need to inform the ASP on their estimated usage level for each resource and a penalty for miss-estimating usage level might be appropriate. In parallel, in order to enforce the ASP to satisfy the required QoS, penalty cost (discounts) should be added, leading to the Fixed plus Usage Based and Service Level Guarantees pricing scheme.

The last pricing scheme we discuss are TB, which we mentioned already in Section 1.3.1. In the network literature, the TB idea was presented in the context of pricing Internet services. It prices the token rate and the bucket parameters, in order to differentiate between different users. For pricing computer resources, this differentiation method is unnecessary, since the nature of computer resources is that they are provided in a tailored manner. Thus, the seller goes through negotiations with each buyer, resulting in different values of contracts' parameters.[3]

[3]This negotiation is practical since today a typical seller serves 20 to 50 customers from one Internet Data

The implementation of TB methods in the control of other computer services deserves a further attention. For example, in order to implement TB for databases, we focus on the pricing of the memory access that are the database's main bottleneck. Since the characteristics of the memory usage are similar to bandwidth characteristics (one resource with a given capacity), the implementation of TB is similar to its implementation in network. However, in order to implement this method to CPU usage (computer power) we consider buyers that need an average of 10 CPUs but once a month need to run their Enterprise Resource Planning (ERP) software, which uses 100 CPUs per 3 hours. This differs from bandwidth since it includes the length of usage parameter (which is normally negligible in network), thus its implementation requires 3 parameters: rate, bucket and length of bucket usage.

A.3.2 Analysis of Traditional Pricing Schemes

In this Subsection we discuss the traditional pricing schemes, the analysis of TB pricing schemes is given in the dissertation's body.

One major advantage of the Fixed Price Only Pricing Scheme is, as suggested by McKnight and Boroumand in (McKnight and Boroumand, 2000), that many users are interested in an accurate prediction of the charges they are going to face with. Another advantage is that an accurate information on demand and service level distribution is unnecessary. The FPO Pricing Scheme needs only the information on different demand and service level pairs. Such information as peak demand and the requested service level in times of peak demand, (or expected demand and the requested service level in times of expected demand), is enough in order to compute the maximum resources needed to satisfy these demand service-level pairs. Its major disadvantage is that the cost of service is independent of the usage level.

Recall that the main advantage of shared services is an increase in resources usage levels, which cannot be achieved without smoothing demand that follows from demand aggregation. Sellers that use a Fixed Price Only Pricing Scheme cannot enforce usage level on their customers, and therefor cannot commit on service levels. In view of the above, in addition to the plausible assumption that buyers will not agree to share their resources without service level guarantees, the Fixed Price Only Pricing Scheme seems worthless for pricing of any shared service with

Center.

some capacity constraint. Yet, it might be beneficial for pricing of either dedicated services (such as storage size of a database) or shared resources without an active capacity constraint or when there is no importance for service level.

One of the difficulties in implementing the Fixed plus Usage Based scheme is the technological challenge in tracking buyers' usage. To solve this, a simpler version of the FUB can use some demand measures, such as: peak demand, demands' 95 percentile, etc., as surrogate of the actual usage. So, the needs and price per unit vectors will be defined according to this measures. With this version in mind, it is important to be aware that for such measures customer's marginal measure value is not their "real" measure value, due to the correlation between users' demands. Moreover, ASPs might be more interested in the value of the marginal measure, e.g. a company with peak demand on Sunday probably add no resource requirements, thus can get a lower price per average until of usage. To compute customer marginal measure, we borrow an idea from auction theory. For example, which uses peak demand as the usage surrogate measure: Compute $R_{\max}$ – the maximal resource level that is used by all customers during period t. Then, compute $R^i_{\max}$ – the level of resources that is used by all customers without the i^{th} one, during period t. Finally, r_i – the marginal contribution of the i^{th} customer to the resource requirement, is equal to the difference $R_{\max} - R^i_{\max}$, which is the i^{th} customer's effective peak demand.

The above idea can be used also to predict the additional resource requirements that a new customer will contribute to the resource pool. This will require knowledge of this customer's demand process and its correlation with demands of existing customers. Moreover, a generalization of this idea can be implemented in the Fixed plus Usage Based and Service Level Guarantees as well.

One of the strengths of the FUSL pricing scheme follows from the discussion in Section A.1. This discussion claims that it might be beneficial to consider the pricing of computer services as a problem of choosing penalty functions for charging both parties (ASP and their Customers) according to a comparison between a vector of their actual performance to the forecasted ones. In such a case, resource's price should consider the percentage of usage each customer is estimated to consume, but the pricing contract should be a function of the real usage and service levels in front of the planned ones. Of course, these penalty functions should

capture the actual usage rate for each resource as follows from the difference between estimated and actual demand. Moreover, buyers' penalty functions should take into account the penalty functions that the ASP is subject to, since a high demand for bandwidth by one customer reduces the available bandwidth for other customers and by so decreasing the QoS they get.

A.3.3 Qualitative Comparison of Traditional Pricing Schemes

Next we qualitatively compare the different pricing schemes suggested in this Section. But, before comparing them, it is important to note that the Fixed Price Only Pricing Scheme is a private case of the Fixed plus Usage Based Pricing Scheme, which is, in turn, a private case of the Fixed plus Usage Based and Service Level Guarantees Pricing Scheme. Thus, both the complexity and information requirements for the implementation of the proposed pricing schemes are increasing as we go from the Fixed Price Only to the elaborate Fixed plus Usage Based and Service Level Guarantees. For TB pricing schemes, there is a high level of information required by the buyer in order to choose the right rate and depth parameters. This information can be buyers' demand distribution, but as shown in Chapters 3 and 4, even then finding the optimal parameters is hard. However, given the rate and depth parameters, the information required by the seller is low and the implementation of TB pricing is easier.

This ease of implementation is partly a result from that TB pricing schemes is not flexible enough to allow large variations in demand patterns. This can be seen even in the case of deterministic demand, where buyers' optimal rate and depth parameters choice is independent of their price (as long as the bucket price is not absurdly high). In a first look, one might think that the Fixed Price Only scheme is flexible enough to allow for large demand variations. However, since demand is not known by the seller, a peak in demand might result in a decline of service because resources level is not high enough. Moreover, the same reasoning explains that the flexibility to demand variation is higher in the Fixed plus Usage Based and Service Level scheme than in the Fixed plus Usage Based scheme, where seller has no incentive to operate enough resources to satisfy peaks in demand. Today, the problem of flexibility to demand peaks is solved by over subscribing of buyers (Chang et al., 2001).

Table A.2 summarizes the discussion in this Subsection.

In light of the above comparison, our future work will focus on the implementation of the

Table A.2: Comparison Among Suggested Pricing Schemes

PS	Sellers' info. needs	Buyers' info. needs	Ease of implement.	Flex. to Demand Var.
FPO	Low	Low	High	Low
FUB	Low	High	Medium	Medium
FUSL	High	High	Low	High
TB	High	Low	Medium	Low

Fixed plus Usage Based and Service Level Guarantees (FUSL), and TB pricing schemes, since both look as if they can induce the right behavior from both parties (buyers and seller).

A.4 Proof of Theorem 6

In order to prove Proposition 6, we recall the Implicit Function Theorem in Lemma 17. Then we show that the conditions required for applying the Implicit Function Theorem are satisfied, and therefore s^* can be represented as an implicit function of r. Subsequently, we use this Theorem to compute the first derivative of s^* regarding r, and show that it is positive, proving part A of Theorem 6. Finally, we use the Implicit Function Theorem to compute the second derivative of s^* regarding r. Using this, and the strict convexity of the semi-invariant moment generating function $\gamma_x(s) \equiv \ln(G_x(s))$, (the semi invariant moment generating function is strictly convex if x is not a constant; for a proof see exercise 7.7 in (Gallager, 1996)) we show that $1/s^*(r)$ is strictly convex as well, proving part B of Theorem 6.

Lemma 17 *Implicit Function Theorem: For $F(r,s)$ defined on a rectangular: $r \in [a,b]$, $s \in [c,d]$. Let $M(r^*,s^*)$ be a point within this rectangular that satisfies:*

1. $F(r^*,s^*) = 0$.
2. $F(r,s)$ *is continuous and differentiable around* $M(r^*,s^*)$.
3. $F'_s(r^*,s^*) \neq 0$.

