Stochastic Modeling in Call Centers Management

Oualid Jouini

ÉCOLE CENTRALE DES ARTS ET MANUFACTURES
« ÉCOLE CENTRALE PARIS »

THÈSE

présentée par

Oualid JOUINI

pour l'obtention du

GRADE DE DOCTEUR

Spécialité : Génie Industriel
Laboratoire d'accueil : Laboratoire Génie Industriel

SUJET :

STOCHASTIC MODELING IN CALL CENTERS OPERATIONS MANAGEMENT

Soutenue le 11 décembre 2006,
devant le jury composé de

Philippe CHEVALIER Professeur, Université Catholique de Louvain, Belgique	**Président**
Yannick FREIN Professeur, INPG, Grenoble	**Rapporteur**
Ger KOOLE Professeur, Vrije Universiteit Amsterdam, Pays-Bas	**Rapporteur**
Fabrice CHAUVET HDR, Chef du pôle Simulation et Optimisation, Gaz de France	**Examinateur**
Yves DALLERY Professeur, Ecole Centrale Paris, Paris	**Directeur de thèse**

Laboratoire Génie Industriel
École Centrale Paris
Grande Voie des Vignes
92295 Châtenay-Malabry Cedex

2006-33

A ma famille,

Remerciements

Je tiens à exprimer toute ma gratitude à Yannick Frein et Ger Koole pour avoir accepté l'évaluation de cette thèse en tant que rapporteurs. Je remercie, également, Philippe Chevalier pour l'honneur qu'il me fait en présidant le jury de cette thèse, et Fabrice Chauvet pour avoir accepté d'en être examinateur.

Je remercie Yves Dallery pour la haute qualité de l'encadrement et du soutien qu'il m'a accordés tout au long de ces trois années. Grâce à lui, je dispose d'une meilleure méthodologie ainsi que d'un meilleur recul par rapport à la recherche que j'ai effectuée. Ses qualités humaines ont largement contribué à l'aboutissement de ce travail. Je tiens à lui exprimer ma vive reconnaissance et tout mon respect pour tout ce qu'il a fait pour moi.

Je tiens à remercier la société Bouygues Telecom dont la Direction Recherche a été à l'origine des problématiques étudiées dans cette thèse. Je suis fier d'avoir travaillé avec Fabrice Chauvet dont l'intérêt pour cette thèse a contribué sensiblement à une excellente collaboration, facilitée par ses excellentes compétences. Je remercie également Rabie Nait Abdallah, Eric Bouzou et Thierry Prat qui ont manifesté leur intérêt à ce travail.

Je remercie particulièrement Mohammed Salah Aguir qui m'a été d'une aide inestimable. Je lui exprime toute ma gratitude pour le temps qu'il m'a consacré et pour avoir suivi de près l'évolution de cette thèse.

Je garderai un excellent souvenir des discussions diverses que j'ai eues avec mes collègues du laboratoire les deux Zied, les deux Ali, Yacine, Nabil, Salma, Jean-Philippe, Ludovic, ... Ils n'ont cessé de m'encourager pendant les périodes difficiles. Je les remercie vivement pour ceci et je tiens à ce qu'ils sachent que leur amitié compte, et comptera, beaucoup pour moi.

Enfin, je rends hommage à tous les membres du Laboratoire Génie Industriel, véritable communauté. En particulier Anne, Sylvie, Corinne et, bien sûr, Jean-Claude.

Oualid

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we give a general introduction of the thesis. First, we provide a background about the call center industry. Second, we highlight some issues related to the design and management of call centers. Third, we describe the work of the thesis and present its main contributions. Finally, we present the structure of the manuscript.

1.1 Background

Call centers have emerged as the primary vehicle for firms to interact with consumers, transforming consumer service jobs once characterized by variety and personal relationships into routinized and high speed operations. Call centers are used to provide services in many areas and industries: banks, insurance companies, emergency centers, information centers, help-desks, tele-marketing and more. Technological development has allowed remote service delivery using various channels of telecommunication. The definition of a call center is continuously changing, but the core fundamentals of a customer making a call (via a phone, email, web site, fax or Interactive Voice Response) to a center (collection of resources) will remain constant. Call center, contact center or customer interaction center operate on identical principals of meeting customer needs in real-time or near real-time.

In 1972, Continental Airlines asked the Rockwell Collins division of Rockwell International (now Rockwell Automation) to develop the first automated call distributor, thus launching the contact center industry. Initially, little thought was given to the use of contact centers to acquire and retain business. The change came in the 1990s, with the advent of software-based routing and Customer Relationship Management (CRM) applications, which increased the marketing possibilities of contact centers. Today, all Fortune 500 companies have at least one contact center. They employ an average of 4,500 agents across their sites. More than $300 billion is spent annually on contact centers around the world, see McKinsey Quarterly [1]. In North America, 2.9 million agents are employed at 55,000 facilities. The number of agents in the rest of the world is predicted to increase by 10% a year from its current position of 3 million.

The current success of call centers is due to the technological advances in information and communications systems, see Pinedo et al. [110]. The most important call centers equipments are the Interactive Voice Response (IVR), the Automated Call Distribution (ACD), and the Computer Telephone Integration (CTI).

The IVR is a menu system that a customer accesses when connecting to a call center. The IVR routes a call to the most appropriate person or desk. The structure of the menu system can be a simple list of two or three items, or a more elaborate decision tree. This tool enables the system and the operator to provide the service in minimum time. The technology is relatively inexpensive when compared to the time wasted in the transfers of customers via live operators. Large banks spend between $1.75 to $2.00 for an operator handled call center transaction and between $0.25 to $0.75 for an IVR transaction, see NACCS [2]. The Automated Call Distribution (ACD) is a service provided by telephone companies that makes physically dispersed operators appear to a caller as residing at one location. The phone company handles the necessary switching

in order to make this happen. Finally, the Computer Telephone Integration (CTI) refers to the combination of computers and telephone systems. Most of modern call centers today are using some form of CTI technology. The CTI allows for example for the Intelligent Call Routing. The Intelligent Call Routing is an application that reads the phone number of an incoming call, retrieves information concerning the caller from a database, and presents it to the operator when he takes the call.

The large-scale emergence of call centers has been also enabled by the development of consulting services and softwares such as routing devices and databases. Naturally, the growth of the call center industry has created a fertile source of management issues. From the cost perspective, the capacity management is the most critical issue. In call centers, human resource costs account between 60% to 70% of operating expenses. This feature explains the huge body of operations management papers dealing with capacity management. The financial importance makes the running of call centers a challenge. The managers have to reduce the labor cost, but not to the detriment of the customers. We should not forget that call centers were born for a basic need: to answer customers in an efficient way so as they do not switch to the concurrence. An extra layer of complexity comes from the human resource management. Agents are indeed human beings. Hence, they need to feel strongly supported by the company so that undesirable and costly issues such as turnover are as low as possible. In most call centers, the lack of motivation and the bad conditions of work makes workforce turnover recurrent.

The goal of the present thesis is to contribute to the operations management research of call centers. We aim to enhance our understanding of such complex systems, so as we gain useful guidelines for the practitioners. This thesis is in part the result of a collaboration with *Bouygues Telecom*. *Bouygues Telecom* is a mobile telephony service provider based in France, operating one of the largest call center operations in France, and answering around 60,000 calls daily at 6 internal call centers with a total number of 2500 service representatives.

1.2 Context

The continued growth of both importance and complexity of modern call centers has been came along an extensive and growing literature. Numerous related academic surveys focusing on various disciplines were published. The main disciplines related to call centers are Mathematics and Statistics, Operations Research, Industrial Engineering, Information Technology, Human Resource Management, as well as Psychology and Sociology. This thesis is pertaining to operational issues and mathematical models. We refer the reader to Pinedo et al. [110] for the basics of call centers management. Important surveys are the paper of Koole and Mandelbaum [85] and

its extended version Gans et al. [40] where the authors survey in particular the literature dealing with queueing models that support the operations management of call centers. In addition, we recommend the overview of Whitt [138], and that of Mandelbaum [95] where the author provides a large number of research papers devoted to call centers issues.

Decisions Levels

We distinguish three main issues dealing with the operations management in call centers. The first issue involves strategic or long-term decisions for the design of the facility. The second issue is related to medium-term aggregate planning of services. The third issue deals, in turn, with short-term decisions made on a daily or weekly basis.

The strategic decisions involve the allocation of resources (equipments) as well as the layout and location of the facilities. Included in this category of decisions are those specificating how to partition customers into classes and how the different communication channels are to be used for serving the customers: for example, which types of customers are to be answered by automates, internal agents, external agents (outsourcing), etc. The medium-term decisions involve the development of a semi-annual or annual manpower plan. The plan will have as inputs the anticipated demand for different skill sets over the planning horizon, the costs of training, and the time to train. Forecasts are usually made monthly and mathematical models are used to determine the appropriate staffing levels on an aggregate basis. Thereafter, we address the shift scheduling problem of servers. The short-term decisions deal with refinements and adjustments that are executed within a short time period and trigged by external factors. One may clearly see that all decision levels are correlated and should be addressed simultaneously. Unfortunately, such an analysis is too complex. The research projects have often investigated them separately.

Call Centers Modeling

Due to the uncertainty governing the call center environment (customers and agents behaviors), the literature has standardly addressed its issues using stochastic models, and in particular queueing models. The most existing work deals with simple queueing models. Although, researchers have started recently investigating important phenomena such as abandonments, retrials and non-Markovian processing times, several questions are still open and much remains to be done. In the natural way, some research projects rely on simulation to analyze complex stochastic models of call centers, see Wallace and Whitt [128]. Other papers resort to heavy-traffic approximations. By heavy-traffic, we mean that the system approaches saturation, so that the queues are non-empty most of the time. Thereafter, diffusion approximations can be used to analyze the system behavior. The asymptotic analysis is motivated by large call centers where the number of agents and the arrival rate are very large. Guided by the asymptotic behavior,

several useful insights for practical management could be derived. We refer the reader to the original work of Halfin and Whitt [52] for a background on the subject. For more recent works focusing on call centers, interesting papers are those of Garnett et al. [44], Armony and Maglaras [12] and [13], Borst et al. [28], and references therein.

Call centers can be broadly classified into two contexts: multi-skill call centers and full-flexible call centers. A multi-skill call center handles several types of calls, and agents may have different skills. The typical example, see Gans et al. [40], is an international call center where incoming calls are in different languages. We distinguish some important issues dealing with multi-skill call centers: are all agents cross-trained with all skills or are the agents only trained for a subset of skills? In the later case, what are the subsets of skills that will be considered and how many agents will have each subset of skills? Another central issue is the skill-based-routing (SBR) algorithms: how to do the routing of calls to agents in an effective manner? Related studies include those by Garnett and Mandelbaum [43], Akşin and Karaesmen [8], Akşin et al. [9], Hopp and van Oyen [57], and references therein. At the same time, addressing such problems is complex and of great interest for practitioners. In modern call centers, it is indeed common to have multiple types of calls and multiple types of agents. These topics are out of the scope of this thesis. The second context is call centers where all agents are able to handle all types of calls, referred to as full-flexible call centers.

Our concern in this thesis is full-flexible call centers. Therefore assistance to customers can be provided by any agent. This would be a plausible assumption for many real cases, especially for unilingual call centers where the complete flexibility is not as difficult as in multilingual call centers. Furthermore, we assume for the models we consider in this thesis that all agents are totally identical statistically. In other words, they can answer all questions coming from customers with the same efficiency, both quantitatively and qualitatively, even in case of different types of customers. There are two reasons for that. The first reason is related to the nature of the call centers we are considering here, and which is the case for many other call centers applications. The difference between customer types is only qualitative, i.e., it is not related to the statistical behavior of customers but to their importance for the company. Let us give an illustration of a manager who partitions customers into two different classes: If the company owns every month from one customer an amount of money crossing a given threshold, then that customer is “VIP", otherwise he is considered to be less important. In concrete terms, we assume for our models that the queries asked by customers do not differ from one type of customers to another. Therefore, it would be credible to assume common requirements for service. The second reason is due to the complexity of the analysis when assuming different behaviors in the statistical

sense. Our main objective in this thesis is to investigate simple but at the same time interesting models that allow us to better understand the system behavior and gain practical guidelines.

Call centers are characterized to handle either inbound calls, or outbound calls, or a mix of both types. Inbound call centers handle incoming calls that are initiated by customers, as help desk and reservation services. However, outbound call centers handle outgoing calls that are initiated by agents. In our models, we deal with inbound call centers.

Motivation

In the following, we highlight some motivations with regard to the models under consideration. Until 2004, the organization of the *Bouygues Telecom*'s call center was a common pooled organization where any call could be addressed by any agent. In 2004, the managers of *Bouygues Telecom* decided to move to a new organization known as customer portfolio management. Since Human Resources (HR) represent a large part of the cost of operating a call center, managing efficiently the HR is a key issue. In particular, we deal with the issue of partitioning the HR of a large call center into a set of independently managed teams. The advantage of this new organization pertains to the better management of the team organization which leads to a higher motivation of the customer representatives and a reduction in the turnover of the workforce.

In the most of our models, we incorporate an important feature which is the reneging (abandonment) phenomenon. The time before reneging (the patience) is defined as the maximal amount of time that a customer is willing to wait for service. If service has not begun within that time, the customer abandons (leaves the system.) Incorporating reneging in theoretical models is of value. In reality, it is natural that a waiting customer will wait for only a limited time, and will hang up within that time. Ignoring reneging leads to overstaffing and pessimistic estimation of queueing delays. Garnett et al. [44] show using numerical examples that models with and without abandonment tend to give very different performance measures even if the abandonment rate is small. Models including reneging are therefore more close to reality, and necessary to obtain more accurate managerial insights. In our models, we assume that times before reneging are identically distributed, independently from their position in queue. The motivation of this assumption was reported in Gans et al. [40]. In call centers, the tele-queueing experience is, indeed, fundamentally different from that of a physical queue, in the sense that customers do not see others waiting and need not be aware of their “progress" (position in the queue.)

In this thesis, we also discuss various scheduling policies subject to satisfying some given service levels. Randolph [113] classifies scheduling policies into those using dynamic scheduling

rules and those using static schedule rules. A dynamic schedule is a discipline that is continuously updated as customers arrive and are processed. However, a static schedule is independent of the state of the system, it is beforehand defined and never altered. Each one of the above classes of policies can be further classified into two major types: agent scheduling and customer routing. As defined in Garnett and Mandelbaum [43], agent scheduling is described by a control decision taken whenever an agent turns idle and there are queued customers: which customer, if any, should be routed to this agent. Whereas, a scheduling policy based on a customer routing rule, is defined by a control decision taken whenever a customer arrives: which idle agent, if any, should serve this customer, if not, to which queue should the customer of interest be routed. In our analysis, we consider dynamic scheduling policies based on the second type of control decision, i.e., dynamic assignment rules of new arrivals to queues. In addition, our policies are based on priority schemes. In general, the provision of differentiated service levels relies on the use of priority queues. Schrage and Miller [117] have shown that scheduling policies similar to multiclass priority queues allow to achieve high performances, often nearly as good as those under optimal policies. Also, the priority schemes are easy to implement, which explain their prevalence in practice.

Recently, call centers have started experimenting by informing arriving customers about anticipated delays. The main reason of the experience is to alleviate congestion and reduce customer dissatisfaction with waiting. Information about anticipated delays is especially important in service systems with invisible queues (tele-queue) such as call centers. In such systems, the uncertainty involved in waiting is higher than that in systems with visible queues. Upon arrival and during their waiting, customers have no means to estimate queue lengths or progress rate. So, the feelings of frustration and anxiety increase over their sojourn in queue. In addition, we point out a particular vicious circle. When a new arrival customer perceives that his anticipated delay is too long, he may balk upon arrival without joining the system. This feature would considerably reduce customers reneging in queue, which allows to make the system more stable in the sense that the variability of queueing delays is reduced. The latter would in turn improve the quality of delays information we give to customers, which reduces even more customers reneging, and so on.

As mentioned above, stochastic processes and queueing models are helpful for the quantitative analysis of call centers. Birth-death processes and in general Markov chains are a rich and important class in modeling numerous phenomena in queueing systems. For instance, they allow to account for customers balking and reneging in call centers. The analytical studies in the literature were intended to obtain useful information for the decision making process, basically

related to the design, the control, and the measurement of effectiveness of the systems. With an equal interest, we underline the usefulness of monotonicity properties of performance measures. They are important for understanding and solving optimization problems of queueing systems. Optimization models are being used increasingly in the design of a variety of systems where queueing phenomena arise. Examples include flexible manufacturing systems, as well as service systems and telecommunications networks. For such problems, it is important to know the convexity properties of the performance measures with respect to the design variables. In call centers, the design variables on which the service provider could act are essentially the staffing level, the arrival rate (outsourcing) and the buffer size. In some cases, it could be possible for him to act on processing times (for example by increasing or decreasing the training quality of the agents.)

1.3 Description and Main Contributions

The current thesis can be divided into two parts. The first part directly addresses issues of the operations management in call centers. The second part is rather focusing on stochastic models while having useful applications for the quantitative analysis of call centers.

The first part

In the first part of the thesis, we focus on two different decision levels: long-term and operational decisions. The long-term issue is addressed in the second chapter. In the third and fourth chapters, we focus on two different operational decisions given a call center structure and staffing level.