Then: Around M there is a neighborhood with a unique function $s = f(r)$, such that $F(r,f(r)) = 0$, and:

- $s^* = f(r^*)$.
- $f(r)$ *is continuous around* r^*.
- $f(r)$ *is differentiable around* r^* *and* $f'(r^*) = -\frac{F'_r(r^*,s^*)}{F'_s(r^*,s^*)}$

Proof. of Theorem 6. Let us show that s^* can be represented as an implicit function of r: By definition $s^* = \arg\left\{E_x\left(e^{s^*x}\right) = 1\right\}$, also $E_x(e^{sx}) = e^{-sr}E_u(e^{su}) = e^{-sr}G_u(s)$. Thus:

$$F(r, s^*) \equiv G_u(s^*) - e^{s^*r} = 0 \tag{A.1}$$

which satisfies condition 1 of the Implicit Function Theorem. Condition 2 of the theorem holds by our assumptions on $G_u(s)$. Verifying the third condition, $F'_s(r, s^*) \neq 0$, requires us to compute this partial derivative:

$$F'_s(r, s^*) = G'_u(s^*) - re^{s^*r}.$$

We will show that

$$G'_u(s^*) - re^{s^*r} > 0. \tag{A.2}$$

In order to show this, we recall that if $\gamma(s^*) = 0$ then $G_x(s^*) = 1$. In the point s^*, the moment generating function of x, $G_x(s^*)$, as defined in (3.2), is increasing thus:

$$G'_x(s^*) > 0.$$

We compute this derivative from 3.2:

$$G'_x(s^*) = e^{-rs^*}\left(-rG_u(s^*) + G'_u(s^*)\right) > 0$$

thus,

$$\left(-rG_u(s^*) + G'_u(s^*)\right) > 0. \tag{A.3}$$

Now, from condition 1 in (A.1) $G_u(s^*) = e^{sr}$, and substituting $G_u(s^*)$ with e^{sr} in (A.3) proves condition 3 in (A.2). Thus, s^* can be represented as an implicit function of r.

In order to prove part A of Theorem 6, we first compute $\frac{ds^*}{dr}$, proving the first claim of part C of the Theorem, using Lemma 17:

$$\begin{aligned} F_r'(r,s^*) &= -s^* e^{s^* r} \\ F_s'(r,s^*) &= G_u'(s^*) - re^{s^* r} \\ \frac{ds^*}{dr} &= -\frac{F_r'(r^*,s^*)}{F_s'(r^*,s^*)} = \frac{s^* e^{s^* r}}{G_u'(s^*) - re^{s^* r}}. \end{aligned}$$

Since s^* is positive, to show that it is strictly increasing we need to show that $F_s'(r,s^*) > 0$. However, we already showed this in (A.3) in our proof that condition 3 of Lemma 17 holds. Thus, part A of Theorem 6 follows.

For notational convenience, we suppress the $*$ superscript and the dependence in r from $s^*(r)$, and the dependence of $G_u(s)$ in s, during the proof of part B of Theorem 6. This proof is equivalent to proving that:

$$\begin{aligned} \frac{d^2(1/s)}{dr^2} &= -\frac{d\left(s'/s^2\right)}{dr^2} \\ &= \frac{2s\left(s'\right)^2 - s''s^2}{s^4} > 0. \end{aligned}$$

Thus, since $s > 0$, we will show that:

$$2\left(s'\right)^2 - s''s > 0, \tag{A.4}$$

which requires the computation of s's second derivative. Fortunately, this can be done using the Implicit Function Theorem that holds in a rectangle around (r,s), and the chain rule. This

will also prove the second claim of part C of the Theorem:

$$\begin{aligned}
\frac{d^2s}{dr^2} &= \frac{\partial s^{'}}{\partial r}+\frac{\partial s^{'}}{\partial s}\frac{ds}{dr} \\
&= \frac{\partial}{\partial r}\frac{se^{sr}}{G_u^{'}-re^{sr}}+s^{'}\frac{\partial}{\partial s}\frac{se^{sr}}{G_u^{'}-re^{sr}} \\
&= \frac{e^{sr}\left[s^2\left(G_u^{'}-re^{sr}\right)+se^{sr}\left(1+rs\right)\right]}{\left(G_u^{'}-re^{sr}\right)^2}+s^{'}\frac{e^{sr}\left[\left(1+sr\right)\left(G_u^{'}-re^{sr}\right)-s\left(G_u^{''}-r^2e^{sr}\right)\right]}{\left(G_u^{'}-re^{sr}\right)^2} \\
&= \frac{se^{sr}\left(sG_u^{'}+e^{sr}\right)+s^{'}e^{sr}\left(G_u^{'}-re^{sr}+srG_u^{'}-sG_u^{''}\right)}{\left(G_u^{'}-re^{sr}\right)^2}.
\end{aligned}$$

Substituting $s^{''}$and $s^{'}$ into (A.4) we need to show:

$$\begin{aligned}
2\left(\frac{se^{sr}}{G_u^{'}-re^{sr}}\right)^2-s\frac{se^{sr}\left(sG_u^{'}+e^{sr}\right)+s^{'}e^{sr}\left(G_u^{'}-re^{sr}+srG_u^{'}-sG_u^{''}\right)}{\left(G_u^{'}-re^{sr}\right)^2} &> 0 \\
2se^{sr}-s\left(sG_u^{'}+e^{sr}\right)-s^{'}\left(G_u^{'}-re^{sr}+srG_u^{'}-sG_u^{''}\right) &> 0 \\
se^{sr}-s^2G_u^{'}-\frac{se^{sr}}{G_u^{'}-re^{sr}}\left(G_u^{'}-re^{sr}+srG_u^{'}-sG_u^{''}\right) &> 0 \\
se^{sr}G_u^{'}-rse^{2sr}-s^2\left(G_u^{'}\right)^2+rs^2e^{sr}G_u^{'}-se^{sr}\left(G_u^{'}-re^{sr}+srG_u^{'}-sG_u^{''}\right) &> 0 \\
-s^2\left(G_u^{'}\right)^2+s^2e^{sr}G_u^{''} &> 0,
\end{aligned}$$

where the second row holds from multiplying by $\left(G_u^{'}-re^{sr}\right)^2>0$ from (A.2), and dividing by $se^{sr}>0$. The third row holds due to cancellation of se^{sr} and substituting $s^{'}$ from 4.3. The fourth row follows from multiplying by $G_u^{'}-re^{sr}>0$. Finally, the last row results from cancellation. Therefore, after dividing by $s^2>0$ we need to show:

$$-\left(G_u^{'}\right)^2+e^{sr}G_u^{''}>0$$

recalling that in s^*, condition 1 in (A.1) holds, that is $e^{sr}=G_u$, we need:

$$G_uG_u^{''}-\left(G_u^{'}\right)^2>0, \tag{A.5}$$

However, this follows from the fact that u's semi-invariant moment generating function is strictly convex, concluding the proof of part B of Theorem 6. ■

A.5 Proof of Lemma 4

We define, and recall the following notation:

Notation 18 *s_1 start of the first subcycle of the cycle: $s_1 = \inf_{i \geq 0} \{L_i \leq 0\}$, (note $s_1 = s^1$, from Definition 1) .*
f_1 end of the first subcycle: $f_1 = \inf_{i \geq s_1} \{L_i > 0\}$.
T length of a cycle: $T = \inf_{i \geq s_1} \{L_i = d\}$, (again $T = T^1$, from 1) .
s_k for $k \geq 2$, the start of the k^{th} subcycle, if such exists, or the end of the cycle otherwise: $s_k = \inf_{i \geq f_{k-1}} \{\{L_i \leq 0\} \cup \{L_i = d\}\}$.
f_k for $k \geq 2$, the end of the k^{th} subcycle, if such exists, or the end of the cycle otherwise: $f_k = \inf_{i \geq s_k} \{L_i \geq 0\}$.
t_k length of the k^{th} subcycle, if such exist, or zero otherwise: $t_k = f_k - s_k$.
t length of a generic actual subcycle
$E(t)$ expected length of a generic subcycle (and note $E(t_k) \neq E(t)$): $E(t) = \frac{\sum_{k=1}^{\infty} E(t_k | t_k > 0) P(t_k > 0)}{\sum_{k=1}^{\infty} P(t_k > 0)}$
n numbers of subcycles in a cycle,
$E(n)$ expected number of subcycles per cycle: $E(n) = \sum_{k=1}^{\infty} P(t_k > 0)$.
M expected number of periods with losses and usage higher than r during a cycle.

Note that the definitions for s_k and f_k are such that if the k^{th} subcycle never occurs they are equal to T, the stopping time for the end of the cycle. Thus, $\lim_{k \to \infty} s_k = \lim_{k \to \infty} f_k = T$.

Proof. of Lemma 4:

$$M = E\left[\sum_{k=1}^{\infty} \left(\sum_{s_k \leq i < f_k} I\{u_i > r\} \right)\right] \tag{A.6}$$

where $I\{\cdot\}$ is the indicator function with value one if the event happens and zero otherwise, and we use the convention $\sum_{T \leq i < T} I\{\cdot\} = 0$.

We exchange the order of the expectation and the sum to get:

$$M = \sum_{k=1}^{\infty} E\left(\sum_{i=s_k}^{f_k-1} I\left\{u_i > r\right\}\right). \tag{A.7}$$

For any subcycle we can write, and using the convention$\sum_{T<i\leq T} I\left\{\cdot\right\} = 0$:

$$\begin{aligned}\sum_{i=s_k}^{f_k-1} I\left\{u_i > r\right\} &= I\left\{u_{s_k} > r\right\} + \sum_{i=s_k+1}^{f_k} I\left\{u_i > r\right\} - I\left\{u_{f_k} > r\right\} \\ &= I\left\{s_k \neq T\right\} + \sum_{i=s_k+1}^{f_k} I\left\{u_i > r\right\} \end{aligned} \tag{A.8}$$

Since $I\left\{u_{s_k} > r\right\} = 1$ whenever a downcross happens (i.e., $s_k \neq T$) and $I\left\{u_{f_k} > r\right\} = 0$ for any k (since at f_k the L process always increases).

In order to express the summation above we use (3.16) Wald's equality (see, for example Chapter 3 of (Gallager, 1996)), which is reminded here. Wald's equality states that for a stopping time N and a function $g\left(x_i\right)$ of a the iid random variable x_i, such that $E\left[g\left(x_i\right)\right] = G$:

$$E\left[\sum_{i=1}^{N} g\left(x_i\right)\right] = GE\left(N\right).$$

We can use Wald's equality in our case when we consider s_k and f_k as stopping times and $I\left\{u_i > r\right\}$ as the function $g\left(x_i\right)$, with expectation $\overline{F_u}\left(r\right)$.

We use the strong Markov property for random walks (that is, that the random walk starts afresh at each stopping time) to write Wald's equality for all the stopping times defined, starting from s_1, the beginning of the first subcycle:[4]

$$\begin{aligned} E\left[\sum_{i=s_1}^{f_k} I\left\{u_i > r\right\}\right] - E\left[\sum_{i=s_1}^{s_k} I\left\{u_i > r\right\}\right] &= \overline{F_u}\left(r\right)\left[E\left(f_k\right) - E\left(s_k\right)\right] \\ &= \overline{F_u}\left(r\right) E\left(t_k\right). \end{aligned}$$

[4] We need to use Wald's equality starting at s_1, since s_1 is a defective stopping time (with positive probability to equal infinity) for the non-regulated random walk. However, at each s_k the random walk starts a fresh and we can use Wald's equality, counting from s_1 and until the required stopping times (s_k, or f_k).

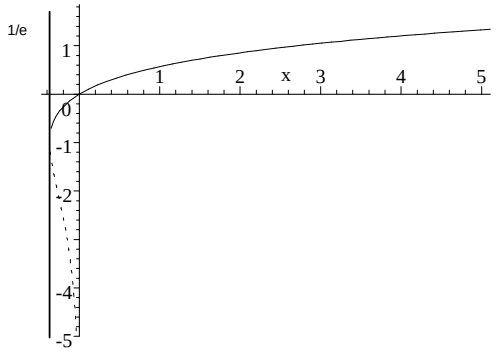


Figure A-1: The real branches of the *LambertW* (x) function. The continuous line is of the principal (zero) branch, and the dotted one is the -1th branch. Both branches meet for $x = -1/e$, where their value is -1.

Substituting this into (A.7) and (A.8) we can write:

$$\begin{aligned} M &= \sum_{k=1}^{\infty} \left(\left[I\left\{ s_k \neq T \right\} + \overline{F_u}\left(r \right) E\left(t_k \right) \right] \right) \\ &= E\left(n \right) + \overline{F_u}\left(r \right) \sum_{k=1}^{\infty} E\left(t_k \right), \end{aligned}$$

concluding the proof. ■

A.6 The LambertW (x) Function

The LambertW(x) function is the solution of

$$\text{LambertW}(x) e^{\text{LambertW}(x)} = x \tag{A.9}$$

It has two branches with real solutions, and other branches with imaginary parts. Figure A-1, shows the real branches of the *LambertW* (x) function. The continuous line is of the principal (zero) branch, and the dotted one is the -1th branch. Both branches meet for $x = -1/e$, where their value is -1. Additional names for this function are the ProductLog, Omega, and W

function. For a discussion of the $LambertW\,(x)$ function, and its usages see (Corless et al., 1996) and (Corless et al., 1997).

A.7 Proofs of Propositions 8, 9, and 10

Proof. of Proposition 8:

When buyer's demand is normally distributed, there is a closed-form solution to s^*:

$$s^* = \frac{2\,(r-\mu)}{\sigma^2}$$

Now, we can substitute this into (4.11):

$$Z\,(r) = Rr + \frac{RC\sigma^2}{2\,(r-\mu)},$$

and equating Z first derivative to zero:

$$0 = R - \frac{RC\sigma^2}{2\,(r-\mu)^2}.$$

Thus, $r^* - \mu$ satisfy:

$$(r-\mu) = \frac{\pm\sqrt{2C\sigma^2}}{2},$$

recalling that in order for $s^* > 0$ it is required that $r > \mu$ the only feasible root is given by:

$$r_B^* = \mu + \frac{\sqrt{2C\sigma^2}}{2},$$

as claimed in (4.16). Clearly, this solution also satisfies the second-order conditions, and the optimal d^* is given in (4.17), which follows from (4.10). ■

Proof. of Proposition 9:

Using (4.19) the buyer's problem, when the service constraint is active (and ignoring the

non-negativity constraint) can be replaced with:

$$\min_{d,r} Dd + Rr$$

$$s.t$$

$$\exp - \left[\frac{2\left(r-\mu\right)}{\sigma^2} d + 0.583 \left(2\frac{r-\mu}{\sigma} \right) \right] = 1 - \alpha.$$

Solving the constraint for d we get:

$$d = -\frac{\sigma^2 \ln\left(1-\alpha\right)}{2\left(r-\mu\right)} - 0.583\sigma$$

So the objective function can be written as:

$$Z\left(r\right) = Rr + \frac{RC\sigma^2}{2\left(r-\mu\right)} - 0.583\sigma,$$

which has exactly the same optimal r as in the proof of Proposition 8 and is given in (4.16). The optimal d^* is easily computed and so is the optimal cost. ■

Proof. of Proposition 10:

Based on (4.21), d is given by:

$$d = -\frac{1}{s} \ln \left(\frac{\lambda\left(1-\alpha\right)}{\lambda - s} \right) \tag{A.10}$$

Thus, the first solution is optimal if d given above is nonpositive. This happens if the ln operation results in a nonnegative number, i.e.:

$$\begin{aligned} \frac{\lambda(1-\alpha)}{\lambda - s} &\geq 1 \\ \lambda - \lambda\alpha &\geq \lambda - s \\ \frac{s}{\lambda} &\geq \alpha \end{aligned}$$

In equation (A.15) we will show that:

$$s^*(r) = \frac{\lambda\left(LambertW\left(\beta\right)+1\right)}{LambertW\left(\beta\right)},$$

Proving that the first solution is optimal when (4.22) holds. Otherwise, we substitute (A.10) into the cost function, transforming the buyer's problem to finding the minimum of:

$$Z(r) = Rr - \frac{D}{s}\ln\left(\frac{\lambda}{\lambda - s}(1-\alpha)\right) = Rr - Dr - \frac{D}{s}\ln(1-\alpha), \tag{A.11}$$

where the second equality follows from the definition of s^*, as the root of:[5]

$$\frac{\lambda}{\lambda - s} = e^{sr} \tag{A.12}$$

Now, equating Z's first derivative to zero:

$$0 = R - D\left(1 - \frac{\ln(1-\alpha)s'}{s^2}\right) \tag{A.13}$$

Applying (4.3), and using (A.12) to compute s' for the exponential case, we get:

$$\begin{aligned} G_u'(s) &= \frac{\lambda}{(\lambda - s)^2} = \frac{e^{sr}}{\lambda - s} \\ s' &= \frac{se^{sr}}{e^{sr}\left(1/(\lambda - s) - r\right)} = \frac{s(\lambda - s)}{1 - r(\lambda - s)} \end{aligned}$$

Substituting the expression for s' into the FOC in (A.13), we get:

$$\begin{aligned} 0 &= R - D\left(1 - \frac{(\lambda - s)\ln(1-\alpha)}{s(1 - r(\lambda - s))}\right) \\ 0 &= K + \frac{(\lambda - s)\ln(1-\alpha)}{s(1 - r(\lambda - s))} \end{aligned}$$

where $K = R/D - 1$ as given in (4.23), and note that $R > D$, thus $K > 0$. From the equation above we can find r as a function of s:

$$\begin{aligned} K(1 - r(\lambda - s)) &= -\frac{(\lambda - s)\ln(1-\alpha)}{s} \\ -r(\lambda - s) &= -\frac{(\lambda - s)\ln(1-\alpha) + Ks}{Ks} \\ r &= \frac{(\lambda - s)\ln(1-\alpha) + Ks}{Ks(\lambda - s)} \end{aligned} \tag{A.14}$$

[5] Recall that s^* is the positive solution to $G_x(s) = 1$.

Now, we can substitute r from above into (A.12) to get:

$$\frac{\lambda}{\lambda - s} = e^{s\frac{\ln(1-\alpha)(\lambda-s)+Ks}{Ks(\lambda-s)}} = e^{\frac{\ln(1-\alpha)(\lambda-s)+Ks}{K(\lambda-s)}} = (1-\alpha)^{1/K} e^{s/(\lambda-s)},$$

which its solution can be easily verified to be:

$$s^*(r) = \frac{\lambda\left(LambertW\left(\beta\right)+1\right)}{LambertW\left(\beta\right)}, \tag{A.15}$$

where β is given in (4.24).

In order to find the right branch of the $LambertW$ function, to maintain s increasing with r, we note that for s to increase in β the $LambertW$ should decrease with β. Moreover, since both r and β decrease (increase) with k (α), the s we need is on the decreasing (with β) branch of the $LambertW$ function. This brunch is the -1^{th} branch of the $LambertW$, see page 17 of (Corless et al., 1996).