In the second chapter, we study a call center design problem where a transition occurs from a completely pooled structure to a dedicated team-based organization. As one would not expect, we show using simple queueing models that the new organization may be more efficient by outweighing the economies of scale associated to the original organization. The efficiency is in terms of both speed and quality of the answer we provide to customers. We incorporate in our analysis the most attractive feature of call centers, which is the human element. We show how a better human resource management may lead to various benefits up to contradict a classical result of queueing theory (in favor of pooled systems). In addition, we slightly modify the new organization by profiting from a flow of customers that can be addressed by any team of agents (out-portfolio flow.) We show how this new element makes the organization even more efficient. In practice, several other call center cases may be characterized by a kind of an out-portfolio flow. For example in a bank call center, an out-portfolio flow may be seen as that of the customers who ask general questions about their bank account or to order some simple service

operations, etc. The application of the team-based organization had very significant effects at the *Bouygues Telecom*'s call center. The quality of answers has improved thereby reducing call backs by 25%. The proportion of disconnected calls (because of a full queue) was divided by 2 (in our work we assumed an infinite queue for simplicity.) Finally, no supplementary agents were hired in spite of the increase of the total number of customers by 15% (equivalent labor cost savings of about 5 million euros per year.)

In the third chapter, we consider a two-class call center model. We discuss various scheduling policies subject to satisfying differentiated service levels. The service levels are related to the probability of being lost and the variance of the waiting time in queue. The nature of the constraints we consider are characterized to be of value in practice, however they are not too much addressed in the literature. We aim even in case of unfavorable situations that may occur, to reach a fixed balance between customer types service levels, independently of the available service capacity. Worrying about the fairness with regard to customers, we also focus on achieving low values of the variance of waiting times. The interesting side of the policies we suggest comes from their simplicity, they are predictable, easy to implement, and do not require information about the workload process. Several studies as in Jongbloed and Koole [63] and Avramidis et al. [18] have shown that the workload process is hard to predict in call centers. Thereby, such policies would be of great value. Our analysis yields to quantitative insights, as well as useful principles for the control problem. To support our analysis, we derived various structural results that investigate the relationship of scheduling policies with the achieved performance measures. To the best of our knowledge, the results are not given beforehand.

In the fourth chapter, we consider call centers models not too far from those of the third chapter, whereas we tackle a different issue of operations management. A central outcome of the fourth chapter deals with the critical issue of the impact of delays information on customers behavior. This is at the same time interesting and difficult due to the attractive human element governing the call center environment. Starting from each model (single and multiclass), we detail and justify the quantitative building of the new model with delays information. In our models, customers have the opportunity to balk in response to their anticipated delay. We model that effect for the simple single class call center. We then extend the model of Whitt [135] by letting already informed customers renege even after having chosen to join the queue. We propose a method for approximating the new reneging experience by pertaining it to the quality of the delay information. We describe how balking in the second model may reduce customers reneging. In practice, this feature would tip the scales in favor of the second model, because reneging customers are the costliest. For example, a customer who balks has a higher

probability to call back than that of a customer who reneges. A reneging customer leaves the system with frustration and loosing trust in the service provider. However, a balking customer leaves the system based on the information we communicate to him. This information would avoid to loose business because it is perceived by balking customers as an invitation to call back when the system will be able to serve them within a reasonable delay. As extension, we turn to analyze a quite complex multiclass priority system where the anticipated delay for a given type of a customer may be affected by future arrivals of other types. To our knowledge, the computation of the state-dependent virtual delays for the low priority customers is new. We use a two-dimensional Markov chain in order to derive them.

The second part

In the second part of the thesis, we focus on the analysis of stochastic processes and queueing theory. The analysis does not address in a direct manner a given issue of call center operations. However, it is an upstream stage which provides useful applications for the quantitative analysis of call centers.

The topic addressed in the fifth chapter is of interest in the fields of birth-death processes and Markovian queues. We give closed-form expressions for the moments of the so-called upcrossing and downcrossing times as well as conditional versions of these. Our approach is different from that in some classical works in the sense that we are not considering the correspondence between continuous birth-death processes and continued fractions. Based on the Chapman-Kolmogorov equations and via Laplace transforms, several new expressions are derived. Also, we discuss various straightforward applications for the quantitative analysis of Markovian queues. The results are in particular of value when characterizing state-dependent queueing delays in call centers.

In the sixth chapter, we derive some monotonicity properties for the probability of being served in an $M/M/s/K + M$ queue. Such results are helpful for the optimization problem of queueing systems. We use both sample path as well as analytical approaches to derive our results. The model we consider allows for reneging, which makes it to be relevant for call centers applications. As we already mentioned, a major drawback in many call center models is assuming customers to be infinitely patient. In this chapter, we analyze the simplest abandonment model, assuming that service times and times before reneging are exponentially distributed. Although such assumptions may be violated (see Zohar et al. [146]) and an appropriate model should be the $M/GI/s + GI$ queue, our model is still of interest in practice as mentioned by Whitt [139], and Pierson and Whitt [109]. The authors have shown, using various simulation experiments, that the $M/M/s + M$ model provides a good approximation of the $M/GI/s + GI$ model.

1.4 Structure of the Manuscript

In this section, we present the structure of the manuscript. We briefly describe the different chapters separately and give their corresponding published or working papers.

In Chapter 2, we investigate the benefits of migrating from a call center where all agents are pooled and customers are treated indifferently by any agent, towards a call center where customers are grouped into clusters with dedicated teams of agents. This Chapter is based on Jouini, Dallery and Nait-Abdallah [69].

In Chapter 3, we consider a priority call center model with two impatient classes of customers, VIP and less important ones. We focus on developing scheduling policies that assign customers upon arrival to parallel queues, high and low priority queues. The performance measures of interest are the probability of being lost (due to reneging) and the variance of the waiting time for the customers who are served. An extended version of this chapter is the working paper Jouini, Pot, Dallery and Koole [70].

In Chapter 4, we study the effect of informing customers about their anticipated delays. We propose a method for modeling the customer reaction with regard to delays information. Then, we conduct comparison analysis between performance measures of both models with and without information. This chapter is based on the working paper Jouini, Dallery and Akşin [68].

In Chapter 5, we consider ordinary and conditional first passage times between pairs of states in general birth-death processes. By adopting classical methodologies, we derive closed-form expressions for the moment of the defined random variables. This chapter is based on the paper Jouini and Dallery [65] (submitted for publication).

In Chapter 6, we consider a Markovian multiserver queue with a finite waiting line in which customers may renege. We focus on a performance measure similar to that considered in Chapter 3, namely the probability for a new customer to enter service. We investigate monotonicity properties of first and second order of this performance with respect to the buffer size. The paper version of this chapter is Jouini and Dallery [66].

In Chapter 7, we close the thesis by giving general concluding remarks and highlighting directions for future research.

Chapter 2

Analysis of the Impact of Team-Based Organizations in Call Centers Management

In this chapter, we address a design issue related to long-terms decisions in call centers. We investigate the benefits of migrating from a call center where all agents are pooled and customers are treated indifferently by any agent, towards a call center where customers are grouped into clusters with dedicated teams of agents. Each cluster is referred to as a portfolio. The purpose of this chapter is to examine how the benefits of moving to this new organization can outweigh its drawback. The drawback comes from the fact that there is less pooling effect in the new organization than in the original one. The benefit comes from the better human resource management that results in a higher efficiency of the agents, both in terms of speed and in terms of the quality of the answer they provide to customers. Also, we extend the analysis to the case where there is an additional flow of calls called out-portfolio flow. It is shown that this feature makes the new organization even more efficient.

The paper version of this chapter is Oualid Jouini, Yves Dallery and Rabie Nait-Abdallah [69]. It was accepted for publication in *Management Science*.

2.1 Introduction

The work in this chapter is the result of a collaboration with the French mobile phone company *Bouygues Telecom*. Our purpose is to provide some insights into the impact of internal organization of call centers on their performances. The *Bouygues Telecom* call center handles an average of 100,000 phone calls daily. Some of the calls are treated by an automated operator. Agents, also called customer representatives, deal with about 60% of these contacts. There are also about one million contacts per year handled by mail, e-mail and fax. Here, we investigate the adequacy of migrating from a call center where all agents are pooled and customers are treated indifferently, towards a call center where customers are grouped into clusters with dedicated agents. In our terminology, each cluster will be called a portfolio. Customers that do not fit into a precise portfolio generate the so-called out-portfolio flow, and must wait in a lower priority out-portfolio flow queue. Managers of *Bouygues Telecom* believe that the challenge is not only to answer quickly but also to answer customers correctly. In this sector (mobile telephony), it is not rare to see customers switching from one company to another as a consequence of low quality responses provided by customer representatives. Agents are the interface between the company and the customers; hence, customer satisfaction is closely linked to agents performance. Managers need to motivate their employees so that the assistance they provide to customers is efficient, both in terms of speed and quality of answers. On the other hand, employees need to feel strongly supported by the company so that the turnover is as low as possible. In fact, turnover means training new employees, and it implies more costs.

The aim of *Bouygues Telecom* through migrating into customer portfolio management is to better manage their employees and as a consequence to satisfy customers more efficiently. This management approach makes agents more responsible towards their own customers. Moreover, partitioning agents into groups creates competition, which increases agents motivation. These factors result in overall agents efficiency improvement, both quantitatively and qualitatively. By quantitative efficiency, we mean the speed (processing time) in providing assistance to the customers. By qualitative efficiency, we mean the quality provided by the agents when addressing the customers request. In the present chapter, we argue that these advantages may outweigh the variability that results from the loss in economy of scale originally associated with the pooled system. In addition, in the proposed organization, all portfolios and corresponding sets of dedicated agents are identical (statistically.) Therefore, issues such as training and forecasting can be done in a homogeneous manner. Also, having homogeneous teams yields a more efficient human resource management. In fact, it allows the call center manager to compare the teams performances, which results in a “global competition".

Such a managerial approach has been widely and successfully used in industry and is also likely to be of interest in service activities such as call center operations. It is, indeed, one of the key success factors of the so-called World Class Manufacturing. For example, Schonberger [116] refers to it as cellular manufacturing and describes its benefits as follows: "Cells create responsibility centers where non existed before. The cell leader and the work group may be charged with making improvements in quality, cost, delays, etc."

The remainder of this chapter is structured as follows. In Section 2.2, we review two kinds of literature closely related to our work. The first one is on pooling, and the second is on integrating human factors in queueing systems, and in particular in call centers. In Section 2.3, we give a comprehensive presentation of the problem we study in this chapter. In Section 2.4, we develop a simple queueing model that is then used to address the issue of benefits versus costs of migrating from the pooled organization to the dedicated organization when there is no out-portfolio flow. We provide some interesting insights on the tradeoff between reduction of the pooling effect and agents efficiency improvement, both quantitatively and qualitatively. In Section 2.5, we extend this analysis to the situation where there is an out-portfolio flow. To do that, we first develop some approximate queueing models of the call center operating with a mix of portfolio and out-portfolio flows. One additional insight is that the drawback of not having a totally pooled system is less important in this context. Finally, we conclude and propose some directions for future research.

2.2 Literature Review

In this section, we review some papers related to the work of this chapter. Our work is pertaining to two areas of literature, one dealing with pooling and the other with human factors in queueing systems. The literature dealing with pooling falls mainly into two categories: pool queues or pool servers, see Mandelbaum and Reiman [97]. Kleinrock [82] is one of the first researchers who gave a depiction of these alternative structures of pooling. He began by considering a collection of m identical $G/G/1$ queues, each of which has a single server with service rate μ and faces a job stream at rate λ. Pooling only queues would change this collection into a $G/G/m$ queue, which has m servers, each server with service rate μ and a job stream at rate $m\lambda$. Pooling only servers would change the last $G/G/m$ queue into a $G/G/1$ queue with arrival rate $m\lambda$ and service rate $m\mu$. In this chapter we only deal with queue pooling issues.

As mentioned in the introduction of this thesis, we are not dealing with multi-skill call centers. Our concern in this chapter, as well as in the following ones, is full-flexible call centers. We only consider queueing models in which all agents (servers) have all skills, i.e., all agents are flexible

enough to answer all requirements of service. It is a plausible assumption for *Bouygues Telecom* as for many other call centers. Often, it is the case for unilingual call centers where the complete flexibility is not as difficult as in multilingual call centers. In this chapter, we deal with the issue of the level of pooling in full-flexible call centers, i.e., are the agents all gathered into a single large team or are they partitioned into a set of independent teams? Several papers discuss the effectiveness of pooling in call centers, see for example Tekin et al. [125] and references therein. Beyond service systems, pooling effect problems arise in various applications, such as manufacturing, and computer network systems. Pooled systems are usually preferred. The standard argument for combining queues is due to the economies of scale, which absorbs stochastic variability (Borst et al. [28].)

While it is easy to see that pooled systems are more effective than independent ones, this intuition was for a long time based on experience and numerical data rather than rigorous mathematical proof. Smith and Whitt [120] were the first to formally prove this result, when combining systems with identical service time distributions. They applied analytic methods for the $M/M/C/C$ loss systems (Erlang-B, no waiting room) and the $M/M/C$ delay systems (Erlang-C, infinite waiting room.) By using sample-path methods, they also showed that efficiency increases through combining queues in systems with general arrival processes and general service time distributions. Benjaafar [24] extended theses results by providing performance bounds on the effectiveness of several pooling scenarios. When we allow service rates in separate systems to become different, combining queues can be counterproductive (Smith and Whitt [120], Benjaafar [24].) Van Dijk and van der Sluis [127] presented a case-study simulation supporting this outcome. Using approximations for $M/G/C$ performance measures, Whitt [136] explored the tradeoff between economies of scale associated with larger systems and the benefit of having customers with shorter service times separated from customers with longer service times.

All of the above results do, in no way, take into account the human element. This takes us to the second area of literature close to our work. Human element is the main characteristic of call centers and contact centers. Both customers and agents are people. Even though it is natural to focus on understanding human behavior, few papers integrate this aspect to analyze call centers and, in general, queueing systems. We refer the reader to the survey of Gans et al. [40] where we find some references examining queueing models of call centers that incorporate customers behaviors, such as, abandonments and retrials. Some other models include the link between agents and customers experiences. In 1987, two papers have launched discussions about human factors in queueing systems. The first is Larson's [90] paper which goes beyond the classical interest on delays and points out the psychological experiences of people in queues. He argues

the importance of perceptions of fairness, and shows for instance how the violation of the first come, first served order may contribute to customers dissatisfaction. The second is Rothkopf and Rech's [115] paper, which deals with the question of combining queues. The authors in that paper discuss the tradeoff between pooled and separated systems by including customers reactions and jockeying between separate queues (a customer can change to one queue while he was waiting in another.) Moreover, they show how separate systems may lead to servers that are more responsible towards their own customers. It may also allow for a faster service due to the degree of specialization gained through experience. To our knowledge, they were the first to emphasize this issue.

Fischer et al. [38] conclude that call center management requires a mix of disciplines that are not typically found in organizations. The review of Boudreau et al. [30] follows through this new area. They propose a framework which is a fertile source of research opportunities. They justify by real examples that operations management itself, without human resource management, can not well analyze systems such as those we are dealing with, and vice versa. In others words, there is a mutual impact between the two fields, and taking into account this fact yields to more realistic and precise insights. In particular, Boudreau et al. [30] consider that more realistic operations management models need to integrate human factors, such as; turnover, motivation and team structure. In fact, a team setting allows for better communication, and may allow for more responsible and motivated agents. In a recent paper, Boudreau [29] underlines once again the significant opportunities for fruitful research at the boundaries between the traditional topics of operations management and human resource management. The present chapter addresses this issue in a call center context. We explore how managing agents by creating separate pools might lead the agents performing more efficiently.

2.3 Problem Setting

In this section, we present the general problem under consideration in this chapter. Consider a company operating a fairly large call center. The call center provides assistance to the customers of the company. Customers call the company whenever they need assistance and their request is addressed by a set of agents. Recall that in the setting of this work, we assume that the call center is operated in such a way that all agents have the same skill. Therefore assistance to the customers can be provided by any agent. In other words, all agents are totally identical (statistically) in the sense that they can answer all questions coming from the customers with the same efficiency, both quantitatively and qualitatively.

2.3.1 Current Organization Mode

Let us describe the behavior of the call center under the current organization mode. The call center is operated in such a way that at any time, any call can be addressed by any agent. So, whenever a call arrives, it is addressed by one of the available agents, if any. If not, the call is placed into a queue and will be addressed as soon as possible. There is a single queue and waiting calls are answered on a first come, first served (FCFS) basis. For simplicity, we assume that the queue has no capacity constraint and that customers do not abandon while waiting. Under this organization, the agents have a given efficiency. The quantitative efficiency is measured by the distribution of the processing times, which represents the time it takes for an agent to answer a call. Note that the randomness of the processing times comes in particular from the variety of questions asked by the customers. The qualitative efficiency is measured by the probability of successfully answering the question of the customer. We assume that if the call has not been addressed in an adequate manner, the customer will call back to get assistance from another agent. This concept of call resolution probability was argued by de Véricourt and Zhou [36] in a call routing problem. As for the global efficiency of the call center under the current organization, its positive side comes from the pooling effect. Its negative side is in terms of human resource (HR) management, given that, it is usually very difficult to have an efficient management of a large set of agents in a large call center.

2.3.2 New Organization Mode

Let us describe the following new organization mode. The set of agents is split into a set of independent teams. The teams are homogeneous in the sense that they have the same number of agents and that all agents have the same skills. In other words, there is no specialization. Let n be the number of independent teams.

In the new organization, in addition to the partitioning of the total number of agents in a set of autonomous teams of agents, there is also a partitioning of the customers into a set of n customer portfolios. Again, this partitioning is done in such a way that the portfolios are homogeneous. In other words, the overall request coming from the different customer portfolios are statistically identical. So, whenever a call arrives from a customer of a given portfolio, it is routed to the corresponding team. The behavior at the team level is then exactly identical to that described above for the original large call center. This new organization is equivalent to operating independently n smaller call centers with each call center having its own customers portfolio. Again, it is important to emphasize that under this new organization, all teams and customer portfolios have the same behavior (statistically.)