Clearly, this solution also satisfies the second-order conditions.[6] Thus, plugging the resulting s into (A.14), the resulting r into (A.10), and both r and d into (A.11), after some algebra, concludes the proof of the proposition. ■

A.8 Computation for the Token Bucket

A.8.1 Moments of the Effective Demand

We can express the different moments of the effective demand using its CDF. Here we will do that for the first two moments. The expectation and second moments of the effective demand are given by:

$$E\left(u^e\right) = \int_0^{d+r} u^e \left[f_U\left(u^e\right)\left[1-F_{\tilde{L}}\left(u^e-r\right)\right] + \left[1-F_U\left(u^e\right)\right] f_{\tilde{L}}\left(u^e-r\right)\right] du^e \tag{A.16}$$

$$E\left(\left(u^e\right)^2\right) = \int_0^{d+r} \left(u^e\right)^2 \left[f_U\left(u^e\right)\left[1-F_{\tilde{L}}\left(u^e-r\right)\right] + \left[1-F_U\left(u^e\right)\right] f_{\tilde{L}}\left(u^e-r\right)\right] du^e \tag{A.17}$$

[6] This can be easily verified by taking the second derivative (regarding to r) from (A.13), which results in the same condition as in (A.4) in the proof of Theorem 7, (in section 4.3).

Thus, the first moment can be calculated from:

$$\begin{aligned} E\left(u^e\right) &= E\left(u; u \in [0, d+r]\right) + r + E\left(\widetilde{l}\right) \\ &\quad - \int_r^{d+r} u^e \left[f_U\left(u^e\right) F_{\widetilde{L}}\left(u^e - r\right) + F_U\left(u^e\right) f_{\widetilde{L}}\left(u^e - r\right)\right] du^e, \end{aligned}$$

and the second one from:

$$\begin{aligned} E\left(\left(u^e\right)^2\right) &= E\left(u^2; u \in [0, d+r]\right) + E\left(\left(\widetilde{l}\right)^2\right) \\ &\quad - \int_r^{d+r} \left(u^e\right)^2 \left[f_U\left(u^e\right) F_{\widetilde{L}}\left(u^e - r\right) + F_U\left(u^e\right) f_{\widetilde{L}}\left(u^e - r\right)\right] du^e, \end{aligned}$$

where we note $E\left(x; x \in [a, b]\right)$, for the expectation of the random variable x taken on the range $x \in [a, b]$. Note that the lower limit of the integrals above can be set to r since $f_{\widetilde{L}}\left(u^e - r\right)$ and $F_{\widetilde{L}}\left(u^e - r\right)$ are zero for $u \leq r$.

A.8.2 Brownian Approximation

To simplify the notation, we do not use the tilde and the superscript e for the two-sided regulated bucket level and the effective demand, respectively.

Normal Demand

When demand is normally distributed, we assume that the probability of demand lower than 0 is negligible, and we can combine the densities of the normal distribution and the approximation

to the bucket level density to write, in the range $u^e \in [r, d+r]$:

$$\begin{aligned}
f_{U^e}(u) &\approx \frac{1}{(1-e^{\theta d})\,\sigma\sqrt{2\pi}}\left[\exp\left(-\frac{(u^e-\mu)^2}{2\sigma^2}\right)\left[\exp\left(\theta\left(u^e-r\right)\right)-\exp\left(\theta d\right)\right]\right] \\
&\quad -\frac{\theta e^{\theta(u^e-r)}}{(1-e^{\theta d})\,\sigma\sqrt{2\pi}}\int_{u^e}^{\infty} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\,du \\
&= \frac{1/\left(\sigma\sqrt{2\pi}\right)}{\left(1-\exp\left(2d\left(\mu-r\right)/\sigma^2\right)\right)}\exp\left(-\frac{(u^e-\mu)^2}{2\sigma^2}\right)\exp\left(\frac{2\left(\mu-r\right)\left(u^e-r\right)}{\sigma^2}\right) \\
&\quad -\frac{1/\left(\sigma\sqrt{2\pi}\right)}{\left(1-\exp\left(2d\left(\mu-r\right)/\sigma^2\right)\right)}\exp\left(-\frac{(u^e-\mu)^2}{2\sigma^2}\right)\exp\left(\frac{2d\left(\mu-r\right)}{\sigma^2}\right) \\
&\quad \frac{-2\left(\mu-r\right)\exp\left(2\left(\mu-r\right)\left(u^e-r\right)/\sigma^2\right)}{\left(1-\exp\left(2d\left(\mu-r\right)/\sigma^2\right)\right)\sigma^3\sqrt{2\pi}}\int_{u^e}^{\infty}\exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right)du \\
&= \frac{1}{\left(1-\exp\left(2d\left(\mu-r\right)/\sigma^2\right)\right)\sigma\sqrt{2\pi}}\exp\left(\frac{4\left(\mu-r\right)\left(u^e-r\right)-(u^e-\mu)^2}{2\sigma^2}\right) \\
&\quad -\frac{1}{\left(1-\exp\left(2d\left(\mu-r\right)/\sigma^2\right)\right)\sigma\sqrt{2\pi}}\exp\left(\frac{4d\left(\mu-r\right)-(u^e-\mu)^2}{2\sigma^2}\right) \\
&\quad \frac{-2\left(\mu-r\right)\exp\left(2\left(\mu-r\right)\left(u^e-r\right)/\sigma^2\right)}{\left(1-\exp\left(2d\left(\mu-r\right)/\sigma^2\right)\right)\sigma^3\sqrt{2\pi}}\int_{u^e}^{\infty}\exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right)du
\end{aligned}$$

The CDF can be approximated as:

$$F_{U^e}(u^e) \approx \begin{cases} 0 & \text{for } u^e < 0 \\ F_U(u^e) & \text{for } 0 \le u^e < r \\ \frac{1-e^{\theta(u^e-r)}}{1-e^{\theta d}} + \frac{e^{\theta(u^e-r)}-e^{\theta d}}{1-e^{\theta d}}\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{u^e} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx & \text{for } [r \le u^e < d+r] \\ 1 & \text{for } d+r \le u^e \end{cases}$$

So, we can approximate the first moment:

$$\begin{aligned}
E(u^e) &= \int_0^{d+r} u f_{u^e}(u)\,du \\
&\approx \int_0^r u f_u(u)\,du + \int_r^{d+r} u\left[f_U(u)\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}} + \left[1-F_U(u)\right]\frac{-\theta e^{\theta(u-r)}}{(1-e^{\theta d})}\right]du.
\end{aligned}$$

Or we can use the alternative computation (since the effective demand is nonnegative):

$$\begin{aligned}
E\left(u^e\right) &= \int_0^{d+r} 1 - F_{U^e}\left(u^e\right) du^e \\
&\approx \int_0^{r} 1 - F_U\left(u\right) du + \\
&\quad \frac{1}{1-e^{\theta d}} \int_r^{d+r} \left\{ e^{\theta(u^e - r)} - e^{\theta d} - \frac{\left(e^{\theta(u^e - r)} - e^{\theta d}\right)}{\sigma\sqrt{2\pi}} \int_{-\infty}^{u^e} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \right\} du^e \\
&= \int_0^{r} 1 - F_U\left(u\right) du + \\
&\quad \frac{1}{1-e^{\theta d}} \left\{ e^{-\theta r} \left(\frac{e^{\theta(d+r)} - e^{\theta r}}{\theta} \right) - d e^{\theta d} - \int_r^{d+r} \frac{\left(e^{\theta(u^e - r)} - e^{\theta d}\right)}{\sigma\sqrt{2\pi}} \int_{-\infty}^{u^e} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx du^e \right\}
\end{aligned}$$

Which is equal to:

$$
\begin{aligned}
= & \int_0^r 1 - F_U(u)\, du + \\
& \frac{1}{1-e^{\theta d}} \left\{ (1 - F_U(0)) \left[\left(\frac{e^{\theta d} - 1}{\theta} \right) - d e^{\theta d} \right] - \int_r^{d+r} \frac{\left(e^{\theta (u^e - r)} - e^{\theta d} \right)}{\sigma \sqrt{2\pi}} \int_0^{u^e} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx du^e \right\} \\
= & \int_0^r 1 - F_U(u)\, du + \\
& \frac{1}{1-e^{\theta d}} \left\{ (1 - F_U(0)) \left[\left(\frac{e^{\theta d} - 1}{\theta} \right) - d e^{\theta d} \right] - \int_0^r \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}} \int_r^{d+r} \left(e^{\theta (u^e - r)} - e^{\theta d} \right) du^e dx \right\} \\
& - \frac{1}{1-e^{\theta d}} \int_r^{d+r} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}} \int_x^{d+r} \left(e^{\theta (u^e - r)} - e^{\theta d} \right) du^e dx \\
= & \int_0^r 1 - F_U(u)\, du + \frac{(1 - F_U(0))}{1-e^{\theta d}} \left[\left(\frac{e^{\theta d} - 1}{\theta} \right) - d e^{\theta d} \right] \\
& - \frac{1}{1-e^{\theta d}} \left\{ \int_0^r \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}} \left[\left(\frac{e^{\theta d} - 1}{\theta} \right) - d e^{\theta d} \right] dx \right\} \\
& - \frac{1}{1-e^{\theta d}} \int_r^{d+r} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}} \left(\left(\frac{e^{\theta d} - e^{\theta (x-r)}}{\theta} \right) - (d + r - x) e^{\theta d} \right) dx \\
= & \int_0^r 1 - F_U(u)\, du + \frac{(1 - F_U(r))}{1-e^{\theta d}} \left[\left(\frac{e^{\theta d} - 1}{\theta} \right) - d e^{\theta d} \right] \\
& - \frac{e^{\theta d} (1/\theta - d - r) (F_U(d + r) - F_U(r))}{1 - e^{\theta d}} \\
& - \frac{1}{1-e^{\theta d}} \int_r^{d+r} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}} \left(\left(\frac{-e^{\theta (x-r)}}{\theta} \right) + x e^{\theta d} \right) dx
\end{aligned}
$$

$$
\begin{aligned}
= & \int_0^r 1 - F_U(u)\, du + \frac{(1 - F_U(r))}{1-e^{\theta d}} \left[\left(\frac{e^{\theta d} - 1}{\theta} \right) - d e^{\theta d} \right] \\
& - \frac{e^{\theta d} (1/\theta - d - r) (F_U(d + r) - F_U(r))}{1 - e^{\theta d}} \\
& - \frac{1}{1-e^{\theta d}} \int_r^{d+r} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}} \left(\left(\frac{-e^{\theta (x-r)}}{\theta} \right) + x e^{\theta d} \right) dx
\end{aligned}
$$