In the research study we performed with *Bouygues Telecom*, the size of the original call centers (total number of agents) was in the order of 2000, and they were considering team sizes ranging from 40 to 100 agents. Because all agents are not always present, this would mean that the number of agents simultaneously present in the call center would be in the order of 1000 and the corresponding number of agents present in each team would be ranging from 20 to 50. The reason advocated for moving to this new organization was along the line of the World Class Manufacturing literature. Namely, that the human resource management could be performed in a much better way at a small team level rather than at the global call center level. Agents motivation and responsibility would increase. Performance measures, both quantitative (processing times) and qualitative (rate of calls successfully addressed), could be examined more appropriately and could be used for internal team management. Due to the team/portfolio one-to-one link, a customer not satisfied with the answer he got from the agent would call back and the additional burden would fall on the same team. Also, the fact that all teams are homogeneous would allow for performance comparisons between the different teams resulting in a “global competition". Incentives could be given to agents based on the global performance of the team.

2.3.3 Research Objectives

In this research project, our goal is to study the tradeoff between the pros and cons of moving from the original organization to the team-based organization, also referred to as the portfolio organization. To do that, we consider a simple stochastic model of the original pooled organization. This model captures the original behavior of the call center when all agents are pooled. Under this situation, the call center has a nominal behavior in terms of efficiency (quantitative and qualitative efficiencies.) It achieves a given quality of service (QoS). We actually consider two different QoS measures: the average waiting time and the 80/20 rule, which is an industry standard for telephone service, see Gans et al. [40]. Under the 80/20 rule, at least 80% of customers must wait no longer than 20 sec.

In this work, we study the increase of efficiency required so that the team-based organization achieves the same QoS as the pooled system with the same total number of agents, i.e., no cost increase. We successively consider the case where the improvement comes only from the increase of the quantitative efficiency (decrease of the average processing time) and then the case where the improvement comes only from the increase of the qualitative efficiency (increase of the average rate of successful answer.) To perform the different analyzes described above, we consider a generic model that captures the important features needed for the comparison between

the pooled organization and the team-based organization. As it is often the case in call center modeling, our analysis is based on the use of a stationary queueing model (Gans et al. [40].) We use standard assumptions on the nature of the underlying processes: Poisson arrival processes, and exponential service time. These assumptions are plausible for *Bouygues Telecom* as for many other cases of call centers, especially for the arrival processes. In addition, we assume that calls not satisfied successfully occur randomly and therefore the splitting of the output flow follows a Bernoulli process. Moreover, delays before customers call back are assumed to be i.i.d. random variables. This allows us to use simple results from standard queueing theory. The generic model under consideration (for both the pooled organization and the team-based organization) is illustrated on Figure 2.1. The results of our study are presented in Section 2.4.

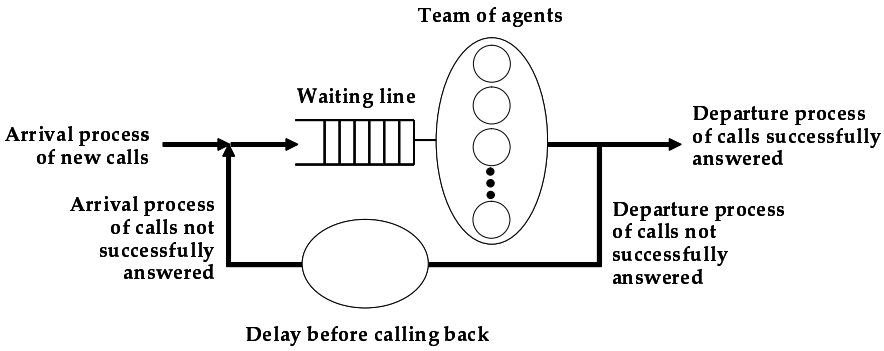

Figure 2.1: The generic model

2.3.4 Out-Portfolio Flow

There was another important feature in the *Bouygues Telecom* call center. Not all the calls received at the call center could be identified as belonging to a given portfolio of customers. Therefore, the actual situation was that in addition to the flow of calls coming in for each portfolio, there was an additional flow of independent calls, referred to as out-portfolio calls. In the original pooled organization, the calls were treated with lower priority. For the team-based organization, all out-portfolio calls are sent to a single queue. These calls can be served by any agent of any team, but portfolio calls have (non-preemptive) priority over out-portfolio calls. This means that when an agent becomes available, he deals with a call from his portfolio first (the first call in line.) If the queue is empty, this agent provides service to a call from the out-portfolio queue (the first in line.) Under this more general setting, we also investigate the improvement in either quantitative or qualitative efficiency that would be required to counterbalance the unpooling effect. The results of our study are presented in Section 2.5.

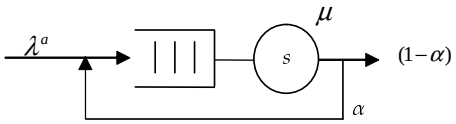

Figure 2.2: Pooled System

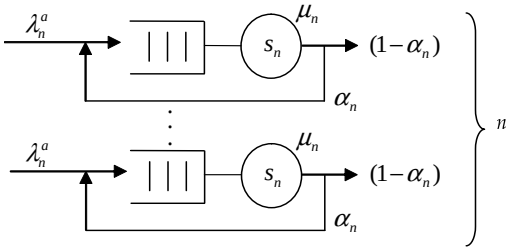

Figure 2.3: Dedicated System

2.4 Analysis of the Efficiency of the Team-Based Organization

In this section, we restrict our attention to the situation where there is no out-portfolio flow. In Section 2.4.1, we present the relevant queueing models and determine the performance measures of interest. In Sections 2.4.2 and 2.4.3, we analyze the required increase in terms of quantitative or qualitative efficiency that must be achieved in order to counterbalance the reduction of the pooling effect. Finally, in Section 2.4.4, we provide some further insights on the advantages of the team-based organization.

2.4.1 Modeling and Performance Analysis

Consider first the queueing model of the original call center. The model consists of a single infinite queue and a set of s identical servers representing the agents. Service times are assumed to be exponentially distributed with rate μ. The arrival process of first-attempt calls (primary calls) is assumed to be Poisson with an arrival rate of λ^a. There is a probability α that the customer is not satisfied with the answer he got and therefore will call again. Thus, $(1-\alpha)$ represents the probability that a call is successfully answered. Delays before customers call back are assumed to be i.i.d. random variables with a general distribution. For tractability purposes, we assume independence between successive calls, both in terms of service times and probability of success. Let λ be the overall arrival rate to the queue, i.e., the sum of the primary calls and the feed-back calls. Under stability conditions, $\lambda = \lambda^a/(1-\alpha)$. This simple model falls into the class of product-form networks analyzed by Baskett et al. [22]. As a result, the stationary behavior of this queueing model does not depend on the distribution of the call-back delays. They can thus be ignored. The resulting model is shown on Figure 2.2. It is equivalent to a simple $M/M/C$ queue with $C = s$ servers, a Poisson arrival rate $\lambda = \lambda^a/(1-\alpha)$ and a service rate μ. This model will be referred to as the Pooled System.

Consider now the modeling of the team-based organization with a partitioning of the original call center into n autonomous teams, each one being associated with a customer portfolio. It is assumed that n is such that s is a multiple of n. Recall that all teams and all customer portfolios

are statistically identical. Under this organization, the call center can be modelled as a set of n independent and identical queueing models. In the following, we focus our attention on the generic model of any portfolio/team subsystem. The assumptions are similar to those above. The model consists of a single infinite queue and a set of s_n identical servers. Service times are assumed to be exponentially distributed with rate μ_n. The arrival process of first-attempt calls (primary calls) is Poisson with an arrival rate of $\lambda_n^a = \lambda^a/n$. The probability that a call is not successfully answered is given by α_n. The resulting model is shown on Figure 2.3. It is equivalent to multiple simple $M/M/C$ queues, with $C = s_n$ servers, a Poisson arrival rate $\lambda_n = \lambda_n^a/(1-\alpha_n)$ and a service rate μ_n. This model will be referred to as the Dedicated System. Note that the Pooled System can be viewed as the particular case of the Dedicated System for $n = 1$. In the *Bouygues Telecom* call center, like in most call centers, the arrival rate of calls varies over time. Therefore, we use queueing models to estimate stationary system performance of half-hour intervals. We assume constant number of agents, and constant arrival and service rates, as well as a system that achieves a steady-state quickly within each half-hour interval of time, see Gans et al. [40].

Consider the Dedicated System. Let $W_n(t)$ be the Probability Distribution Function (PDF) of the waiting time in the queue. Let $r_n = \lambda_n/\mu_n = \lambda_n^a/(1-\alpha_n)\mu_n$ be the traffic intensity, and $\rho_n = r_n/s_n$ the server utilization (proportion of time each server is busy.) Note that the condition for existence of a steady-state solution is $\rho_n < 1$; that is, the mean total arrival rate must be less than the mean maximal service rate of the system. As in Gross and Harris [47], Equation (2.1) gives the probability that the waiting time in the queue is less than t.

$$W_n(t) = 1 - \frac{r_n^{s_n}\, p_n^0}{s_n!\,(1-\rho_n)} e^{-(s_n\mu_n - \lambda_n)\,t}. \tag{2.1}$$

Equation (2.2) gives the expression of the average waiting time in the queue.

$$W_n = \left(\frac{r_n^{s_n}}{s_n!\,(s_n\mu_n)\,(1-\rho_n)^2}\right) p_n^0. \tag{2.2}$$

Equation (2.3) gives the expression of p_n^0, which is the stationary probability of finding no customers in the system.

$$p_n^0 = \left(\sum_{i=0}^{s_n-1} \frac{r_n^i}{i!} + \frac{r_n^{s_n}}{s_n!\,(1-\rho_n)}\right)^{-1}. \tag{2.3}$$

The above equations give performance measures of the Dedicated System with n teams. The Pooled System is a special case of the Dedicated System. Performance measures of the Pooled System are obtained by using $n = 1$, $s_1 = s$, $\mu_1 = \mu$, and $\lambda_1^a = \lambda^a$ in Equations (2.1), (2.2) and

(2.3).

Due to the variability effect in the arrival and service processes, the comparison between the Pooled System and the Dedicated System will always show that, for any positive integer $n \geq 2$, the Pooled System outperforms the Dedicated System under the same situation, i.e., $s_n = s/n$, $\mu_n = \mu$, $\lambda_n^a = \lambda^a/n$, and $\alpha_n = \alpha$. Under these conditions, it is intuitively clear that the Dedicated System is less efficient because a call may wait for one server (of one team) while another server (of another team) is idle; such a situation does not occur in the Pooled System.

In this section as in the next one, we perform the study of the quantities of interest as a function of the number of dedicated pools, n. Alternatively, we could have chosen to perform this study according to the size of the dedicated pools, s/n. However, because the total staffing level s is fixed in our study, the two analyzes are totally equivalent and the conclusions drawn in terms of n can readily be interpreted in terms of s/n.

2.4.2 Evaluation of Service Rate Percentage Increase

We start from a Pooled System with a given QoS ($W(t)$ or W), and our purpose is to evaluate the required service rate in a Dedicated System with n pools in order to ensure the same QoS ($W_n(t) = W(t)$ or $W_n = W$.) The total staffing level, the total arrival rate of first-attempt calls, and the call back proportion are all held constant.

In the Pooled System, the arrival rate of first-attempt calls is $\lambda^a = 177.36$ calls per min, the call back proportion is $\alpha = 10\%$, the service rate is $\mu = 0.2$ calls per min, and the number of agents is $s = 1000$. In this system, 80% of customers wait no more than 20 sec, and the corresponding average waiting time is $W = 0.18$ min. In the Dedicated System, each call center has a staffing level $s_n = s/n$, an arrival rate of first-attempt calls of $\lambda_n^a = \lambda^a/n$, and a call back proportion of $\alpha_n = \alpha = 10\%$. We vary n from 1 to 50. For each number n of separated call centers, we calculate the service rate μ_n, so that, the average waiting time is $W_n = 0.18$ min. We repeat the same analysis for the QoS in terms of the 80/20 rule. The results are presented in Table 2.1.

Figure 2.4 shows the curves of the required percentage of service rate increase, calculated as $100 \times (\mu_n - \mu)/\mu$, according to the number of pools n in order to reach $W_n = 0.18$ min and $W_n(20sec) = 80\%$, respectively. Figure 2.4 shows that, for both types of QoS, the required increase of service rate does not grow in a dramatic fashion. We notice that for a Dedicated System with $n = 20$ separate teams, the required mean service time is about 4 min and 25 sec in order to reach $W_n = 0.18$ min, and it is about 4 min and 30 sec in order to reach $W_n(20sec) = 80\%$. In a Dedicated System with $n = 10$ separate call centers, the required mean

	$W_n = 0.18$		$W_n(20sec) = 80\%$	
n	μ_n	ρ_n	μ_n	ρ_n
1	0.200	98.53%	0.200	98.53%
2	0.202	97.49%	0.202	97.51%
4	0.206	95.80%	0.205	95.91%
5	0.207	95.07%	0.207	95.24%
8	0.212	93.13%	0.211	93.51%
10	0.214	91.99%	0.213	92.51%
20	0.226	87.30%	0.222	88.57%
25	0.231	85.35%	0.227	86.98%
40	0.245	80.40%	0.237	82.99%
50	0.254	77.60%	0.244	80.76%

Table 2.1: Required service rates in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

service time is only about 4 min and 40 sec for both types of QoS. All these values are not too far from the actual mean service time (5 min.) In conclusion, it is possible to even up the performances of a Pooled System by slightly increasing the service rate. In practice, an increase in service rate efficiency in the order of 10% seems very reasonable to achieve because of the competition created by the team-based organization.

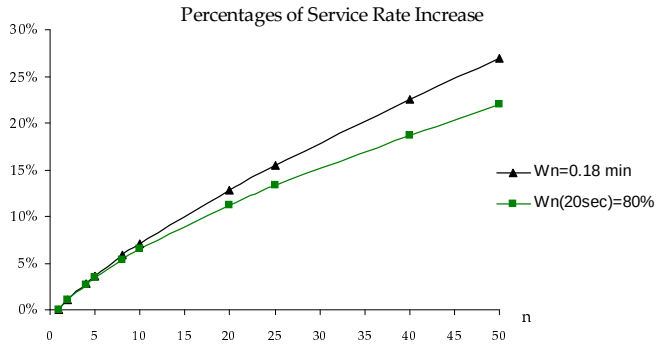

Figure 2.4: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

2.4.3 Evaluation of Percentage of Call Back proportion Decrease

Now, we focus on evaluating the required decrease of the call back proportion in a Dedicated System with n separated call centers, in order to ensure the same QoS ($W_n(t) = W_n(t)$ or $W_n = W$) as in the corresponding pooled configuration.

We consider again the same example of the Pooled System (with the same parameters s, λ^a, α and μ) as in the previous subsection. Each one of the n separated call centers of the Dedicated System has an arrival rate of first-attempt calls of $\lambda_n^a = \lambda^a/n$, a service rate $\mu_n = \mu = 0.2$ calls

per min, and a staffing level $s_n = s/n$. For each n, we calculate the required call back proportion α_n, so that, the average waiting time is $W_n = 0.18$ min. We repeat the same analysis for the QoS in terms of the 80/20 rule. The corresponding results are presented in Table 2.2. We choose to vary n only from 1 to 10, so that, α_n stays positive.

	$W_n = 0.18$		$W_n(20sec) = 80\%$	
n	α_n	ρ_n	α_n	ρ_n
1	10.00%	98.53%	10.00%	98.53%
2	9.02%	97.47%	9.04%	97.49%
4	7.38%	95.74%	7.50%	95.87%
5	6.64%	94.98%	6.83%	95.18%
8	4.61%	92.96%	5.06%	93.40%
10	3.36%	91.76%	4.00%	92.38%

Table 2.2: Required call back proportions in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

In Figure 2.5, we plot the required percentages of the call back proportion decrease, calculated as $100 \times (\alpha - \alpha_n)/\alpha$, versus the number of pools n in order to reach $W_n = 0.18$ min and $W_n(20sec) = 80\%$, respectively. Once again, we see from Figure 2.5 that the required percentage decrease of the call back proportion grows with the number of pools n in a not so strong way, and that the curves for each type of QoS are similar. For example, in a Dedicated System with 10 pools, we have to decrease the call back proportion α_n by about 60% with regard to $\alpha = 10\%$ in order to reach $W_n(20sec) = 80\%$. Note that it is possible to achieve this required decrease in practice, especially when the quality of response within the pooled configuration is quite poor. The reason of the improvement is that the agents in the team-based organization are more responsible for their own customers than in the case of the pooled organization. Agents will try to provide answers that are as good as possible, in order to diminish the call back flow, and as a consequence, improve the performance of their team.

2.4.4 Synthesis

The results of the previous sections have shown that migrating towards separated call centers may not be as bad an idea as it seems. In addition to the analysis reported above, we have performed a more systematic analysis to confirm the robustness of our conclusions. This analysis is reported in Appendix A.1. It shows that the qualitative results discussed above are valid for a large range of parameters typical of those that would be encountered in real situations.

In addition, it would be realistic to assume that the better team management enabled by the new organization implies an improvement of both parameters, i.e., an increase of the service rate and a decrease of the call back proportion. To see the combined improvement of the two factors,

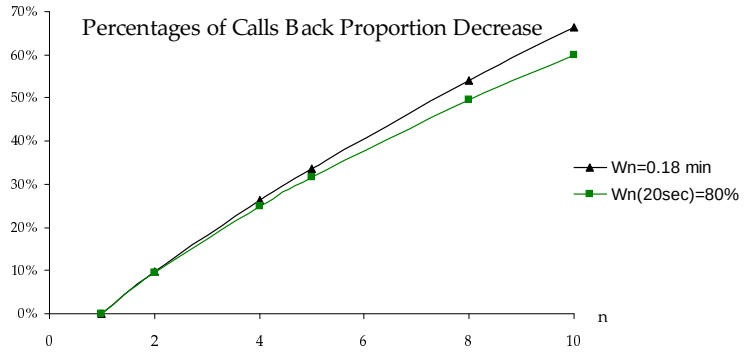

Figure 2.5: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve $W_n = 0.18$ min and $W_n(20sec) = 80\%$

we perform the same analysis as that of Section 2.4.2 by also incorporating an improvement on the call back proportion. We consider two cases corresponding to two values of α: $\alpha = 8\%$ and $\alpha = 5\%$, corresponding to a 20% and 50% improvement in the call back proportion with respect to the initial value of $\alpha = 10\%$, respectively. The results are provided in Tables 2.3 and 2.4 for the range of values of n of interest, i.e., n from 20 to 50. The results show that by having an improvement on both efficiencies, the required performance improvement on each one is not as high as when focusing on each one separately.

	$W_n = 0.18$			$W_n(20sec) = 80\%$		
n	μ_n	% of improvement	ρ_n	μ_n	% of improvement	ρ_n
20	0.221	10.53%	87.20%	0.218	8.90%	88.52%
25	0.226	13.08%	85.24%	0.222	10.90%	86.92%
40	0.240	20.09%	80.26%	0.232	16.23%	82.93%
50	0.249	24.45%	77.45%	0.239	19.44%	80.70%

Table 2.3: Assuming an improvement of the call back proportion by 20% ($\alpha = 8\%$)

	$W_n = 0.18$			$W_n(20sec) = 80\%$		
n	μ_n	% of improvement	ρ_n	μ_n	% of improvement	ρ_n
20	0.214	7.22%	87.06%	0.211	5.55%	88.44%
25	0.219	9.71%	85.08%	0.215	7.50%	86.83%
40	0.233	16.58%	80.07%	0.225	12.69%	82.84%
50	0.242	20.85%	77.24%	0.232	15.81%	80.60%

Table 2.4: Assuming an improvement of the call back proportion by 50% ($\alpha = 5\%$)

Let us now focus on the mix of efficiency improvements for a given value of n. Table 2.5 presents the results pertaining to the case $n = 10$. When we migrate to a Dedicated System with $n = 10$ separated call centers, we need either to increase the service rate μ_n by about 7.11% with

regard to $\mu = 0.2$ call per min, or to decrease the call back proportion α_n by about 66% with regard to the initial $\alpha = 10\%$ in order to achieve $W_n = 0.18$ min. Another solution is to increase μ_n by 3% and decrease α_n by about 37% at the same time. In such a case, it should come as no surprise that we improve the performances in the dedicated systems rather than deteriorate them. Team management effects may change both parameters and may go beyond the simple fact of outweighing the increase of variability.

μ_n Percentage Increase	α_n Percentage Decrease
0.00%	66.43%
1.00%	56.51%
2.00%	46.79%
3.00%	37.26%
4.00%	27.92%
5.00%	18.76%
6.00%	9.78%
7.00%	0.97%
7.11%	0.00%

Table 2.5: Percentages of call back proportion decrease according to percentages of service rate increase in a Dedicated System with $n = 10$ in order to achieve $W_n = 0.18$ min

Figure 2.6 shows the variation of the percentage decrease of α_n according to the percentage increase of μ_n. The graph suggests that improving α_n is linear according to improving μ_n. However it is not the case in general. In fact, let us take the particular case of a Dedicated System with a collection of n separated $M/M/1$ queues ($s_n = 1$.) In this case, the average waiting time is given by

$$W_n = \frac{\rho_n}{\mu_n(1-\rho_n)}, \tag{2.4}$$

where ρ_n is the server utilization, $\rho_n = \lambda_n/\mu_n$. Since $\lambda_n = \lambda_n^a/(1-\alpha_n)$, we deduce from Equation (2.4) that

$$\alpha_n = 1 - \frac{\mu_n W_n + 1}{\mu_n^2 W_n}\,\lambda_n^a. \tag{2.5}$$

If W_n and λ_n^a are held constant, Equation (2.5) shows that α_n is not linear according to μ_n. However, we obtain an almost linear behavior when the number of pools n is not very large and therefore the number of servers per pool s_n is not very small, which is the case in our call center. It is an interesting result, since, if we can assume linearity between these two parameters, we are able to approximate them through simple formulas.

Another advantage of the team-based organization is its robustness with respect to errors in the estimation of the arrival rate of primary calls. For example, consider again the Pooled System, $n = 1$, described in Section 2.4.2 and the Dedicated System, $n = 10$, obtained by increasing the service rate in order to ensure the same QoS as in the Pooled System. We denote

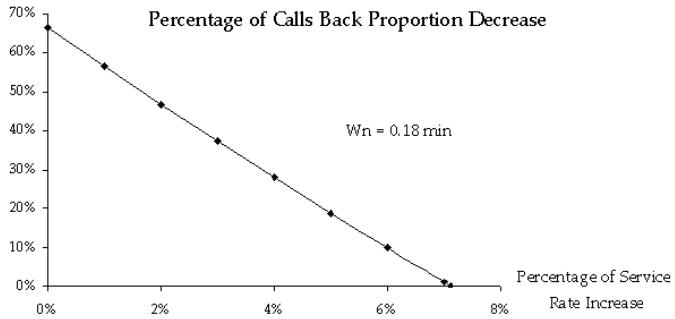

Figure 2.6: Percentages of call back proportion decrease according to percentages of service rate increase in a Dedicated System with $n = 10$ in order to achieve $W_n = 0.18$ min

by $\lambda^{a,real}$ the real first-attempt arrival rate. Figure 2.7 plots the average waiting time W_n versus $\lambda^{a,real}$ for the Pooled System, $n = 1$, and for the Dedicated System, $n = 10$. We observe that the QoS of the Pooled System is much more affected than the one of the Dedicated System by an underestimation of the first-attempt calls arrival rate. Let us give an explanation. Under the original expected first-attempt arrival rate, the server utilization in the Pooled System, 98.53%, is much closer to 1 than the one in the Dedicated System, 91.99%. If the first-attempt calls arrival rate is underestimated, the deterioration of the quality of service is increasing faster when the server utilization is closer to 1, since the queue becomes less and less stable. For example, assume that we underestimate the total arrival rate of first-attempt calls (which is now $\lambda^a = 177.36$ calls per min for both systems) by only 1.41%. Then the real server utilization of the Pooled System becomes 99.92% and the one of the Dedicated System becomes 93.28%. As a consequence, the average waiting time of the first system goes beyond 5 min and the one of the second system is only 0.27 min. This shows that the team-based organization is more robust than the original pooled organization. This is a very attractive feature that gives another strong argument in favor of the team-based organization.

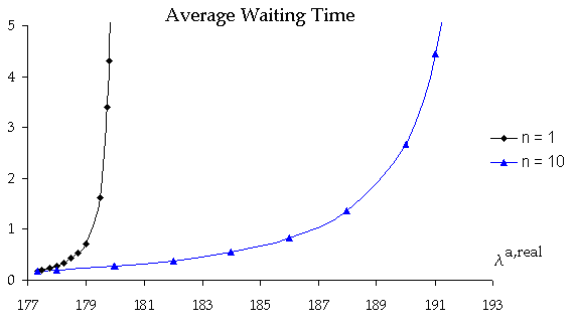

Figure 2.7: Average waiting time of a Pooled System ($n = 1$) and a Dedicated System ($n = 10$) according to the total arrival rate of first-attempt calls

2.5 Call Center Models with Out-Portfolio Flow

In this section, we address the same issues as in Section 2.4 by considering two new models (a pooled model and a dedicated model) of call centers. They differ from the above models by taking an anonymous flow (out-portfolio flow) of calls into account. The latter consists of calls for which one cannot associate a portfolio when they enter the call center. An anonymous call can be a call of a customer of *Bouygues Telecom* who does not communicate his phone number to the Computer-Telephone Integration (CTI), a person who is not a customer of *Bouygues Telecom*, etc. In Section 2.5.1, we present the models and we develop approximations to estimate the performance measures of interest. In Sections 2.5.2 and 2.5.3, we analyze the required improvement in terms of quantitative or qualitative efficiency that must be achieved in order to outweigh the loss in economy of scale. Finally, in Section 2.5.4, we further investigate the consequence of having an out-portfolio flow on the behavior of the Dedicated System.

2.5.1 Modeling and Performance Analysis

Consider first the queueing model of the original large call center with two types of customers: identified customers (portfolio or PTF customers) and non-identified (anonymous) customers (out-portfolio or OPTF customers.) PTF customers have priority over OPTF customers in the sense that agents are providing assistance to PTF customers first. The priority rule is non-preemptive, which simply means that an agent currently serving an OPTF customer while a PTF customer joins the waiting queue will complete this service before turning to the PTF customer. Note that the priority rule in call centers is non-preemptive. It is not common to interrupt the service of a customer to let another one with higher priority start service. The model consists of two infinite queues and a set of s identical servers representing the set of agents. All agents are able to answer all types of customers. Each type of customer has its own queue. Service times are assumed to be exponentially distributed and independent of each other with rate μ for both types of customers. The arrival process of first-attempt calls (primary calls) is assumed to be Poisson with a total arrival rate of λ^a. The proportion of OPTF first-attempt calls is p. So, the total arrival rate of first-attempt PTF calls is $\lambda^{a,PTF} = (1-p)\lambda^a$, and that of the OPTF calls is $\lambda^{a,OPTF} = p\lambda^a$. There is a probability α that the customer is not satisfied with the answer he got and therefore will call again. We assume that α is the same for both types of customers. We make the same detailed assumptions as those presented in Section 2.4.1. Following similar arguments, the behavior of this call center can be approximated by a simple $M/M/C$ queue with two classes of customers (PTF and OPTF), $C = s$ servers, a Poisson arrival rate of PTF customers $\lambda^{PTF} = (1-p)\lambda^a/(1-\alpha)$, a Poisson arrival rate of OPTF customers

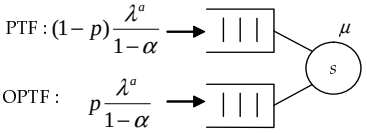

Figure 2.8: Portfolio Pooled System

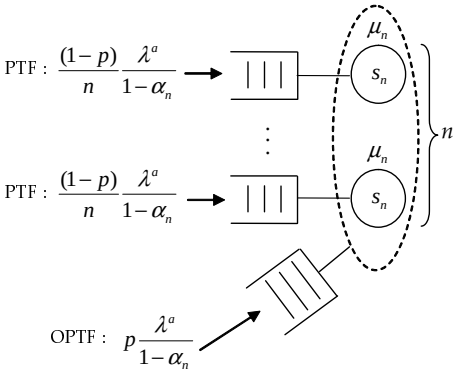

Figure 2.9: Portfolio Dedicated System

$\lambda^{OPTF} = p\lambda^a/(1-\alpha)$ and a service rate μ. PTF customers have non-preemptive (head-of-line) priority over OPTF customers. Within each queue, the discipline is FCFS. This model, referred to as the Portfolio Pooled System, is illustrated on Figure 2.8. Note however that in this more general situation, this model is only an approximation of the behavior of the Portfolio Pooled System. This is due to the fact that the model does no longer belong to the class of product-form networks analyzed by Baskett et al. [22], because of the priority of the PTF customers over the OPTF ones. Note also that the Portfolio Pooled System reduces to the Pooled System studied in Section 2.4 when $p = 0\%$.

Consider now the modeling of the team-based organization with a partitioning of the original call center (with out-portfolio customers) into n autonomous teams, each one being associated with a customer portfolio. It is assumed that n is such that s is a multiple of n. All teams and all customer portfolios are statistically identical. Each team has s_n identical servers, and has its own infinite queue for its own PTF customers. There is a single infinite queue for all OPTF customers. An OPTF customer is served only when at least one of the agents (of any team) is idle and no PTF customers are waiting in the corresponding portfolio queue. The assumptions are similar to those above. The arrival process of PTF first-attempt calls to each PTF queue is Poisson with an arrival rate of $\lambda_n^{a,PTF} = (1-p)\lambda^a/n$. The arrival process of OPTF first-attempt calls to the OPTF queue is Poisson with an arrival rate $\lambda^{a,OPTF} = p\lambda^a$.

The behavior of this call center can be approximated by a set of n identical parallel $M/M/C$ systems with $C = s_n$ servers. Each $M/M/C$ system has its own arrival process corresponding to its PTF customers. This arrival process is Poisson with a rate $\lambda_n^{PTF} = \lambda_n^{a,PTF}/(1-\alpha_n) = ((1-p)/n)(\lambda^a/(1-\alpha_n))$. In addition, there is an additional queue for the OPTF customers. The arrival process to this queue is Poisson with a rate $\lambda^{OPTF} = \lambda^{a,OPTF}/(1-\alpha_n) = p(\lambda^a/(1-\alpha_n))$. The OPTF customers can be served by any server of any of the parallel $M/M/C$ queues. However, PTF customers have non-preemptive (head-of-line) priority over OPTF customers. The resulting

model is shown on Figure 2.9. This model will be referred to as the Portfolio Dedicated System. Note that the Portfolio Dedicated System reduces to the Portfolio Pooled System when $n = 1$, and to the Dedicated System when $p = 0\%$.

We now focus on evaluating the stationary performances of the two above models. Exact performance measures of the Portfolio Pooled System (viewed as a non-preemptive priority $M/M/C$ queue) in terms of the waiting time distribution can be found in Kella and Yechiali [78]. However, the exact quantitative analysis of the Portfolio Dedicated System is complicated. For that, we developed a set of models which allow us to approximate its performance measures.

In the following, we present two approximations for the Portfolio Dedicated System. The first enables us to calculate a pessimistic estimate (upper bound) of the waiting times of the PTF customers, while the second one enables us to calculate a pessimistic estimate (upper bound) of the waiting times of the OPTF customers. The reason for choosing pessimistic estimates is such that the improvements in efficiency that will follow from our analysis can be viewed as a lower bound on the improvements in efficiency that will actually be required in the exact analysis. These approximate models are of the same nature as that of the Portfolio Pooled Model. Their performance measures can then be exactly calculated in the same way. Validations of these approximations are presented in Appendix A.2.

Pessimistic Model for PTF customers The pessimistic model for PTF customers is obtained from the Portfolio Dedicated model by splitting the flow of OPTF into a set of n independent flows. The resulting model consists of a set of n independent and identical $M/M/C$ systems with $C = s_n$ servers and a service rate μ_n. As in the original model, the arrival process of PTF customers is a Poisson process with rate $\lambda_n^{PTF} = ((1-p)/n)(\lambda^a/(1-\alpha_n))$. The arrival process of OPTF customers is a Poisson process with rate $\lambda_n^{PTF} = (p/n)(\lambda^a/(1-\alpha_n))$. The OPTF customers can be served by any of the s_n servers from the corresponding team. However, PTF customers have non-preemptive (head-of-line) priority over OPTF customers. The model is shown on Figure 2.10.

In this model, there is not a single OPTF queue as opposed to the Portfolio Dedicated System. The OPTF flow is equally divided and assigned to each one of the separate OPTF queues. Thus, it may happen that an OPTF customer is delayed to access a server, while a server of another team is available. This delay may later on delay the access of a PTF customer to a server because the OPTF customer will now be served and the priority is non-preemptive. Another way to look at this approximate model is to see it as an unpooling of the OPTF flow, which indirectly causes additional delays to the PTF customers. Thus this model provides a pessimistic estimate of the waiting times of the PTF customers.

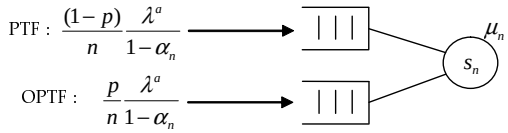

Figure 2.10: Pessimistic model for the PTF customers

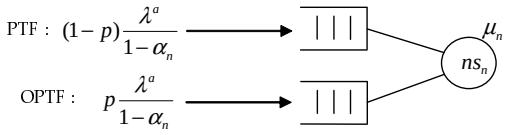

Figure 2.11: Pessimistic model for the OPTF customers

Pessimistic Model for OPTF customers The pessimistic model for OPTF customers is obtained from the Portfolio Dedicated System by merging the flow of PTF into a single flow and simultaneously merging all the servers into a single pool of ns_n servers. This model is similar to the Portfolio Pooled System. The arrival process to the single PTF queue is Poisson with rate $\lambda_n^{PTF} = (1-p)(\lambda^a/(1-\alpha_n))$. The arrival process to the single OPTF queue is a Poisson process with rate $\lambda_n^{PTF} = p(\lambda^a/(1-\alpha_n))$. PTF customers have non-preemptive priorities over OPTF customers. The model is shown on Figure 2.11.

Because of the pooling effect of the PTF customers, the average waiting time of OPTF customers will be higher. In the Portfolio Dedicated System, it may happen that an OPTF customer gets access to a server while a PTF customer is waiting in a queue (not served by this server), which reduces the waiting time of the OPTF customer. In the pessimistic model, this will never occur due to the pooling of all PTF queues and all servers. Thus this model provides a pessimistic estimate of the waiting time of OPTF customers.

In the following subsections, we evaluate the impact of migrating from a Portfolio Pooled System towards a Portfolio Dedicated System. We concentrate on the evaluation of efficiency improvement (both qualitative and quantitative) required to counterbalance the reduction of the pooling effect of the team-based organization. We define a global quality of service QoS^{global} for all types of customers as $QoS^{global} = (1-p)\, QoS^{PTF} + p\, QoS^{OPTF}$. In what follows, we only focus on the QoS measured in terms of the average waiting time. Similar results could be obtained for the 80/20 rule.