Let us look at the last integral, without the division in $\left(1-e^{\theta d}\right)$:

$$\begin{aligned}
\int_r^{d+r} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\left[\frac{e^{\theta(x-r)}}{\theta}-xe^{\theta d}\right]dx &= \int_0^{d+r} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\left[\frac{e^{\theta(x-r)}}{\theta}-xe^{\theta d}\right]dx \\
&= \frac{e^{-\theta r}e^{\left(\theta^2\sigma^2/2+\mu\theta\right)}}{\theta}\int_0^{d+r}\frac{\exp\left(-\frac{\left(x-\mu-\theta\sigma^2\right)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}dx \\
&\quad -e^{\theta d}E_u\left(x;x\in[0,d+r]\right) \\
&= \frac{e^{-\theta r}e^{\left(\theta^2\sigma^2/2+\mu\theta\right)}}{\theta}\Phi\left(\frac{d+r-\mu-\theta\sigma^2}{\sigma}\right) \\
&\quad -\frac{e^{-\theta r}e^{\left(\theta^2\sigma^2/2+\mu\theta\right)}}{\theta}\Phi\left(-\frac{\mu+\theta\sigma^2}{\sigma}\right) \\
&\quad -e^{\theta d}E_u\left(x;x\in[0,d+r]\right),
\end{aligned}$$

where we note $E_u\left(x;x\in[a,b]\right)$ as the expectation of the demand taken on the range $[a,b]$, and $\Phi\left(x\right)$ is the CDF of a standard normally distribution evaluated at the point x. To summarize the above computation, the first moment of the effective demand when demand is normal distributed is given by:

$$\begin{aligned}
E\left(u^e\right) &= \int_0^r 1-F_U\left(u\right)du+\frac{\left(1-F_U\left(r\right)\right)}{1-e^{\theta d}}\left[\left(\frac{e^{\theta d}-1}{\theta}\right)-de^{\theta d}\right] \\
&\quad -\frac{e^{\theta d}\left(1/\theta-d-r\right)\left(F_U\left(d+r\right)-F_U\left(r\right)\right)}{1-e^{\theta d}}-e^{\theta d}E_u\left(x;x\in[0,d+r]\right) \\
&\quad +\frac{e^{-\theta r}e^{\left(\theta^2\sigma^2/2+\mu\theta\right)}}{\theta}\left[\Phi\left(\frac{d+r-\mu-\theta\sigma^2}{\sigma}\right)-\Phi\left(\frac{-\mu-\theta\sigma^2}{\sigma}\right)\right] \\
&= \int_0^r 1-\Phi\left(\frac{u-\mu}{\sigma}\right)du+\frac{\left(1-\Phi\left(\frac{r-\mu}{\sigma}\right)\right)}{1-e^{\theta d}}\left[\left(\frac{e^{\theta d}-1}{\theta}\right)-de^{\theta d}\right] \\
&\quad -\frac{e^{\theta d}\left(1/\theta-d-r\right)\left(\Phi\left(\frac{d+r-\mu}{\sigma}\right)-\Phi\left(\frac{r-\mu}{\sigma}\right)\right)}{1-e^{\theta d}}-e^{\theta d}E_u\left(x;x\in[0,d+r]\right) \\
&\quad +\frac{e^{-\theta r}e^{\left(\theta^2\sigma^2/2+\mu\theta\right)}}{\theta}\left[\Phi\left(\frac{d+r-\mu-\theta\sigma^2}{\sigma}\right)-\Phi\left(\frac{-\mu-\theta\sigma^2}{\sigma}\right)\right].
\end{aligned}$$

Exponential Demand

The density of the effective demand can be written as:

$$f_{U^e}(u) \approx \begin{cases} 0 & \text{for } u < 0 \\ \lambda e^{-\lambda u} & \text{for } 0 \leq u < r \\ \lambda e^{-\lambda u}\left[\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}}\right] + \left[e^{-\lambda u}\right]\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)} & \text{for } r \leq u \leq r+d \\ 0 & \text{for } d+r < u \end{cases}$$

and the expected effective demand as:

$$\begin{aligned} E(u^e) &\approx \int_0^r u\lambda e^{-\lambda u}du + \int_r^{r+d} u\left[\lambda e^{-\lambda u}\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}} + \left[e^{-\lambda u}\right]\frac{-\theta e^{\theta(u-r)}}{(1-e^{\theta d})}\right]du \\ &= \Bigg[_0 - \frac{e^{-\lambda u}}{\lambda}(\lambda u+1) - \frac{e^{-\theta r}}{1-e^{\theta d}}\Bigg]_r^{r+d}\left[\frac{1}{\lambda-\theta}e^{-(\lambda-\theta)u}((\lambda-\theta)u+1)\right] \\ &\quad + \frac{e^{\theta d}}{1-e^{\theta d}}\Bigg]_r^{r+d}\frac{e^{-\lambda u}}{\lambda}(\lambda u+1) \\ &= \frac{1-e^{-\lambda r}(\lambda r+1)}{\lambda} - \frac{e^{-\theta r}}{1-e^{\theta d}}\frac{1}{\lambda-\theta}e^{-(\lambda-\theta)(r+d)}((\lambda-\theta)(r+d)+1) \\ &\quad + \frac{e^{-\theta r}}{1-e^{\theta d}}\frac{1}{\lambda-\theta}e^{-(\lambda-\theta)r}((\lambda-\theta)r+1) \\ &\quad + \frac{e^{\theta d}}{\lambda(1-e^{\theta d})}\left[e^{-\lambda(r+d)}(\lambda(r+d)+1) - e^{-\lambda r}(\lambda r+1)\right] \\ &= \frac{1-e^{-\lambda r}(\lambda r+1)}{\lambda} - \frac{e^{-\lambda r}}{1-e^{\theta d}}\frac{1}{\lambda-\theta}e^{-(\lambda-\theta)d}(\lambda-\theta)d \\ &\quad - \frac{e^{-\lambda r}}{1-e^{\theta d}}\frac{1}{\lambda-\theta}\left(1-e^{-(\lambda-\theta)d}\right)((\lambda-\theta)r+1) \\ &\quad + \frac{e^{\theta d-\lambda r}}{\lambda(1-e^{\theta d})}\left[\left(e^{-\lambda d}-1\right)(\lambda r+1) - e^{-\lambda d}\lambda d\right] \end{aligned}$$

and the second moment:

$$
\begin{aligned}
E\left((u^e)^2\right) &\approx \int_0^r u^2\lambda e^{-\lambda u}du + \int_r^{r+d} u^2\left[\lambda e^{-\lambda u}\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}} + \left[e^{-\lambda u}\right]\frac{-\theta e^{\theta(u-r)}}{(1-e^{\theta d})}\right]du \\
&= \int_0^r u^2\lambda e^{-\lambda u}du + \int_r^{r+d} u^2\left[\frac{e^{-\theta r}\left(\lambda-\theta\right)e^{-(\lambda-\theta)u}}{1-e^{\theta d}} - \frac{e^{\theta d}}{1-e^{\theta d}}\lambda e^{-\lambda u}\right]du \\
&= \Bigg[_0^r - \frac{e^{-\lambda u}}{\lambda^2}\left(\lambda^2u^2+2\lambda u+2\right) + \frac{e^{\theta d}}{1-e^{\theta d}}\Bigg]_r^{r+d}\frac{e^{-\lambda u}}{\lambda^2}\left(\lambda^2u^2+2\lambda u+2\right) \\
&\quad -\frac{e^{-\theta r}}{1-e^{\theta d}}\Bigg[_r^{r+d}\left[\frac{e^{-(\lambda-\theta)u}}{(\lambda-\theta)^2}\left(\left(\left(\lambda-\theta\right)u\right)^2+2\left(\lambda-\theta\right)u+2\right)\right] \\
&= \frac{2-e^{-\lambda r}\left[(\lambda r+1)^2+1\right]}{\lambda^2} \\
&\quad +\frac{e^{\theta d-\lambda r}}{\lambda^2\left(1-e^{\theta d}\right)}\left[e^{-\lambda d}\left[\left(\lambda\left(r+d\right)+1\right)^2+1\right]-\left[(\lambda r+1)^2+1\right]\right] \\
&\quad -\frac{e^{-\theta r}}{1-e^{\theta d}}\frac{e^{-(\lambda-\theta)r}}{(\lambda-\theta)^2}\left[e^{-(\lambda-\theta)d}\left[\left(\left(\lambda-\theta\right)\left(r+d\right)+1\right)^2+1\right]-\left[\left(\left(\lambda-\theta\right)r+1\right)^2+1\right]\right] \\
&= \frac{2-e^{-\lambda r}\left[(\lambda r+1)^2+1\right]}{\lambda^2} \\
&\quad +\frac{e^{\theta d-\lambda r}}{\lambda^2\left(1-e^{\theta d}\right)}\left[e^{-\lambda d}\left[\left(\lambda\left(r+d\right)+1\right)^2+1\right]-\left[(\lambda r+1)^2+1\right]\right] \\
&\quad -\frac{1}{1-e^{\theta d}}\frac{e^{-\lambda r}}{(\lambda-\theta)^2}\left[e^{-(\lambda-\theta)d}\left[\left(\left(\lambda-\theta\right)\left(r+d\right)+1\right)^2+1\right]-\left[\left(\left(\lambda-\theta\right)r+1\right)^2+1\right]\right]
\end{aligned}
$$

A.8.3 The Enhancement

To simplify the notation, we omit the superscript e for the effective demand. We only provide here the results for the exponential case, since there is no closed form for the result of the normal demand case. However, the required integration can be easily done for the normal demand case as well.

Exponential Demand

The density of the effective demand can be written as:

$$f_{U^e}(u) \approx \begin{cases} 0 & \text{for } u < 0 \\ \lambda e^{-\lambda u} & \text{for } 0 \leq u < r \\ \lambda e^{-\lambda u}\left[\frac{e^{\theta(u-r)}-e^{\theta d}}{1-e^{\theta d}}\right] + \left[e^{-\lambda u}\right]\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)} & \text{for } r \leq u \leq r+d \\ 0 & \text{for } d+r < u \end{cases}$$

recall $P\left(\widetilde{L}=0\right) = \frac{1-\lambda r}{\left(1-e^{-2\lambda d(1-\lambda r)}\right)}$

$$f_{U^e}(u) \approx \begin{cases} 0 & \text{for } u < 0 \\ f_U(u) & \text{for } 0 \leq u < r \\ f_U(r)\left[1-P\left(\widetilde{L}=0\right)\right] + \delta P\left(\widetilde{L}=0\right) & \\ \left(\frac{1/\lambda - r}{1/\lambda\left(1-e^{-\theta d}\right)} + \frac{1-e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)\left(1-\frac{1/\lambda-r}{1/\lambda\left(1-e^{-\theta d}\right)} - \frac{r-1/\lambda}{e^{\theta d}-1}\frac{\lambda r+e^{-\lambda r}-1}{(1-e^{-\lambda r})\lambda}\right)}\right)\lambda e^{-\lambda u} + \left(1-e^{-\lambda u}\right) & \\ & \text{for } r \leq u \leq r+d \\ 1 & \text{for } d+r < u, \end{cases}$$

where $\theta = 2\left(r - 1/\lambda\right) / \left(1/\lambda\right)^2 = 2\lambda\left(\lambda r - 1\right) > 0$, and the expected effective demand:

$$
\begin{aligned}
E^E\left(u^e\right) \approx & \int_0^{r+d} u\lambda e^{-\lambda u} du - P\left(\widetilde{L}=0\right)\int_r^{r+d} u\lambda e^{-\lambda u} du \\
& -c\int_r^{r+d} u\left[\lambda e^{-\lambda u}\frac{1-e^{\theta(u-r)}}{1-e^{\theta d}} - e^{-\lambda u}\frac{-\theta e^{\theta(u-r)}}{\left(1-e^{\theta d}\right)}\right] du \\
& +P\left(\widetilde{L}=0\right)\left[1-F_U\left(r\right)\right] r + P\left(\widetilde{L}=d\right)\left[1-F_U\left(r+d\right)\right]\left(r+d\right) \\
= & \left[\int_0^{r+d}\left[-\frac{e^{-\lambda u}}{\lambda}\left(\lambda u+1\right)\right] - P\left(\widetilde{L}=0\right)\right]_r^{r+d}\left[-\frac{e^{-\lambda u}}{\lambda}\left(\lambda u+1\right)\right] \\
& -\frac{c}{1-e^{\theta d}}\left[\int_r^{r+d}\left[-\frac{e^{-\lambda u}}{\lambda}\left(\lambda u+1\right)\right] + \frac{ce^{-\theta r}}{1-e^{\theta d}}\right]_r^{r+d}\left[-\frac{1}{\lambda-\theta}e^{-(\lambda-\theta)u}\left(\left(\lambda-\theta\right)u+1\right)\right] \\
& +P\left(\widetilde{L}=0\right)e^{-\lambda r} r + P\left(\widetilde{L}=d\right)e^{-\lambda(r+d)}\left(r+d\right) \\
= & \frac{1-e^{-\lambda(r+d)}\left(\lambda\left(r+d\right)+1\right)}{\lambda} - \frac{P\left(\widetilde{L}=0\right)e^{-\lambda r}}{\lambda}\left[\left(\lambda r+1\right) - e^{-\lambda d}\left(\lambda\left(r+d\right)+1\right)\right] \\
& -\frac{ce^{-\lambda r}}{\left(1-e^{\theta d}\right)\lambda}\left[\left(\lambda r+1\right) - e^{-\lambda d}\left(\lambda\left(r+d\right)+1\right)\right] \\
& +\frac{ce^{-\lambda r}}{1-e^{\theta d}}\frac{1}{\lambda-\theta}\left[\left(\left(\lambda-\theta\right)r+1\right) - e^{-(\lambda-\theta)d}\left(\left(\lambda-\theta\right)\left(r+d\right)+1\right)\right] \\
& +P\left(\widetilde{L}=0\right)e^{-\lambda r} r + P\left(\widetilde{L}=d\right)e^{-\lambda(r+d)}\left(r+d\right)
\end{aligned}
$$

and the second moment:

$$
\begin{aligned}
E\left((u^e)^2\right) &\approx \int_0^{r+d} u^2\lambda e^{-\lambda u}du - P\left(\widetilde{L}=0\right)\int_r^{r+d} u^2\lambda e^{-\lambda u}du \\
&\quad -c\int_r^{r+d} u^2\left[\lambda e^{-\lambda u}\frac{1-e^{\theta(u-r)}}{1-e^{\theta d}} - e^{-\lambda u}\frac{-\theta e^{\theta(u-r)}}{1-e^{\theta d}}\right]du \\
&\quad +P\left(\widetilde{L}=0\right)\left[1-F_U\left(r\right)\right]r^2 + P\left(\widetilde{L}=d\right)\left[1-F_U\left(r+d\right)\right](r+d)^2 \\
&= \int_0^{r} u^2\lambda e^{-\lambda u}du - \left[P\left(\widetilde{L}=0\right)-\frac{c}{1-e^{\theta d}}\right]\int_r^{r+d} u^2\lambda e^{-\lambda u}du \\
&\quad +\frac{ce^{-\theta r}}{1-e^{\theta d}}\int_r^{r+d} u^2\left(\lambda-\theta\right)e^{-(\lambda-\theta)u}du \\
&\quad +P\left(\widetilde{L}=0\right)e^{-\lambda r}r^2 + P\left(\widetilde{L}=d\right)e^{-\lambda(r+d)}(r+d)^2 \\
&= \Bigg[_0^{r+d}\left[-\frac{e^{-\lambda u}}{\lambda^2}\left(\lambda^2u^2+2\lambda u+2\right)\right] \\
&\quad -\left[P\left(\widetilde{L}=0\right)-\frac{c}{1-e^{\theta d}}\right]\Bigg[_r^{r+d}\left[-\frac{e^{-\lambda u}}{\lambda^2}\left(\lambda^2u^2+2\lambda u+2\right)\right] \\
&\quad +\frac{ce^{-\theta r}}{1-e^{\theta d}}\Bigg[_r^{r+d}\left[\frac{e^{-(\lambda-\theta)u}}{\left(\lambda-\theta\right)^2}\left(\left(\left(\lambda-\theta\right)u\right)^2+2\left(\lambda-\theta\right)u+2\right)\right] \\
&\quad +P\left(\widetilde{L}=0\right)e^{-\lambda r}r^2 + P\left(\widetilde{L}=d\right)e^{-\lambda(r+d)}(r+d)^2 \\
&= \frac{1}{\lambda^2}\left[2-e^{-\lambda(r+d)}\left(\lambda^2(r+d)^2+2\lambda(r+d)+2\right)\right] \\
&\quad -\frac{\left[P\left(\widetilde{L}=0\right)\left(1-e^{\theta d}\right)-c\right]e^{-\lambda r}}{\lambda^2\left(1-e^{\theta d}\right)}\left[\left(\left(\lambda r+1\right)^2+1\right)-e^{-\lambda d}\left(\left[\lambda\left(r+d\right)+1\right]^2+1\right)\right] \\
&\quad +P\left(\widetilde{L}=0\right)e^{-\lambda r}r^2 + P\left(\widetilde{L}=d\right)e^{-\lambda(r+d)}(r+d)^2.
\end{aligned}
$$