2.5.2 Evaluation of Service Rate Percentage Increase

We start from a Portfolio Pooled System with a given quality of service W^{global}, and our purpose is to evaluate the required service rate μ_n in a Portfolio Dedicated System with n identical teams in order to ensure the same global average waiting time $W_n^{global} = W^{global}$. The total staffing level, the total arrival rate of first-attempt calls, and the call back proportion are all held constant.

In the Portfolio Pooled System, the arrival rate of first-attempt calls is $\lambda^a = 177.36$ calls per

min, the call back proportion is $\alpha = 10\%$, the service rate is $\mu = 0.2$ calls per min, and the number of servers is $s = 1000$. The server utilization is then $\rho = 98.53\%$. We choose the same parameters as in the Pooled System that we studied in Section 2.4. In addition, we vary the proportion of OPTF customers, $p = 0\%$, $p = 5\%$, $p = 10\%$, or $p = 20\%$. As expected, W^{global} does not depend on p. In fact, the Portfolio Pooled System is a workconserving system. It is not the case that one server is idle while a customer (PTF or OPTF) is waiting for service. If we vary p, the order of service of a given customer may change, but the overall average waiting time remains unchanged. We give the mathematical explanation in Appendix A.3. More discussions about this result for more general models will be addressed in Chapters 3 and 6. Here, $W^{global} = 0.18$ min for all values of p. For each value of p, we now consider the corresponding Portfolio Dedicated System. We vary n from 1 to 50. For each n, we evaluate the required service rate μ_n (using pessimistic models), so that, the global average waiting time is $W_n^{global} = W^{global} = 0.18$ min. We present the results in Table 2.6.

	$p = 0\%$		$p = 5\%$		$p = 10\%$		$p = 20\%$	
n	μ_n	ρ_n	μ_n	ρ_n	μ_n	ρ_n	μ_n	ρ_n
1	0.2	98.53%	0.2	98.53%	0.2	98.53%	0.2	98.53%
2	0.202	97.49%	0.201	98.23%	0.2	98.36%	0.2	98.39%
4	0.206	95.80%	0.202	97.38%	0.201	97.95%	0.201	98.23%
5	0.207	95.07%	0.203	96.84%	0.202	97.67%	0.201	98.13%
8	0.212	93.13%	0.207	95.08%	0.204	96.54%	0.202	97.74%
10	0.214	91.99%	0.21	93.92%	0.206	95.60%	0.202	97.39%
20	0.226	87.30%	0.221	89.06%	0.217	90.80%	0.209	94.16%
25	0.231	85.35%	0.226	87.05%	0.222	88.74%	0.214	92.13%
40	0.245	80.40%	0.24	81.96%	0.236	83.54%	0.227	86.81%
50	0.254	77.60%	0.249	79.09%	0.244	80.62%	0.235	83.81%

Table 2.6: Required service rate increase in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

Moreover, we calculate by simulation the exact values of μ_{20} for $p = 5\%$, 10% and 20%, in order to get some indications of their deviations regarding the values given by the proposed models. Recall that for $p = 0\%$, the Portfolio Dedicated System behaves like separate Erlang-C models. So, the corresponding μ_{20} is obtained by an exact numerical result. For $p = 5\%$, the required service rate given by simulation is $\mu_{20} = 0.217$ instead of 0.221 given by our approximation models, for $p = 10\%$ it is 0.210 instead of 0.217, and for $p = 20\%$ it is 0.202 instead of 0.209. This shows us that the costs of partitioning given by our models are not too far from those given by simulation, at least for the reasonable parameters of the Portfolio Dedicated System we have chosen. In Figure 2.12, we plot for each value of p, the curve of the percentage of required service rate increase, calculated as $100 \times (\mu_n - \mu)/\mu$, in the Portfolio Dedicated System according to

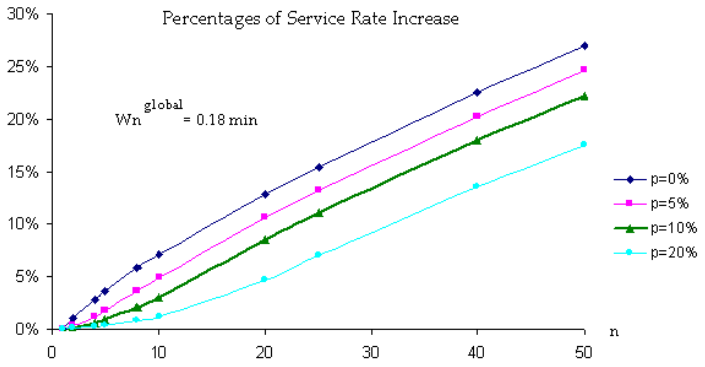

Figure 2.12: Percentages of required service rate increase according to number of pools n in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

the number of pools n, in order to reach a global quality of service of $W_n^{global} = 0.18$ min. We notice that for a given p, we have the same qualitative results as in Section 2.4.2. The required increase of the service rate is not very important and it is feasible to reach in practice. This is due to the competition element of the team-based organization.

The new interesting insight here is that the necessary increase of the service rate to compensate the loss of pooling effect decreases when the proportion of OPTF customers increases. Particulary, migrating towards a Portfolio Dedicated System (with any $p > 0\%$) is always less costly than migrating towards a Dedicated System ($p = 0\%$). For example, consider a Portfolio Dedicated System with $n = 10$. If the OPTF proportion is $p = 5\%$, we need to increase the service rate by 4.91%. However, with a proportion $p = 20\%$ we only need to increase the service rate by 1.17%. We explain this advantage by the fact that the OPTF flow is used to reduce idle periods of servers while customers are waiting in the Portfolio Dedicated System. Idle times would not exist in the case of $p = 100\%$ (Pooled System.) This will be explained with more details in Section 2.5.4.

2.5.3 Evaluation of Percentage of Call Back proportion Decrease

Let us again start from the Portfolio Pooled System of Section 2.5.2. We aim to evaluate the call back proportion α_n for a Portfolio Dedicated System with n identical teams, in order to get $W_n^{global} = W^{global} = 0.18$ min as in the Portfolio Pooled System. We again vary p ($p = 0\%$, $p = 5\%$, $p = 10\%$, or $p = 20\%$.) For each p, we vary n from 1 to 10. We choose to vary n only from 1 to 10, so that, α_n stays positive. We present the results in Table 2.7.

In Figure 2.13, we plot for each value of p, the curve of the required percentage of call back proportion decrease, calculated as $100 \times (\alpha - \alpha_n)/\alpha$, in the Portfolio Dedicated System according to n, in order to reach a global quality of service of $W_n^{global} = 0.18$ min. Again, we get the same

	$p = 0\%$		$p = 5\%$		$p = 10\%$		$p = 20\%$	
n	α_n	ρ_n	α_n	ρ_n	α_n	ρ_n	α_n	ρ_n
1	10.00%	98.53%	10.00%	98.53%	10.00%	98.53%	10.00%	98.53%
2	9.02%	97.47%	9.72%	98.23%	9.86%	98.37%	9.93%	98.46%
4	7.38%	95.74%	8.91%	97.35%	9.46%	97.94%	9.77%	98.28%
5	6.64%	94.98%	8.39%	96.80%	9.20%	97.66%	9.67%	98.17%
8	4.61%	92.96%	6.61%	94.95%	8.09%	96.48%	9.29%	97.76%
10	3.36%	91.76%	5.39%	93.73%	7.12%	95.47%	8.94%	97.39%

Table 2.7: Required call back proportion decrease in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

qualitative results as in Section 2.4.3. In addition, we notice that the cost in terms of the required decrease of call back proportion is decreasing according to p. It is due again to the OPTF flow. When p increases, the variability in the Portfolio Dedicated System decreases. For instance, consider a Portfolio Dedicated System with $n = 10$. If the OPTF proportion is $p = 5\%$, we need to decrease the call back proportion by 46.07%. However, with a proportion of $p = 20\%$ we need to decrease the call back proportion by only 10.57%.

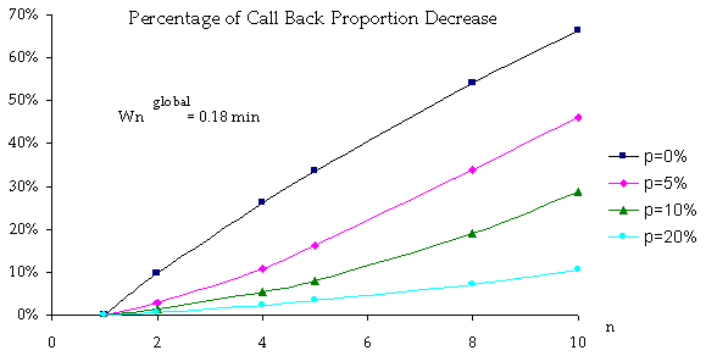

Figure 2.13: Percentages of call back proportion decrease according to number of pools n in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

2.5.4 Synthesis

In the previous sections, we showed that the reduction of pooling effect when migrating from a Pooled System to a Dedicated System could be outweighed by team management benefits. The discussion and insights presented in Section 2.4.4 are still valid in the more general setting of a call center with an out-portfolio flow. However, there is a new important insight. It clearly appears that having an out-portfolio flow may reduce the drawback of migrating from the Pooled System to the Dedicated System. This is due to the fact that as opposed to the PTF flows, the OPTF flow is not decomposed into several independent flows, each one being associated with a

specific team. Thus the OPTF flow maintains the benefits of the pooling effect. The out-portfolio flow can then be seen as an "idle time killer" in a Portfolio Dedicated System: an out-portfolio call is distributed only when an agent is idle and no customer of his portfolio is waiting. As out-portfolio calls have less priority, this allows reducing idle time without significantly penalizing the QoS of portfolio calls.

The Portfolio Dedicated System can thus be considered as a particular case of partial pooling. We call this configuration "partial calls pooling" because a proportion p of incoming calls (pooled calls or out-portfolio calls in the *Bouygues Telecom* case) can be served by any agent while the remaining calls $(1-p)$ are dedicated to specific agents. It clearly appears from the graphs presented in Sections 2.5.2 and 2.5.3 that this improvement in efficiency required to counterbalance the reduction of pooling effect decreases as p increases. Now, one additional very attractive feature is that this decrease is not linear in p. Let us, for instance, consider the case of quantitative efficiency improvement (service rate increase) discussed in Section 2.5.2 (a similar analysis could be done for the qualitative efficiency improvement discussed in Section 2.5.3.)

In order to illustrate this behavior, let us again consider the same basic example of the Portfolio Pooled System (1000 servers, $\lambda^a = 177.36$ calls per minute, $\mu = 0.2$ calls per min, and 10% of call back proportion.) In Figure 2.14, we plot the required percentage of service rate increase in the Portfolio Dedicated System to reach the same performance as the Portfolio Pooled System, as a function of the proportion of out-portfolio flow p (p ranges from 0% to 100%.) There are three graphs corresponding to three Portfolio Dedicated System configurations: $n = 10$, 20, and 40. The graphs confirm the non-linear shape of the curve. This means that with a fairly small percentage of out-portfolio flow, the required efficiency improvement is much smaller than that of the system without out-portfolio flow (corresponding to the case where $p = 0\%$.) In other words, a rather small out-portfolio flow significantly reduces the drawback of the unpooling effect of the PTF customers. Recall also that since we are using pessimistic approximations, the actual curves would be stiffer.

For this case, we also performed an extensive numerical study to validate that the conclusions discussed above remain valid for a large set of parameters, thereby confirming the robustness of our analysis.

2.6 Conclusions and Perspectives

We focused on a fundamental problem in the design and management of stochastic service systems. We investigated the impact of team-based organizations in call centers management. Agents of call centers are the interface between the company and the customers. Thus, man-

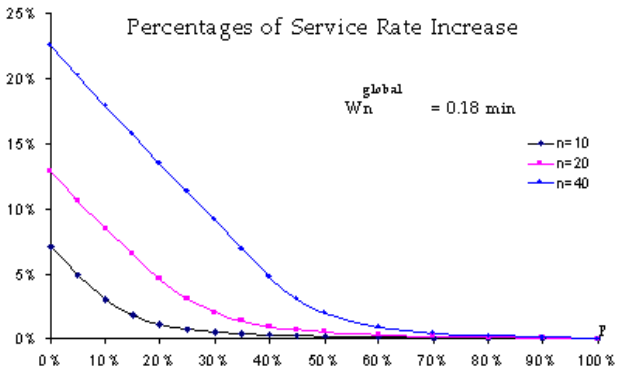

Figure 2.14: Percentages of required service rate increase according to OPTF proportion p in a Portfolio Dedicated System in order to achieve $W_n^{global} = 0.18$ min

agers have to support and motivate their employees, so that, the assistance they provide to the customers is efficient. Partitioning agents into groups creates competition and makes agents more responsible, which motivates them to provide both rapid and improved responses.

In this chapter, we argued how team management benefits, that come from the portfolio/team one-to-one link, may outweigh the economy of scale associated with the pooled organization. First, we study partitioning of a large call center into identical and separated call centers, where agents of a same team are dedicated to one portfolio of customers. Queueing models involved in this part of the study are simple. They give us important insights and help us understand the behavior of more complicated systems. We show that the costs of migrating towards a Dedicated System are not as important as it may appear. In practice, combining the benefits of the team-based organization in terms of both improved service rate efficiency and reduced call back proportion can easily outweigh the loss of the economy of scale. We also present further insights, such as robustness of the Dedicated System regarding errors in the estimation of the arrival rate.

In the second part of the chapter, we extend our analysis to the more general situation with an additional out-portfolio flow. We develop a set of models that give us lower bounds of performance measures. We verify the same qualitative results as in the first part. In addition, we present an interesting insight, that is, a small proportion of out-portfolio calls may be sufficient to approximately attain the same performances as in the Pooled System. This property fits into a general idea in queueing theory. It is like saying that with a small amount of flexibility, an SBR call center may yield most of the benefits of full-flexibility (Chevalier et al. [34].)

The application of customer portfolio management had very significant effects in the *Bouygues Telecom* call center. The quality of answers has been improved reducing callbacks by 25%. The proportion of disconnected calls (because of a full queue) was divided by 2 (in our work we

assumed an infinite queue for simplicity.) In addition, no supplementary agents were hired in spite of the increase of the total number of customers by 15%. This provides an experimental confirmation of the results and insights presented in this chapter.

In a future study, it would be interesting to extend our models by considering abandonments and limited waiting lines and more general service time distributions. We will also try to improve the approximation models discussed here to get more accurate analyzes. Finally, a more ambitious extension would be to investigate the introduction of team-based organization in an SBR call center where agents have specific skills.

Chapter 3

Real-Time Scheduling Policies for Multiclass Call Centers

In this chapter, we address an issue related to the real-time management of call centers. We consider a call center model with two classes of impatient customers, VIP and less important ones. We focus on developing scheduling policies that assign customers upon arrival to parallel queues, high and low priority queues. The policies are developed subject to satisfying constraints on performances related to the ratio of the probabilities of being lost, as well as the variance of waiting times in queue. We propose and compare several real-time scheduling policies in order to reach our objective. The policies are characterized to be simple and easy to implement in practice.

An extended version of this chapter is the working paper Jouini, Pot, Dallery and Koole [70].

3.1 Introduction

This chapter deals with a real-time problem of call centers, namely customer routing and server scheduling. The issue of this chapter is conceptually different of that analyzed in Chapter 2 in the sense that we are not focusing on a design problem. Given a system structure and staffing level, our purpose is to develop routing schemes for arriving calls subject to satisfying some given quality of service constraints.

As in Chapter 2, our concern here is a full-flexible call center. We assume that all agents are flexible enough (polyvalent) to answer all requirements of service. However, we divide customers into two different classes according to their importance, VIP and ordinary customers. The resulting model for our call center here has a V-design according to the canonical designs presented in Garnett and Mandelbaum [43]. In addition, we let customers to be impatient. Incorporating reneging in theoretical models is of value. In reality, it is natural that a waiting customer will wait for only a limited time, and will hang up within that time. Ignoring reneging leads to overstaffing and pessimistic estimation of queueing delays. Garnett et al. [44] show using numerical examples that models with and without abandonment tend to give very different performance measures even if the abandonment rate is small. Models including reneging are therefore more close to reality, and necessary to obtain more accurate managerial insights.

In this chapter, we discuss various real-time (online) scheduling policies subject to satisfying differentiated service levels related to the probabilities of being lost. The target consists on a balance between the achieved service levels of customer classes. This objective is motivated by a situation that often occurs in practice due to the uncertain environment of call centers. It happens when the workload prediction step is incorrectly done. Several studies as in Jongbloed and Koole [63] and Avramidis et al. [18] have shown that the arrival process and the workload are hard to predict in call centers. We show in this chapter that using online scheduling policies is an effective way to compensate workforce for mismatches between different customer classes, and meet targets on service levels. The main advantage of our control policies is that they require no information about the arrival processes in advance. If fluctuations in workload occur, the policies are adapted such that targets on the service levels are met.

In general, the provision of differentiated service levels relies on the use of priority queues. Schrage and Miller [117] have shown that scheduling policies similar to multiclass priority queues allow to achieve high performances, often nearly as good as those under optimal policies. Note, in addition, that the priority schemes are easy to implement, which explain their prevalence in practice. However, it is well known that fixed strict priority policies result either in satisfying target performances for lower priority customers and an over service level for higher priority

ones, or in satisfying the target for high priority customers while having heavily penalized lower priority customers.

Our purpose here is to develop simple and useful dynamic routing policies that are based on priority schemes. On the one hand, optimal routing policies are very complex to obtain because of the mathematical difficulties. On the other hand, they are usually not interesting to implement in practice due to their several requirements about real-time system information, see Koole and Pot [88]. Hence, it might be more beneficial to use simple scheduling policies instead of attempting to use optimal ones. We derive various schemes for dynamic assignment of customers to queues in order to meet our target. The policies we propose are characterized to be workconserving (non-idling.) A policy is defined to be workconserving if there can be no idling servers when there are waiting customers, which is natural for large service systems such as call centers. Through the analysis below, we also focus our interest on the achieved variance of the waiting time in order to differentiate the proposed policies. We do not focus on the achieved first moment of the waiting time. In fact, our proposed policies allow to achieve not very different values of mean waiting times. We thereafter prefer a system with low waiting time variance than a system that is faster on average but highly variable. A further advantage of minimizing the variance of the waiting time is related to the announcement of anticipated delays to customers. Computing the full distribution of the state-dependent waiting time is a very complex task. So, we aim to minimize its variance such that announcing the state-dependent mean waiting time will not be a bad prediction. Further details on call centers with delays announcement will be addressed in Chapter 4.

The interesting side of the policies we suggest comes from their simplicity, they do not require information about the workload process. The analysis yields to quantitative insights, as well as useful principles and guidelines for the control problem. Note that non-workconserving policies, such as thresholds or reservations of agents for important customers, are not considered in this chapter. In our opinion, the restriction does not decrease the usefulness of the analysis because we conjecture that the larger the call centers, the more effective are workconserving policies, see Pot [112]. A further reason is that we want to prevent our analysis from being too complicated and therefore not interesting for practitioners.

Here is how the rest of the chapter is structured. In Section 3.2, we review two kinds of literature close to our work. The first one deals with queueing models incorporating reneging, and the second deals with results about optimal scheduling policies. In Section 3.3, we give a comprehensive presentation of the problem we study in this chapter: Section 3.3.1 is devoted to formulate the queueing model of the call center, Section 3.3.2 gives some preliminary results, and

Section 3.3.3 concretely expresses our objective. In Section 3.4, we develop dynamic scheduling policies that allow to meet our objective. In Section 3.5, we present and discuss simulation results of the proposed policies. In Section 3.6, we investigate some extensions. In Section 3.6.1, we focus on extending our analysis to the case of three customer classes. The main analysis in this chapter focuses on a period of the day, where the system parameters are assumed to be stationary. From a practical side, it would be interesting to extend the analysis to the whole day. In Section 3.6.2, we investigate the conservation of the proportionality between the service levels for a call center working day. In Section 3.7, we give some concluding remarks and future research directions.

3.2 Literature Review

The literature related to this chapter spans into two main areas. The first deals with queueing systems with impatient customers. The second area deals with the control of queueing systems, specifically, the problem of customer routing and server scheduling.

In the following, we highlight some of the literature with regard to the first area. Queueing models incorporating impatient customers have received a lot of attention in the literature. Gross and Harris [47] define the impatience through three different forms. The first is balking, that is, the reluctance of a customer to join a queue upon arrival. The second is reneging, which means the reluctance to remain in queue after joining and waiting. Finally, the third is jockeying between separate queues. Jockeying means that one customer has the possibility to change to one queue while he was waiting in another. In this thesis, we consider the second form of impatience. The other forms are not allowed. To underline the importance of the abandonment modeling in the call center field, the authors in Gans et al. [40] and in Mandelbaum and Zeltyn [98] gave some numerical examples that point out the effect of abandonment on performances. The literature on queueing models with reneging focus especially on performance evaluation. We refer the reader to Ancker and Gafarian [10], Garnett et al. [44], and references therein for simple models assuming exponential reneging times. In Garnett et al. [44], the authors study the subject of Markovian abandonments. They suggest an asymptotic analysis of their model under the heavy-traffic regime. Their main result is to characterize the relation between the number of agents, the offered load and system performances such as the probability of delay and the probability to abandon. This can be seen as an extension of the results of Halfin and Whitt [52] by adding abandonments. Zohar et al. [146] investigate in their work the relation between customers reneging and the experience of waiting in queue. Other papers have allowed reneging to follow a general distribution. Related studies include those by Baccelli and Hebuterne [19],

Brandt and Brandt [31], Ward and Glynn [130], Pla et al. [111] and Zeltyn and Mandelbaum [145]. Moreover, one should mention results about monotonicity and convexity properties. These results are especially relevant for practical guidelines, as well as for obtaining useful structures of control policies. Some related literature to this subject is given in Section 6.1 of Chapter 6.

Let us now focus on the second area of literature closed to our work, that is, the control of queueing systems. Scheduling policies has been studied in great depth within the context of queueing systems. A scheduling policy, or a discipline of service, prescribes the order in which customers are served. It is tied to identifiable characteristics of customers. Arrival time is certainly one of these characteristics, it is the basis for the most familiar disciplines as the first come, first served (FCFS) discipline and the last come, first served (LCFS) discipline. Several other characteristics are the bases for queue disciplines. Customers may be processed according to service times, which may lead to the well known Shortest Remaining Processing Time discipline (SRPT.) The queue discipline may also be based on the customer type, for example, VIP customers are scheduled first. Or the queue discipline may be an hybrid strategy that accounts for more than one characteristic. The focus in control problems is on determining the form of the optimal policy so as to optimize system performance. Randolph [113] classifies scheduling policies into those using dynamic schedule rules and those using static schedule rules. A dynamic schedule is a discipline that is continuously updated as customers arrive and are processed. However, a static schedule is state of system independent, it is beforehand defined and never altered. Each one of the above classes of policies can be further classified into two major types: agent scheduling and customer routing. As defined in Garnett and Mandelbaum [43], agent scheduling is described by a control decision taken whenever an agent turns idle and there are queued customers: which customer, if any, should be routed to this agent. However, a scheduling policy based on a customer routing rule, is defined by a control decision taken whenever a customer arrives: which idle agent, if any, should serve this customer, if not, to which queue should the customer of interest be routed.

In this chapter, under some given performance constraints, we develop dynamic scheduling policies based on the second type of control decision, i.e., the assignment rule of new arrivals to queues. Our basic call center model has a V-design, with two infinite queues. A high priority queue and a lower priority one. We focus on policies that assign a priority level to customers upon arrival. We should note that there are two further possible refinements in priority situations, namely preemption and non-preemption. In preemptive cases, a customer with high priority is allowed to enter service immediately even if another one with lower priority is already present in service at its arrival epoch. However, a priority discipline is said to be non-preemptive if there is

no interruption. A customer with higher priority just goes to the head of the queue and wait his turn. The scheduling policies analyzed in this work are characterized to follow a non-preemptive priority schemes. In addition, we focus on policies that are workconserving, that is, we do not allow agents to be idle while there are waiting customers.

In the following, we present some known results about optimal scheduling policies. Note that due to the complexity of such studies, most of the existing research considers simple queueing models. Moreover, the literature illustrates that optimal results are difficult to obtain. The structure of optimal policies are model dependent and often difficult to generalize to more complicated cases. Schrage and Miller [117] proves that the SRPT policy, which schedules in a preemptive manner the customer with the smallest remaining processing time at every point in time, is optimal with respect to minimizing mean sojourn times in an $M/G/1$ system. Pekoz [108] addresses the analysis of a multiserver non-preemptive priority queue with exponentially distributed interarrival and service times. He finds and evaluates the performance of an asymptotically optimal policy that minimizes the expected queueing delay for high priority customers. Guérin [48] presents a model without waiting queues. It contains a multi-server station, which receives low and high priority arrivals. He develops an admission policy for the low priority customers such that the fraction of blocked high priority customers is bounded and he analyzes the system under that policy. Örmeci [105] considers a Markovian call center model with two classes of customers, one pool of generalists, two pools of specialists, and no waiting rooms. Then, the author derives the structure of dynamic admission policies that maximizes, in the long-run, the total expected discounted revenue. Aguir et al. [7] present an optimization problem for an $M/M/1$ queue with two classes of customers. They prove and characterize a class of optimal static scheduling policies subject to satisfying differentiated performances for customer classes. The proposed policies are based on strict priority rules, and the performances are measured in terms of the mean waiting time and the 80/20 rule. The authors investigate extensions to the multiserver queue, and analyze a dual class of static policies based on agent scheduling. We also refer the readers to Xu et al. [142], Huang [58], Bhulai and Koole [27], and Gans and Zhou [42] for more results. Other papers, related to scheduling problem for multiserver systems under the asymptotic heavy-traffic regime, include those by Gans and van Ryzin [41], Bassamboo et al. [23], Atar et al. [16], Armony [11] and references therein.

3.3 Framework

In this section, we first describe the basic model of our call center. Second, we present notations related to the performances we consider in this chapter, and develop some preliminary results.

Finally based on the preliminary study, we concretely specify our motivation and objective with regard to the scheduling policies we aim to develop.

3.3.1 Model Formulation

We model our call center as a queueing model with two classes of customers; important customers type A, and less important ones type B. The model consists of two infinite queues type 1 and 2, and a set of s identical servers representing the set of agents. All agents are able to answer all types of customers. The call center is operated in such a way that at any time, any call can be addressed by any agent. So upon arrival, a call is addressed by one of the available agents, if any. If not, the call must join one of the queues. Customers are assigned to queues according to the selected scheduling policy, as we shall detail later. Customers in queue 1 have priority over customers in queue 2 in the sense that agents are providing assistance to customers belonging to queue 1 first. The priority rule is non-preemptive (see the motivation in Section 2.5.1 of Chapter 2), which simply means that an agent currently serving a customer pulled from queue 2, while a new arrival customer joins queue 1, will complete this service before turning to queue 1 customer. Within each queue, customers are served in order of their arrivals, i.e., under the FCFS discipline. Interarrival times and service times are assumed to be i.i.d., and follow general distributions. In certain cases, we shall in particular consider the exponential distribution for successive service times. Then, a customer is served with rate μ, independent of the customer type.

In addition, we assume that the customers are impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time he will renege (leaves the queue.) Times before reneging, for type A and B customers, are assumed to be i.i.d. and exponentially distributed with rate γ. Assuming identical distribution of patience within each class, independently from their position in queue, seems to be a plausible assumption for call centers, see Gans et al. [40]. Indeed, the tele-queueing experience in call centers is fundamentally different from that of a physical queue, in the sense that customers do not see others waiting and need not be aware of their “progress" (position in the queue) if the call center does not provide information about queueing delays. The system is workconserving, i.e., a server is never forced to be idle with customers waiting. Finally, retrials are ignored, and reneging is not allowed once one customer starts his service. We do not allow also jockeying between separate queues. Following similar arguments, the behavior of this call center can be viewed as a variety of a $GI/GI/s + M$ queueing system. The symbol M after the $+$ is to indicate the Markovian assumption for times before reneging. Note that owing to abandonments,

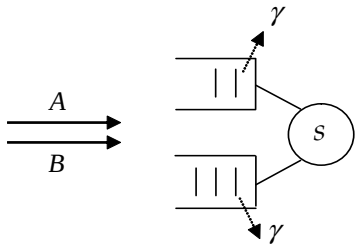

Figure 3.1: The basic model

the system is unconditionally ergodic. The resulting model is shown on Figure 3.1.

In this chapter, the main results concerns the case when service times, as well as times before reneging, are identically distributed for both customer types. There are two reasons for that. The first reason is related to the nature of the call center we consider here, and which is the case in many other call centers applications. In fact, the difference making a type A customer more important for the company than a type B customer is that the first pays more money than the second one. In concrete terms, if our call center owns every month from one customer an amount of money crossing a given threshold, then that customer is of type A, else he is of type B. We consider that the queries asked by customers, as well as the patience experiences do not differ from one type of customers to another. Therefore, it would be credible to assume a common distribution. The second reason is due to the complexity of the analysis when assuming different service and reneging time distributions. Our objective here is to investigate simple models that allow us to better understand the system behavior and to gain general useful guidelines and insights.

Finally, note that assuming exponential distribution for service times and times before reneging is not that bad approximation. It is true that in real call centers cases, impatience times need not be exponential, and they can vary significantly with the type of service, the information provided during waiting, etc. We refer the reader to Zohar et al. [146] and Mandelbaum et al. [96] who show how such assumptions may be violated. However, our model is still of interest in practice as mentioned by Whitt [139], and Pierson and Whitt [109]. The authors have shown, using various simulation experiments, that the $M/M/s+M$ (Erlang-A) model provides a good approximation of the $M/GI/s+GI$ model.

3.3.2 Preliminaries

In this section, we first present notations and definitions about the performances we are interested on. The performances are defined in terms of the fraction of customers who abandon, the mean and the variance of the waiting time in queue. Second, we develop some structural results related

to these performances.

We denote by m the type of one customer, $m \in \{A, B\}$. We assume that at time $t = 0$, the system starts empty. Under some given scheduling policy π, let $n^m(t)$ be the number of type m arrivals during the interval of time $[0, t]$, $t > 0$. Let $a_\pi^m(t)$ be the number of type m customers who abandon the queue, and $b_\pi^m(t)$ the number of those who get service. We define a first service level in terms of the fraction of customers who abandon within each type, as well as for all types of customers. The fraction of type m customers who abandon, say $Q_\pi^m(t)$, during $[0, t]$ is defined by

$$Q_\pi^m(t) = \frac{a_\pi^m(t)}{n^m(t)}. \tag{3.1}$$

As for the overall service level, it is defined by

$$Q_\pi(t) = \frac{a_\pi^A(t) + a_\pi^B(t)}{n_\pi^A(t) + n_\pi^B(t)}. \tag{3.2}$$

During the stationary regime, the service level for type m customers, say Q_π^m, and the overall service level for all types, say Q_π, are given by

$$Q_\pi^m = \lim_{t\to\infty} Q_\pi^m(t), \text{ and } Q_\pi = \lim_{t\to\infty} Q_\pi(t). \tag{3.3}$$

Recall that due to abandonments, the system is ergodic so that the latter limits exist. Let us now define the ratio, c_π, of the stationary service level of customers A over that of customers B. It is given by

$$c_\pi = \frac{Q_\pi^A}{Q_\pi^B}. \tag{3.4}$$

Similarly, we define the mean waiting time in queue for class m and the overall mean waiting time in queue for all customer types. Note that we only define these quantities for the customers who enter service. Under a given scheduling policy π, let $w_{q,\pi}^m(i, t)$ be the waiting time in queue of the i^{th} type m customer who enters service, $0 \le i \le b_\pi^m(t)$. As in the usual way, the mean waiting time in queue $W_{q,\pi}^m(t)$ for type m customers during $[0, t]$ is defined by

$$W_{q,\pi}^m(t) = \frac{1}{b_\pi^m(t)} \sum_{i=1}^{b_\pi^m(t)} w_{q,\pi}^m(i, t). \tag{3.5}$$

Also, we define the overall mean waiting time in queue, during the interval $[0, t]$ by

$$W_{q,\pi}(t) = \frac{1}{b_\pi^A(t) + b_\pi^B(t)} \left(\sum_{i=1}^{b_\pi^A(t)} w_{q,\pi}^A(i, t) + \sum_{i=1}^{b_\pi^B(t)} w_{q,\pi}^B(i, t) \right). \tag{3.6}$$

During the stationary regime, the mean waiting time in queue for type m customers, say $W^m_{q,\pi}$, and the overall mean waiting time in queue for all customers, say $W_{q,\pi}$, are given by

$$W^m_{q,\pi} = \lim_{t\to\infty} W^m_{q,\pi}(t), \text{ and } W_{q,\pi} = \lim_{t\to\infty} W_{q,\pi}(t). \tag{3.7}$$

Now, we are going to underline one important property of performance measures. It is the variance of the waiting time. Minimizing the number of customers lost, or minimizing the mean waiting time in queue are only some important properties among many others. It often has been argued that a system with reasonable and predictable waiting time may be more desirable than a system with lower mean waiting but highly variable. We refer the reader to Lu and Squillante [93] for more details.

As above, we define the variance of the waiting time for type m customers and an overall variance for all types, in both transient and stationary regimes. The variance $Var^m_\pi(t)$ for type m customers during $[0,t]$ is defined by

$$Var^m_\pi(t) = \frac{\sum_{i=1}^{b^m_\pi(t)} \left(w^m_{q,\pi}(i,t) - W^m_{q,\pi}(t)\right)^2}{b^m_\pi(t)}. \tag{3.8}$$

We define the overall variance of the waiting time in queue for the customers who enter service during the interval of time $[0,t]$ by

$$Var_\pi(t) = \frac{\sum_{m\in\{A,B\}} \sum_{i=1}^{b^m_\pi(t)} \left(w^m_{q,\pi}(i,t) - W_{q,\pi}(t)\right)^2}{\sum_{m\in\{A,B\}} b^m_\pi(t)}. \tag{3.9}$$

During the stationary regime, the variance for type m customers, say Var^m_π, and the overall variance for all types, say Var_π, are given by

$$Var^m_\pi = \lim_{t\to\infty} Var^m_\pi(t), \text{ and } Var_\pi = \lim_{t\to\infty} Var_\pi(t). \tag{3.10}$$

Finally, we define the standard deviation, in transient and stationary regimes, for each type and for all types by taking the square root of the variances defined above. We denote these quantities by $\sigma^m_\pi(t)$, $\sigma_\pi(t)$, σ^m_π and σ_π, respectively.