A.9 Computations for the Token Bucket with Rate Control

A.9.1 The CDF

Integrating (5.29), we can find the CDF of the effective demand in the case of TBwRC. For the sake of brevity we use $F_{U^e}(x)$ as the CDF of the effective demand at the point x within

the equations for $F_{U^e}(u^e)$. Thus, the CDF of the effective demand is:

$$F_{U^e}(u^e) = \begin{cases} 0 & \text{for } u^e < 0 \\ \int_0^{u^e}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1-F_L(0^-)\right] & \text{for } u^e \in [0,r) \\ \int_0^{r}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(r^-)\left[1-F_L(0^-)\right] + F_L(-r^+) + \\ \left[1-F_U(r)\right]\left[F_L(0^-)-F_L(-r^+)\right] + \int_0^r\int_{-r^+}^{x^- -r} f_U(x) f_L(y)dydx + \\ f_U(r)\left[1-F_L(0^-)\right] + \left[1-F_U(r)\right]P(L=0) & \text{for } u^e = r \\ F_{U^e}(r) + F_U(u^e) - F_U(r^+) - \int_r^{u^e} f_U(x) F_L(x-r)\,dx + \\ F_L(u^e-r) - F_L(0^+) - \int_r^{u^e} F_U(x) f_L(x-r)\,dx & \text{for } u^e = (r, r+d) \\ F_{U^e}\left((r+d)^-\right) + \left[1-F_U\left((r+d)^-\right)\right]P(L=d) & \text{for } u^e \geq r+d \end{cases} \quad (A.18)$$

$$\begin{cases} 0 & \text{for } u^e < 0 \\ \int_0^{u^e}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1-F_L(0^-)\right] & \text{for } u^e \in [0,r) \\ \int_0^{r}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(r)\left[1-F_L(0^-)\right] + F_L(-r)F_U(r) & \\ \int_0^r\int_{-r^+}^{x^- -r} f_U(x) f_L(y)dydx + \left[1-F_U(r)\right]F_L(0) & \text{for } u^e = r \\ F_{U^e}(r) + F_U(u^e) - F_U(r) - \int_r^{u^e} f_U(x) F_L(x-r)\,dx + \\ F_L(u^e-r) - F_L(0) - \int_r^{u^e} F_U(x) f_L(x-r)\,dx & \text{for } u^e = (r, r+d) \\ F_{U^e}\left((r+d)^-\right) + \left[1-F_U\left((r+d)^-\right)\right]P(L=d) & \text{for } u^e \geq r+d \end{cases}$$

where the equation holds due to algebra and the fact that for CDF $F_X(y) = F_X(y^+)$ (almost surely). After rearranging terms, in the range $u^e \in [r, r+d)$:

$$F_{U^e}(u^e) = \begin{cases} 0 & \text{for } u^e < 0 \\ \int_0^{u^e}\int_0^{x-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1-F_L(0^-)\right] & \text{for } u^e \in [0,r) \\ \int_0^{r}\left[\int_0^{x-} f_U(y) f_L(y-x)\,dy + \int_{-r^+}^{x^- -r} f_U(x) f_L(y)dy\right]dx & \\ +F_U(r)\left[1-F_L(0)-F_L(0^-)+F_L(-r)\right] + F_L(0) & \text{for } u^e = r \\ F_{U^e}(r) + F_U(u^e) - F_U(r) + F_L(u^e-r) - F_L(0) & \\ -\int_r^{u^e}\left[F_U(x) f_L(x-r) + f_U(x) F_L(x-r)\right]dx & \text{for } u^e = (r, r+d) \\ F_{U^e}\left((r+d)^-\right) + \left[1-F_U\left((r+d)^-\right)\right]P(L=d) & \text{for } u^e \geq r+d \end{cases} \quad (A.19)$$

We note these integrations $I_1 = \int_0^r \left[\int_0^{x^-} f_U(y) f_L(y-x)\,dy + \int_{-r^+}^{x^- -r} f_U(x) f_L(y)dy\right]dx$, and $I_2 = \int_r^{u^e}\left[F_U(x) f_L(x-r) + f_U(x) F_L(x-r)\right]dx$ and develop them. For the first integral, it

can be seen that it covers exactly the range for which buyer's usage is between zero and r and the bucket level is between $-r$ and zero, and using the independence of the bucket level and the buyer's usage, we see that $I_1 = [F_U(r) - F_U(0)][F_L(0) - F_L(-r)]$. We also give this result algebraically, where for simplicity we assume $F_U(0) = 0$:

$$\begin{aligned}
I_1 &= \int_0^r \left[\int_0^{x^-} f_U(y) f_L(y-x)\, dy + \int_{-r^+}^{x^- - r} f_U(x) f_L(y) dy \right] \\
&= \int_0^r \int_0^{x^-} f_U(y) f_L(y-x)\, dy dx + \int_0^r f_U(x) \left[F_L\left(x^- - r\right) - F_L\left(-r^+\right) \right] dx \\
&= \int_0^r f_U(x) F_L\left(x^- - r\right) dx - F_U(r) F_L(-r) + \int_0^r \int_0^{x-} f_U(y) f_L(y-x)\, dy dx \\
&= \int_0^r f_U(x) F_L\left(x^- - r\right) dx - F_U(r) F_L(-r) + \int_0^{r^-} \int_{y^+}^r f_U(y) f_L(y-x)\, dx dy \\
&= \int_0^r f_U(x) F_L\left(x^- - r\right) dx - F_U(r) F_L(-r) + \int_0^{r^-} f_U(y) \left[F_L(0) - F_L(y-r) \right] dy \\
&= -F_U(r) F_L(-r) + F_U(r) F_L\left(0^-\right) \\
&= F_U(r) \left[F_L\left(0^-\right) - F_L(-r) \right],
\end{aligned}$$

where the second and third equalities hold from integrating the second integral, exchanging the orders of elements, and since $F_L(-r^+) = F_L(-r)$ (almost surely). The fourth equality holds from exchanging the order of integration. The fifth one holds due to integration, and note that $\int f_L(-x)\, dx = -F_L(x)$. The sixth (almost surely) and seventh follow algebraically.

For the second integral $I_2 = \int_r^{u^e} \left[F_U(x) f_L(x-r) + f_U(x) F_L(x-r) \right] dx$ we recall the formula of integration by parts (for definite integrals):

$$\int_a^b g(x)\, dh(x) = h(b) g(b) - h(a) g(a) - \int_a^b h(x)\, dg(x). \tag{A.20}$$

This can be used when there are no common points of discontinuity in the range $[a, b]$. Applying (A.20), for the first element of I_2 we see:

$$\begin{aligned}
I_2 &= \int_r^{u^e} \left[F_U(x) f_L(x-r) + f_U(x) F_L(x-r) \right] dx \\
&= F_U(u^e) F_L(u^e - r) - F_U(r) F_L(0) - \int_r^{u^e} f_U(x) F_L(x-r)\, dx + \int_r^{u^e} f_U(x) F_L(x-r)\, dx \\
&= F_U(u^e) F_L(u^e - r) - F_U(r) F_L(0)
\end{aligned}$$

Replacing the expressions for the integrals into (A.19), we get:

$$
\begin{aligned}
F_{U^e}(u^e) &= \begin{cases}
0 & \text{for } u^e < 0 \\
\int_0^{u^e}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1 - F_L(0^-)\right] & \text{for } u^e \in [0, r) \\
F_U(r)\left[F_L(0^-) - F_L(-r)\right] & \\
+F_U(r)\left[1 - F_L(0) - F_L(0^-) + F_L(-r)\right] + F_L(0) & \text{for } u^e = r \\
F_{U^e}(r) + F_U(u^e) - F_U(r) + F_L(u^e - r) - F_L(0) & \\
-F_U(u^e) F_L(u^e - r) + F_U(r) F_L(0) & \text{for } u^e = (r, r+d) \\
F_{U^e}\left((r+d)^-\right) + \left[1 - F_U(r+d)\right] P(L = d) & \text{for } u^e \geq r + d
\end{cases} \\
&= \begin{cases}
0 & \text{for } u^e < 0 \\
\int_0^{u^e}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1 - F_L(0^-)\right] & \text{for } u^e \in [0, r) \\
F_U(r)\left[1 - F_L(0)\right] + F_L(0) & \text{for } u^e = r \\
F_{U^e}(r) + F_U(u^e) - F_U(r)\left[1 - F_L(0)\right] - F_L(0) & \\
F_L(u^e - r)\left[1 - F_U(u^e)\right] & \text{for } u^e = (r, r+d) \\
F_{U^e}\left((r+d)^-\right) + \left[1 - F_U(r+d)\right] P(L = d) & \text{for } u^e \geq r + d
\end{cases}
\end{aligned}
$$

which with some additional algebra becomes:

$$
\begin{aligned}
F_{U^e}(u^e) &= \begin{cases}
0 & \text{for } u^e < 0 \\
\int_0^{u^e}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1 - F_L(0^-)\right] & \text{for } u^e \in [0, r) \\
F_U(r)\left[1 - F_L(0)\right] + F_L(0) & \text{for } u^e = r \\
F_U(u^e) + F_L(u^e - r)\left[1 - F_U(u^e)\right] & \text{for } u^e = (r, r+d) \\
F_U\left((r+d)^-\right) + F_L(d^-)\left[1 - F_U((r+d^-))\right] & \\
\left[1 - F_U(r+d)\right] P(L = d) & \text{for } u^e \geq r + d
\end{cases} \\
&= \begin{cases}
0 & \text{for } u^e < 0 \\
\int_0^{u^e}\int_0^{x^-} f_U(y) f_L(y-x)\,dydx + F_U(u^e)\left[1 - F_L(0^-)\right] & \text{for } u^e \in [0, r) \\
F_U(r)\left[1 - F_L(0)\right] + F_L(0) & \text{for } u^e = r \\
F_U(u^e) + F_L(u^e - r)\left[1 - F_U(u^e)\right] & \text{for } u^e = (r, r+d) \\
1 & \text{for } u^e \geq r + d
\end{cases}
\end{aligned}
$$