In what follows, we present some results about the relation between the performance measures of interest and the discipline of service. Let us recall a known result for queueing system with infinitely patient customers, that is, a customer never leaves the queue before beginning service. It is well known that the expected time in system and expected time in queue are independent of the queue discipline. We only need to assume that the remaining total service or work required

at any point during an arbitrary busy period is order-of-service independent. In other words, no service needs are created or destroyed within the system: no renege in the midst of service, no preemption when service times are not exponentially distributed, no forced idleness of servers, and so on. The proof can be easily done by comparing the diagrams of the cumulative work for two different queue disciplines during the busy period, elsewhere both systems behaves identically owing to the workconserving property.
We should note that one may find counterexamples for the above result if service times are not assigned when service begins, see Whitt [134]. However, it is common that service times are associated upon customer arrivals. In such cases, one may still assume that service times are i.i.d. and independent of the arrival process to get the result. We refer the reader to Berger and Whitt [25] for more discussion.

We investigate below some conservation results in a more general queueing system including reneging. By means of Theorem (3.1) already derived in Pot [112], we motivate our consideration for only workconserving policies. Theorem (3.1) concerns a $GI/M/s+M$ system, which has an i.i.d. and generally distributed interarrivals, exponential service times, s servers, and exponentially distributed patience times.

Theorem 3.1 (Pot [112]) *Consider a $GI/M/s+M$ system with non-preemptive service discipline. The average abandonment rate is equal or higher under non-workconserving policies in comparison to workconserving policies.*

In what follows, we emphasize a number of theorems concerning workconserving policies. In Theorems (3.2) and (3.3), we investigate the conservation of the fraction of abandoning customers and the average waiting time in queue with respect to the scheduling policies, respectively. Some consequences are next derived in Corollaries (3.1) and (3.2).

Theorem 3.2 *Consider a $GI/GI/s+M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. Then, the service level Q is constant for any workconserving non-preemptive scheduling policy.*

Proof. We prove the result by coupling arguments. Consider two identical $GI/GI/s+M$ models, say Model 1 and Model 2. The discipline of service in Model 1 (Model 2) is defined by the workconserving non-preemptive policy π_1 (π_2.) We assume that policies π_1 and π_2 are different. Our approach is based on a single sample path. In both models, we create identical successive arrival epochs, as well as identical successive service times. Service times are assigned to servers and not to arrivals. Since times before reneging are exponentially distributed, then the decision for one customer to abandon the queue is not affected by his elapsed waiting time. This

enables us to create randomly, for each customer in queue, a new maximum time of patience at each selection for service (or equivalently successful departure epoch.) Assume that at time $t = 0$ both systems are empty, and let work begins. We denote by D_k the epoch of the k^{th} departure, $k = 1, 2, ..., \infty$.

Both models behave identically until a busy period starts and the following situation occurs: a server becomes idle (service completion) and more than one customer are waiting in queue. Let D_i be the epoch of that service completion (which occur simultaneously in Model 1 and 2.) For both models, let n be the number of waiting customers in queue just before D_i, $n \geq 2$. At D_i, the idle server in Model 1 selects one customer from the queue that can be different from the one selected by the idle server in Model 2. However, the number of customers in queue goes down by 1 for both models, it becomes $n - 1$. Note that the number of customers who abandon the queue is until now identical for both models.

In Model 1, we create for each customer waiting in queue a new maximum patience time. Without altering distributions, since times before reneging are identically distributed, we create the same set of $n - 1$ maximum patience times, and we assign them arbitrary to the customers waiting in Model 2. After D_i, three events are possible: one customer reneges, or a new customer enters the system, or a server becomes idle. Recall that by construction, these events occur simultaneously in both models. Assume that the first event occurs, then the number of customers who abandon the queue goes up by 1 in both models and as a consequence is still identical for them. It is still identical also if another customer abandons the queue. It is the case as long as the number of customers in queue is larger or equal to 1. Assume now that one customer enters the system. Hence, the number of customers in queue goes up by 1 in both models. Note that if another arrival occurs or that one customer abandons the queue, then, the number of customers in queue will increase by 1 or decrease by one, respectively. Thus, the number of customers who abandon the queue vary identically from one model to another. Assume now that one server becomes idle. If the number of customers in queue is less or equal to 1, it is obvious to see owing to the workconserving property, that policies π_1 and π_2 will select the unique available customer, if any. Otherwise, the busy period ends in both models, hence, both policies will select identically new arrivals for service until the beginning of the next busy period. However, if the number of customers in queue is greater or equal to 2, the selected customer for service may be different in both models. As above, we create for the remaining waiting customers in both models, the same set of maximum patience times. Recall that until now, the number of customers who abandon the queue is still identical for both models.

Continuing with the same arguments, we state that during the steady state, the number of

customers who abandon the queue in Model 1 coincides with the one in Model 2. Since by construction of the sample path the number of arrivals are also equal for both models, hence, we conclude that the fraction of abandoning customers is unchanged, $Q_{\pi 1} = Q_{\pi 2}$. This completes the proof of the theorem. □

Note that the result in Theorem (3.2) does not hold if service times are order of service dependent, or if we allow preemption when service times are not exponentially distributed, or if times before reneging are not identically and exponentially distributed. The proof of the conservation of Q for any workconserving policy (with or without preemption) can be easily obtained from Lemma (2) in Jouini and Dallery [66]. In the latter, the authors prove the result for a $GI/M/s/K+M$ queue with limited waiting space. Further details about that result will be given in Chapter 6.

In Theorem (3.3), we focus on the conservation of the waiting time in queue with respect to workconserving non-preemptive scheduling policies. We consider again a $GI/GI/s+M$ queue, and we focus on three different definitions of the average waiting time. Let W_q be the average waiting time in queue for served customers. Let W_q^{ab} be the one for abandoning customers, i.e., the average spending time in queue before leaving the system without being served. Finally, we define W_q^{tot} as the overall waiting time in queue for all customers, i.e., served as well as abandoning customers. In Theorem (3.3), we prove an intuitive result for the conservation of W_q^{tot}. In addition, we show a counterintuitive result for W_q and W_q^{ab}. Although the number of abandonments as shown in Theorem (3.2) does not vary for any workconserving non-preemptive scheduling policy, W_q and W_q^{ab} do vary.

Theorem 3.3 *Consider a $GI/GI/s+M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. When considering the class of workconserving non-preemptive scheduling policies, the following holds*

1. *W_q^{tot} does not depend on the scheduling policy.*
2. *W_q and W_q^{ab} depend on the scheduling policy.*
3. *The upper (lower) bound of W_q is achieved under the FCFS (LCLS) discipline of service.*
4. *The upper (lower) bound of W_q^{ab} is achieved under the LCLS (FCFS) discipline of service.*

Proof. We prove the first statement by coupling arguments. Using the same notations as in the proof of Theorem (3.2), we couple Model 1 and 2 using a single sample path. We showed that the number of waiting customers in queue is identically distributed for both models. Then, the

mean number of customers in queue, say L_q^{tot}, does not depend on the scheduling policy. Let λ be the average rate of arrivals. So, we state from the Little's Law that $\lambda\, W_q^{tot} = L_q^{tot}$. Then, we easily deduce that W_q^{tot} is independent of the scheduling policy, which completes the proof of the first statement.

We use a simple counterexample to prove the second statement. Let us couple Model 1 and 2. Model 1 is working under the FCFS discipline, and Model 2 is working under a workconserving non-preemptive policy different of the FCFS discipline, say π. The policy π works identically as the FCFS discipline except when 2 customers are waiting in queue and a service completion occurs. At that moment, π selects the younger customer (the second one), whereas the FCFS discipline chooses of course the one in the head of the queue (the older customer.) Using a single sample path, both models behave identically until the first time when two customers are waiting in queue and a service completion occurs. Let us stop our clock temporarily. Let D be the epoch of that event. We denote by A_1 and A_2 the first and the second waiting customers in the queue of Model 1, respectively. The same customers are also waiting in the queue of Model 2. Let w_1 and w_2 be the ages in queue of customers A_1 and A_2, respectively. Since A_1 entered in system before A_2, then $w_1 > w_2$. Also, let w_q^{FCFS} and w_q^{π} be the cumulative waiting times in queue for served customers in Model 1 and Model 2, respectively. Up to now, we have $w_q^{FCFS} = w_q^{\pi}$. Let our clock resumes ticking. The idle server in Model 1 selects the customer waiting in the head of the queue, namely A_1. However, the same server in Model 2 selects A_2. Updating the cumulative waiting times for the served customers leads to $w_q^{FCFS} = w_q^{\pi} + (w_1 - w_2)$, hence, $w_q^{FCFS} > w_q^{\pi}$. From the memoryless property of the distribution of times before reneging, we generate a new time before reneging and affect it twice: to A_2 (the unique customer waiting in Model 1) and to A_1 (the unique customer waiting in Model 2.) Only three non-zero probability events are possible: either a service completion occurs, or a customer abandons the queue, or a new customer joins the queue. Note that these events occur simultaneously for both models. If a service completion occurs first, then the idle server will select the unique available customer in queue, which allows w_q^{FCFS} to coincide again with w_q^{π}. If A_2 abandons in Model 1 and A_1 abandons in Model 2, then w_q^{FCFS} is still greater than w_q^{π}. Both quantities will never coincide thereafter. Assume now that a new arrival occurs. From the structure of both policies, we see that future events lead to: either A_1 in Model 2 and A_2 in Model 1 enter service, or A_1 and A_2 abandon before being served, or A_1 abandons and A_2 gets service. The event A_1 gets service and A_2 abandons is not possible. Thereafter, if A_1 abandons, then w_q^{FCFS} and w_q^{π} will never coincides again. Otherwise, we see that these quantities coincide again (in the best case) further to a given combination of events. A central statement is that there is no possible combination,

in any point of the sample path, which may make w_q^{FCFS} strictly lower than w_q^{π}. Following the same explanation until the stationary regime and knowing from Theorem (3.2) that the number of served customers is the same for both models, we finally state that the stationary average waiting time W_q^{FCFS} will be strictly greater than the stationary average waiting time W_q^{π}. This completes the proof of the second statement.

To prove the third statement, we again couple Model 1 and Model 2. The scheduling policy for Model 1 is the FCFS policy. The one for Model 2 is different of FCFS and is denoted by π'. Then, at least for some situations, the oldest customer (waiting in the head of the queue of Model 2) looses the higher priority for service. Taking a single sample path, let us now compare the cumulative waiting times for served customers in Model 1, say w_q^{FCFS} with that in Model 2, say $w_q^{\pi'}$. Initially and as long as both policies (FCFS and π') select identically the waiting customers, w_q^{FCFS} equals $w_q^{\pi'}$. The first time when π' selects a customer in a different manner as that in the FCFS discipline, w_q^{FCFS} is no longer equal to $w_q^{\pi'}$. Since the FCFS discipline selects the oldest customer, hence, w_q^{FCFS} becomes strictly larger than $w_q^{\pi'}$. Each time FCFS and π' select customers for service differently, w_q^{FCFS} becomes more and more larger $w_q^{\pi'}$. In a distant future, knowing that the number of served customers is unchanged under both policies, we state that the largest expected waiting time of served customers, W_q, is achieved under the FCFS policy. Applying above arguments by considering the LCLS discipline and a workconserving non-preemptive policy different of LCLS, we state that the LCLS policy is optimal subject to minimizing the average waiting time of served customers, which finishes the proof of the third statement.

The fourth statement is a direct consequence of the third one. It suffices to recall that the overall cumulative waiting time, defined as $w_q^{tot} = w_q + w_q^{ab}$, is unchanged under both policies. Hence, the policy that maximizes W_q will minimize W_q^{ab}, and vice versa. This completes the proof of the fourth statement and the theorem. □

Note also that although the first moment W_q^{tot} does not depend on the discipline of service, the second moment of the overall waiting time and thus the full distribution does depend on the discipline of service. As shown in Theorem (3.3), the maximum of the average waiting time for served customers, W_q, is achieved under FCFS discipline. However, we conjecture, based on a well known property in queueing literature, that the minimum of its variance is also achieved under the FCFS policy. In practice, if the value of W_q under the FCFS policy is not too far from that under another policy, thereafter, a call center manager will usually prefer the FCFS policy owing to its fairness. We refer the reader to Avi-Itzhak and Levy [17] for more details on the

fairness property in queueing systems.

We finally comment that the result in Theorem (3.3) is still valid when considering also preemptive scheduling policies, however, service times have to be exponentially distributed.

Corollary 3.1 *Consider a $GI/GI/s+M$ queue with two classes of customers A and B. Service times and times before reneging are identically distributed for both types of customers. Then, the overall service level Q and the overall expected waiting time W_q^{tot} are constant for any workconserving non-preemptive scheduling policy.*

Proof. This is an immediate consequence of Theorem (3.2) and the first statement of Theorem (3.3). It suffices to divide arrivals into two streams of customers to get the result. □

Denoting by c_{π_A} and c_{π_B} the respective achieved stationary service level ratios under policies π_A and π_B, the following result holds.

Corollary 3.2 *Consider a $GI/GI/s+M$ queue with two classes of customers A and B. Service times and times before reneging are identically distributed for both types of customers. Let π_A (π_B) be the policy that gives strict non-preemptive priority to customers A (B.) Then, for any workconserving non-preemptive policy, π, the achieved service level ratio in the stationary regime, c_π, satisfies the following relation*

$$c_{\pi_A} \leq c_\pi \leq c_{\pi_B}. \tag{3.11}$$

Proof. Consider a workconserving non-preemptive scheduling policy, say π, different from π_A. Then, in some situations under π, we first select a customer B from the queue whereas there is at least one waiting customer A. A sample path comparison of the system working under π with an identical one working under π_A may easily show that the distribution of the number of customers A waiting in the queue of the first system is greater than that for the second system. Subsequently, the number of customers A who abandon the queue is strictly greater in the first system, and equivalently, $Q_\pi^A > Q_{\pi_A}^A$. From that, one can state that

$$\min_{\pi\in\Pi}\{Q_\pi^A\} = Q_{\pi_A}^A, \tag{3.12}$$

where Π denotes the class of workconserving non-preemptive policies. As shown in Theorem (3.2), the overall fraction of customers who abandon, in the stationary regime, held constant under any workconserving non-preemptive policy. Thus, Equation (3.12) leads to

$$\max_{\pi\in\Pi}\{Q_\pi^B\} = Q_{\pi_A}^B. \tag{3.13}$$

From Equations (3.12) and (3.13), we thereafter deduce that $c_\pi \geq c_{\pi_A}$, which completes the proof of the first part of the corollary. For the second part, it suffices to follow the same arguments as above to state that $\max_{\pi\in\Pi}\{Q^A_\pi\} = Q^A_{\pi_B}$, and $\min_{\pi\in\Pi}\{Q^B_\pi\} = Q^B_{\pi_B}$. Thus, $c_\pi \leq c_{\pi_B}$, which finishes the proof of the corollary. □

Note that values of c_π ranging out of the interval $[c_{\pi_A}, c_{\pi_B}]$ may be achieved through non-workconserving policies, such as thresholds or reservations policies. These policies are indeed useful to discriminate between customer classes. Under such policies, the lower bound for Q^A (Q^B) is equal to that achieved under the policy that gives strict preemptive priority to type A (B) customers. Obviously, the upper bound for Q^A or Q^B is 1. It is reached for a given type by simply refusing service for all customers of that type.

3.3.3 Objective and Motivation

In this section, we motivate our objective with regard to the scheduling policies we aim to find. Before that, we highlight some known limits of the fixed strict priority policy. Consider the policy that assigns customers A to the high priority queue 1 and customers B to the low priority queue 2. It is well known that this policy makes the system highly unbalanced, and the QoS of customers B tends to be very low. Such behavior is undesirable for a call center manager. Unfortunately as mentioned before, it is very often in practice due to the highly uncertain environment of call centers that the workload is either underestimated, or overestimated. In both cases, the staffing step is incorrectly done, and as a consequence, service levels will be much more affected. Let us give an explanation. When the workload is underestimated, the QoS of customers B deteriorates even more. The system becomes less and less stable for that class, since most of the time, the service capacity will be assigned to handle type A customers. The unfairness is still valid when the workload is overestimated. Customers A get a service level that approaches the maximum while there is really no need for that. In such cases, a policy providing a slightly lower service level for customers A and a higher level for customers B would be of interest.

Having in mind above arguments, we are now ready to formulate our objective. We assume that scheduling of agents has already taken place, such that the number of available agents is known in advance. We aim to develop scheduling policies allowing, even in the case of unfavorable situations, to reach a fixed balance, during the stationary regime, for both service levels independently of the available service capacity. We should note that such policies need no information about the arrival process. The service level we consider here is the fraction of abandoning calls. We also keep in mind the advantages of having lower variances of waiting times. In addition,

we focus on studying simple policies that are easy to implement in practice, namely, priority scheduling policies. From practical reasons, we apply some simplifications to the general model. We assume that for each class of customers, interarrivals are i.i.d. and exponentially distributed. We assume also that service times are i.i.d. and exponentially distributed. We will not focus on the overall achieved fraction of abandoning customers, because anyway, this quantity is unchanged for any workconserving policy. In concrete terms, we specify our objective as follows. We aim to develop simple scheduling policies that minimizes the variance of the waiting time in queue of VIP customers, subject to satisfying a target ratio, c^*, of customer classes service levels.

3.4 Real-Time Scheduling Policies

The call center we consider here allows for flexible scheduling through dynamic alternate routing and sequencing, hereafter referred to as dynamic scheduling. Actually, this is easily possible for most call centers due to the technology development of their equipments. However, an interesting and challenging problem is to design scheduling policies that are simple to implement and their performances are acceptable for a call center manager. In this section, we restrict ourselves to develop simple online scheduling policies allowing to achieve an objective ratio, c^*. We consider several techniques that do take transient service levels into account and, hence, can be classified as online updating methods and real-time routing. The principle of our policies is that we adjust them during the evolution of the process. The adjustments depend on the history of the process.