A.9.2 Computing the PDF for the Brownian Motion Approximation

The computation of the PDF for the Brownian approximation is given by:

$$
\begin{aligned}
f_{U^e}(u) \approx & \begin{cases}
\int_0^{u^e} f_U(x)\,\theta e^{-\theta(d-(x-u^e))}dx + f_U(u^e)\left[1-e^{-\theta d}\right] & \text{for } u^e \in [0,r) \\
e^{-\theta(d+r)}F_U(r)\delta + \int_0^r \int_{-r}^{x-r} f_U(x)\theta e^{-\theta(d-y)}dydx\delta & \\
+[1-F_U(r)]e^{-\theta d}\delta & \text{for } u^e = r \\
f_U(u^e)\left[1-e^{-\theta(d-(u^e-r))}\right] + [1-F_U(u^e)]\,\theta e^{-\theta(d-(u^e-r))} & \text{for } u^e = [r, r+d] \\
0 & \text{otherwise}
\end{cases} \\
= & \begin{cases}
\int_0^{u^e} f_U(x)\,\theta e^{-\theta(d-(x-u^e))}dx + f_U(u^e)\left[1-e^{-\theta d}\right] & \text{for } u^e \in [0,r) \\
e^{-\theta(d+r)}F_U(r)\delta + e^{-\theta(d+r)}\int_0^r f_U(x)\left(e^{\theta x}-1\right)dx\delta & \\
+[1-F_U(r)]e^{-\theta d}\delta & \text{for } u^e = r \\
f_U(u^e) - e^{-\theta(d-(u^e-r))}\left[f_U(u^e) + F_U(u^e)\,\theta - \theta\right] & \text{for } u^e = [r, r+d] \\
0 & \text{otherwise}
\end{cases} \\
= & \begin{cases}
\int_0^{u^e} f_U(x)\,\theta e^{-\theta(d-(x-u^e))}dx + f_U(u^e)\left[1-e^{-\theta d}\right] & \text{for } u^e \in [0,r) \\
e^{-\theta(d+r)}\int_0^r f_U(x)e^{\theta x}dx\delta + [1-F_U(r)]e^{-\theta d}\delta & \text{for } u^e = r \\
f_U(u^e) - e^{-\theta(d-(u^e-r))}\left[f_U(u^e) + F_U(u^e)\,\theta - \theta\right] & \text{for } u^e = [r, r+d] \\
0 & \text{otherwise}
\end{cases}
\end{aligned}
$$

Bibliography

Altmann, J. and Chu, K. (2001a). A proposal for a flexible service plan that is attractive to users and internet service providers. *IEEE Infocom2001, Conference on Computer Communications, Anchorage, Alaska, USA, apr., 2001.*

Altmann, J. and Chu, K. (2001b). A proposal for a flexible service plan that is attractive to users and internet service providers. IEEE Infocom2001, Conference on Computer Communications, Anchorage, Alaska, USA.

Austin, R. D. (2001). Web and it hosting facilities. *President and Fellows of Harvard College.*

Balakrishnan, H., Padmanabhan, V., Seshan, S., Stemm, M., and Katz, R. (1998). Tcp behavior of a busy server: Analysis and improvements. *Proc IEEE INFOCOM n'98, Mar. 1998.*

Barford, P. (2001). http://cs-www.bu.edu/pub/barford/ss_lrd.html.

Berger, A. W. (1991). Performance analysis of a rate-control throttle when tokens and jobs queue. *IEEE journal on Selected Areas in Communications*, 9(2).

Berger, A. W. and Whitt, W. (1992). The brownian approximation for rate-control throttle and the g/g/1/c queue. *Discrete Event Dynamic System: Theory and Applications*, 2.

Berger, A. W. and Whitt, W. (1994). The pros and cones of a job buffer in a token-bank rate-control throttle. *IEEE Transactions on Communications*, 42(2/3/4):857–861.

Chang, Y., Guo, X., Kimbrel, T., and King, A. (2001). Optimal allocation policies for web hosting. 11th INFORMS Applied Probability Society Conference, New York City.

Chung, J. and Sirbu, M. A. (2001). Distributed network storage with quality-of-service guarantees. http://www.ini.cmu.edu/ sirbu/pubs/99251/Chuang.htm, Carangie Mellon University.

Corless, R. M., Gonett, G. H., Harc, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the lambertw function. *Advances in Computational Mathematics*, 5:329–359.

Corless, R. M., Jeffrey, D. J., and Knuth, D. E. (1997). A sequence of series for the lambert w function. In Kuschlin, W. W., editor, *Proceedings ISSAC'97, Maui*, pages 197–204.

Courcoubetis, C., D.Stanmoulis, G., Manolakis, C., and Kelly, P. F. (1998). An intelligent agent for optimizing qos-for-money in priced abr connections. FO.R.T.H., University of Crete, Greece, University of Cambridge, England.

Datadomo (2001). http://www.datadomo.com.

Dovrolis, K., Prasad, M., and Ramanathan, V. P. The selection of the token bucket parameters in the ietf guaranteed service class.

Exodus (2001). http://www.exodus.com.

Gallager, R. G. (1996). *Discrete Stochastic Process*. Kluwer Academic Publishers.

genuity (2001). http://www.genuity.com.

Glasserman, P. (1997). Bounds and asymptotics for planning critical safety stock. *Operation Research*, 45(2).

Graves, S. C. (1982). The application of queueing theory to continuos perishable inventory systems. *Management Science*, 28(4).

Gridcomputing (2003). http://www.gridcomputing.com.

Gridforum (2003). http://www.gridforum.org.

Harrison, M. J. (1985). *Brownian Motion and Stochastic Flow Systems*. New York: Wiley.

Hosting (2001). http://www.hosting.com.

Hurwitz, G. (2001). Trends in e-business outsourcing and the rise of the managed hosting model - exodus communication, inc. www.exodus.com, www.hurwitz.com.

Intersystems (2001). http://www.intersystems.com.

Ioannis, P. C. and Yong, L. (2002). Large deviations-based asymptotics for inventory control in supply chains. Department of Manufacturing Engineering, Boston University, Working Paper.

Kelly, P. F. and Tan, D. (1998). Rate control for communication networks: shadow prices, proportional fairness and stability. University of Cambridge, England.

Kushner, H. J. and Dupuis, P. (2001). *Numerical Methods for Stochastic Control Problems in Continous Time*. Springer.

Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15.

MacKie-Mason, J. K. and Varian, H. R. (1994a). Some economics of the internet. http://www.sims.berkeley.edu/ hal/people/hal/papers.html.

MacKie-Mason, J. K. and Varian, H. R. (1994b). Some faq's about usage-based pricing. http://www.sims.berkeley.edu/ hal/people/hal/papers.html.

Maglaras, C. and Zeevi, A. (2003). Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions. To appear in Management Science.

McKnight, W. L. and Boroumand, J. (2000). Pricing internet services: after flat rate. *Telecommunication Policy*, pages 565–590.

Nic (2001). http://www.nic.be.

O'donnell, S. (2001). A first cut at the slicing of the internet dollar. MIT presentation.

Pancha, P. and Zarki, M. E. (1995). Leaky bucket access control for vbr mpeg video. *IEEE*, pages 796–803.

Paxson, V. (1997). End-to-end routing behavior in the internet. *IEEE/ACM Transactions on Networking*, 5(5):601–615.

Paxson, V. (1999). End-to-end internet packet dynamics. *IEEE/ACM Transactions on Networking*, 7(3):272–292.

Paxson, V. and Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244.

Qwest (2001). http://www.qwest.com.

Ross, S. M. (1974). Bounds on the delay distribution in gi/g/1 queues. *Journal of Applied Probability*, 11:417–421.

Schilkrut, A. Z. (2000). *Managing Customer Relationship Channels through Incentives and Service Quality*. PhD thesis, Sloan School of Management, MIT.

Sharcnet (2001). http://www.Sharcnet.ca.

Shenkar, S., Clark, D., Estrin, D., and Herzug, S. (1995). Pricing in computer networks: Reshaping the research agenda. In *In Proc. of TPRC 1995*. http://citeseer.nj.nec.com/shenker95pricing.html.

Siegmund, D. (1985). Sequential analysis: Tests and confidence intervals.

Superb (2001). http://www.superb.net.