Without loss of generality, we only consider an objective ratio, such that, $c^* < 1$. In fact, as we shall explain in the proof of Theorem (3.4), the case $c^* = 1$ reduces to the case of the FCFS discipline of service. As for the case $c^* > 1$, it is on the one hand not relevant for our analysis because we must keep in mind that type A customers are VIP. On the other hand, even if we would like to investigate that case, it suffices to apply the analysis for the case $c^* < 1$ by exchanging customers A by customers B and vice versa.

We propose three scheduling policies, say π_1, π_2 and π_3. The policies are belonging to the class of queue joining policies, which consists in assigning arriving customers to one of the queues, i.e., customer routing rules. The policies are dynamic, i.e., state-dependent, in the sense that upon arrival, a policy determines a rule for the queue assignment. Our policies do not anticipate on future events. They just react to the realization of the ratio that is determined by the history of the process. In addition, the proposed policies are easy to understand and implement in practice. Based on simulation experiments, a comparison analysis with regard to the variance of the waiting time is thereafter addressed in Section 3.5.

Scheduling Policy π_1

The scheduling policy π_1 starts work identically as a strict priority policy giving the higher priority to customers A. After the epoch when the first type B customer finishes his service, we apply the following assignment rule for any new arrival, which we denote by the k^{th} arrival. Let D_k be the epoch of that arrival. Let Q_k^A (Q_k^B) be the achieved service level from $t = 0$ until D_k for customers A (B.) Let c_k be the achieved ratio from $t = 0$ until D_k, $c_k = Q_k^A/Q_k^B$. If $c_k < c^*$, then we give high priority to customers B, that is, if the new arrival is type A, it is routed to queue 2, otherwise, it is routed to queue 1. However if $c_k \geq c^*$, we give high priority to type A customers, that is, if the new arrival is type A, it is routed to queue 1, and if it is type B, it is routed to queue 2. The scheduling policy π_1 is shown on Figure 3.2.

Scheduling Policy π_2

The scheduling policy π_2 starts work identically as π_1 until the first customer B finishes service. Following the same notations as in the last paragraph, let a new arrival enter system. Under π_2, a customer A is always routed to queue 1. However, the assignment rule of customers B is as follows. If $c_k < c^*$, then a new type B arrival is routed to queue 1, otherwise if $c_k \geq c^*$, it is routed to queue 2. The scheduling policy π_2 is shown on Figure 3.3.

Scheduling Policy π_3

The scheduling policy π_3 starts work identically as the policy π_1 and π_2 until the first customer B finishes service. Again, following the same notations, let a new arrival enter system. Under π_3, a customer B is always routed to queue 2. However, the assignment rule of customers A is as follows. If $c_k \geq c^*$, then a new type A arrival is routed to queue 1, otherwise if $c_k < c^*$, it is routed to queue 2. The scheduling policy π_3 is shown on Figure 3.4.

The scheduling policy π_1 can be immediately obtained intuitively. It allows the achieved ratio to be updated upon each arrival such that it converges in the long-run to the objective. The idea behind policy π_2 is that we keep always customers A in the high priority queue, however when it is necessary, we assign customers B to this queue to improve their service level (which deteriorates the service level of customers A.) Such a rule allows to increase the transient ratio and to keep it close to the objective. As a consequence, the ratio would converge in a distant future to the desired value. The policy π_3 can be viewed as another variant. It allows some times to penalize customers A by assigning them to the low priority queue, which again allows to increase the transient ratio. Our choice for policies π_2 and π_3 comes from the assumption, $c^* < 1$.

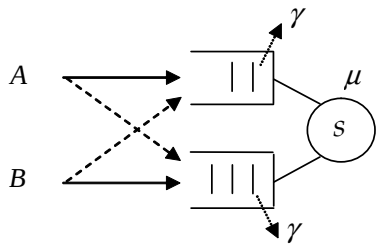

Figure 3.2: Scheduling policy π_1

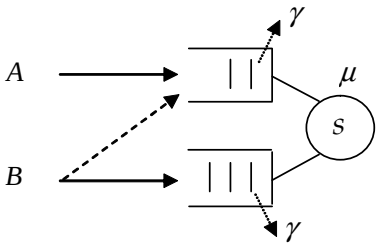

Figure 3.3: Scheduling policy π_2

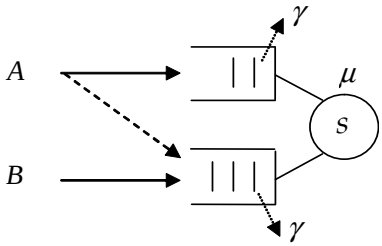

Figure 3.4: Scheduling policy π_3

Theorem 3.4 *Using the above notations, the following holds.*

1. π_1 *reaches* c^* *if and only if* $c_{\pi_A} \leq c^* \leq c_{\pi_B}$.
2. π_2 *reaches* c^* *if and only if* $c_{\pi_A} \leq c^* \leq 1$.
3. π_3 *reaches* c^* *if and only if* $c_{\pi_A} \leq c^* \leq 1$.

Proof. We start by proving the first statement. Let us take our basic model working under the scheduling policy π_1. First, it is easy to see that there are two bounds for achievable ratios, namely c_{π_A} and c_{π_B}. The reason comes from the fact that we are considering workconserving non-preemptive policies. The lower (upper) bound, c_{π_A} (c_{π_B}), is achieved when we give strict priority to customers A (B.) In addition, assigning dynamically customers to the high or the low priority queue, as under policy π_1, will affect the quantities $Q^A_{\pi_1}$ and $Q^B_{\pi_1}$ in the stationary regime, and equivalently the ratio $c_{\pi_1} = Q^A_{\pi_1}/Q^B_{\pi_1}$. Consider a given objective c^* ranging between c_{π_A} and c_{π_B}. During the transient regime, if it happens that the achieved ratio is strictly lower than c^*, then giving the priority to customers B (which is possible under π_1) allows necessarily to go beyond c^* after a given duration of time. Continuing in doing this, the ratio converges to c_{π_B}. Otherwise during the stationary regime, if it happens that the achieved ratio is strictly greater than c^*, then giving the priority to customers A (which is possible under π_1) allows necessarily to go below c^* after a given duration of time. Continuing in doing this, the ratio converges to

c_{π_A}. As the number of arrivals grows, these manipulations would make the difference between the achieved ratios and the objective less and less low. This would allow the stationary ratio to coincide with c^*.

Let us now focus on proving the second statement. Let c_{FCFS} be the achieved ratio under the FCFS policy. With the same explanation as above, one easily states that any ratio ranging from c_{π_A} and c_{FCFS} can be reached by the policy π_2. On the one hand, The lower bound for Q^A is achieved when we give strict priority to A customers. In addition at the same time, the upper bound for Q^B is reached. Thus, the minimum possible achievable target ratio corresponds to the policy c_{π_A}. On the other hand, the lower bound for Q^B is achieved by assigning all type B arrivals to queue 1, which also allows to achieve the upper bound for Q^A. This corresponds to the FCFS policy for all arrival types. Hence, the achieved ratio could not be worse than that under the c_{FCFS} policy. One may easily see that $c_{FCFS} = 1$. In fact, our model working under the FCFS manner, for each customer type and for both customer types, is simply equivalent to a single class model working under the FCFS manner.

Finally, we note that the proof of the third statement is similar to that of the second statement. This completes the proof of the theorem. □

One may construct several auxiliary policies similar to the above ones. For example instead of changing the priority rule at each new arrival epoch, we only change it at the arrival epoch of the customer who finds all servers busy and both queues empty. Then, we keep that rule until the end of the current busy period. With regard to reaching the stationary target ratio, the latter class of policies has the same properties as those for the class of policies π_1, π_2 and π_3. One drawback could be that they are less reactive to correct the transient ratio. A further possibility is to construct similar policies by changing the priority rule cyclically; at a given arrival and based on the transient ratio, we determine the priority rule and we keep it for a given fixed number of the next following arrivals. Once the cycle finishes, we determine the priority rule at the epoch of the arrival that follows the cycle. Again, we keep that rule for the same given fixed number of new arrivals, and so on.

In the following few sentences, we briefly describe an interesting dual class of policies, namely a class of call-selection policies. Whenever a server becomes idle, the policy has to decide which customer from the queue, if any, should be selected for service. It consists to select waiting calls by using so-called waiting time factors. When selecting a call, the idle agent considers the longest waiting customer in each queue. From this set of customers, he chooses the customer of which the product of the waiting time and the waiting time factor is the highest. The factor of a queue

is the same for all agents.

The idea behind these policies was introduced by Lu and Squillante [93], and addressed in details within our context by Pot [112] and Jouini et al. [70]. An advantage of waiting time factors is its flexibility with regard to several different routing policies that are possible. Setting both waiting time factors equal will result in a policy that serves customers from both classes in a FCFS order. Taking one waiting time factor equal to zero will give one of both types the full priority over the other. Waiting time factors between 0 and 1 may lead to the target ratio while giving low variability in the waiting time.

3.5 Simulation Results

Even though the workconserving scheduling policies we present here, lead to the target ratio and does not affect the overall fraction of abandoning calls, it will affect the waiting time variance. Because certainty is usually preferred to uncertainty, minimizing the variance is a logical basis for selecting the appropriate policy. A low variance is very useful when we want to estimate and inform customers about their queueing delays, see Armony et al. [15]. In fact, deriving the state-dependent estimation is too complicate, even more for our context here. The best we can do could be computing the mean value of the state-dependent waiting time. Then, a lower variance will give more credibility to the anticipated mean waiting times, see Jouini et al. [68]. More details about predicting and announcing queueing delays are given in the next chapter. A further advantage for a system with a low waiting time variance is related to its inherent fairness with regard to customers delays. Serving customers within comparable delays represents indeed an important issue for both the manager and the customer.

Both analytic and numeric methods are too complex for a direct analysis of the policies comparison. Thus, we resort to simulation experiments to prove their efficiency and gain useful guidelines. We consider six systems, denoted by System 1, ..., System 6. Systems parameters are chosen so as we get realistic scenarios. The number of servers is $s = 50$. The common service rate is $\mu = 0.2$, i.e., the mean service time for one customer is 5 min. The common reneging rate is $\gamma = 0.33$. From one system to another, we vary the total arrival rate so as we get different "service utilizations", $\frac{\lambda_A+\lambda_B}{s\mu}$. We choose, $\lambda_A = \lambda_B = 4.5$, 4.75, 4.9, 4.95, 5, and 6, respectively. The "service utilization" is increasing (starting from 90% in system 1 until 120% in system 6.) We choose balanced cases for both arrival processes so that we facilitate the comment of the simulation results. Recall that abandonments make our systems unconditionally stable. The simulations are done for the target ratios $c^* = 0.5$, 0.7 and 0.9. We determined for each system the interval $[c_{\pi_A}, c_{\pi_B}]$, and we checked that the values $c^* = 0.5$, 0.7 and 0.9 are ranging in all of

these intervals. For each system, we give the performance measures under policies π_1, π_2 and π_3, as well as those under policy π_A (high priority for type A customers.) The simulations results are presented in Table 3.1 below for $c^* = 0.7$, and in Tables B.1 and B.2 of Appendix B for $c^* = 0.5$ and 0.9, respectively. The rows corresponding to the quantity c are to indicate the achieved stationary ratio under each scheduling policy.

The end of section is devoted to a discussion of the simulation results. As expected, the total fraction of abandoning calls, Q, is independent of the scheduling policy. The negligible deviations in the values presented in Tables 3.1, B.1 and B.2 are due to the simulation duration. These quantities will necessarily coincide when running the simulations for a very long duration. We check from the experiments that the target ratio is always met by policies π_1, π_2 and π_3, which agrees with Theorem (3.4). For each system, the value of the ratio under policy π_A represents a lower bound for the achievable ratio under any workconserving non-preemptive scheduling policy. We can not do better when considering that class of policies.

In what follows, we address a comparison analysis with regard to the standard deviation of the waiting time in queue. We do not conduct a rigorous analysis in the sense that we do not prove our statements. Such a work is of great value. We leave it for a future research. However, we give here some general ideas and intuitive explanations to support the claims we derive.

For type A customers, starting from the lower value, most experiments show that the standard deviation values are structured in turn for policies π_A, π_2, π_3 and π_1. Let us give an intuitive explanation. The reason is basically related to the well known property in queueing theory which claims that the FCFS discipline minimizes waiting time variance (time in queue and in system) when the queue discipline is service time independent. We refer the reader to Randolph [113] for more discussion. The best we can do for customers A under a workconserving non-preemptive policy is to not give at any time the higher priority to customers B. Such situation allows customers A waiting times to be as low as possible. This is the case for policy π_A. Next, since the discipline of service within queue 1 is FCFS, then π_A should lead to the lower variance. With regard to the order of service of customers A, policy π_1 deviates more than π_2 and π_3 from the FCFS discipline. This tells us that π_1 has the higher variance. When comparing policies π_2 and π_3, one may see that on the contrary of π_3, policy π_2 respects the FCFS order for customers A, which allows it to ensure a lower variance than that under π_3.

For type B customers, starting from the lower value, we conclude from the majority of the experiments that the standard deviation values are structured in turn for policies π_3, π_2, π_1 and π_A. When comparing policies π_1, π_2 and π_3, the explanation is identical to that conducted for the first comment. As for policy π_A, the only explanation we have is related to the waiting

		π_A	π_1	π_2	π_3
System 1: $\lambda_A = \lambda_B = 4.5$	c	0.302	0.700	0.700	0.700
	Q^A	1.336%	2.057%	2.056%	2.203%
	Q^B	4.430%	2.938%	2.937%	3.146%
	Q	2.883%	2.497%	2.496%	2.675%
	W^A	0.039	0.055	0.057	0.061
	W^B	0.114	0.077	0.078	0.085
	W	0.076	0.066	0.068	0.073
	σ^A	0.104	0.194	0.164	0.174
	σ^B	0.332	0.255	0.244	0.244
	σ	0.247	0.227	0.207	0.212
System 2: $\lambda_A = \lambda_B = 4.75$	c	0.282	0.700	0.700	0.700
	Q^A	1.840%	3.446%	3.445%	3.449%
	Q^B	6.518%	4.922%	4.922%	4.927%
	Q	4.180%	4.184%	4.184%	4.188%
	W^A	0.053	0.092	0.097	0.096
	W^B	0.170	0.128	0.131	0.134
	W	0.110	0.110	0.114	0.115
	σ^A	0.119	0.260	0.215	0.220
	σ^B	0.407	0.339	0.318	0.308
	σ	0.303	0.303	0.272	0.268
System 3: $\lambda_A = \lambda_B = 4.9$	c	0.267	0.700	0.700	0.700
	Q^A	2.290%	4.465%	4.469%	4.464%
	Q^B	8.564%	6.378%	6.385%	6.377%
	Q	5.427%	5.421%	5.427%	5.420%
	W^A	0.067	0.119	0.126	0.125
	W^B	0.225	0.167	0.170	0.174
	W	0.143	0.143	0.148	0.150
	σ^A	0.131	0.302	0.246	0.252
	σ^B	0.472	0.392	0.365	0.351
	σ	0.351	0.350	0.311	0.306
System 4: $\lambda_A = \lambda_B = 4.95$	c	0.258	0.700	0.700	0.700
	Q^A	2.731%	4.845%	4.874%	4.837%
	Q^B	10.597%	6.921%	6.963%	6.910%
	Q	6.664%	5.883%	5.918%	5.874%
	W^A	0.080	0.129	0.138	0.136
	W^B	0.280	0.182	0.186	0.190
	W	0.176	0.155	0.162	0.163
	σ^A	0.143	0.317	0.257	0.263
	σ^B	0.533	0.411	0.383	0.366
	σ	0.396	0.367	0.326	0.319
System 5: $\lambda_A = \lambda_B = 5$	c	0.254	0.700	0.700	0.700
	Q^A	2.611%	5.216%	5.225%	5.222%
	Q^B	10.076%	7.452%	7.464%	7.460%
	Q	6.343%	6.334%	6.344%	6.341%
	W^A	0.076	0.139	0.148	0.147
	W^B	0.266	0.196	0.199	0.205
	W	0.167	0.167	0.174	0.176
	σ^A	0.139	0.331	0.266	0.274
	σ^B	0.516	0.428	0.397	0.380
	σ	0.383	0.383	0.338	0.332

		π_A	π_1	π_2	π_3
System 6: $\lambda_A = \lambda_B = 6$	c	0.189	0.700	0.700	0.700
	Q^A	5.598%	14.597%	14.516%	14.516%
	Q^B	29.649%	20.853%	20.737%	20.737%
	Q	17.623%	17.724%	17.626%	17.626%
	W^A	0.167	0.389	0.438	0.432
	W^B	0.889	0.574	0.577	0.624
	W	0.475	0.478	0.505	0.524
	σ^A	0.187	0.634	0.428	0.463
	σ^B	0.928	0.793	0.679	0.622
	σ	0.718	0.721	0.567	0.554

Table 3.1: Simulation experiments for $c^* = 0.7$

time values. The larger waiting times of type B customers are achieved under π_A, because of their lower priority. However due to uncertainty in arrivals, reneging and service times, there is a non-zero probability that some customers B enter service without waiting or within a short delay. This allows their waiting times variance to be as a consequence the higher.

Recall that a further investigation is required. We only gave some directions to compare the achieved values of variances. For instance, the comparison should lie in the values of λ_A and λ_B, also in the objective c^*. An objective close to 1 would make our policies work in a similar manner than that of the FCFS discipline, whereas an objective far from 1 (being under or beyond) would make the policies similar to the strict priority policy. Based on the analysis here, we can not distinguish a best policy. However, one may recommend policy π_2. First, it reaches the objective ratio. Second, it gives (in most cases) the lower variance of VIP customers waiting times. Third, it allows to have a “good" variance for customers B, as well as for all customer types.

3.6 Extensions

In this section, we discuss some extensions of the analysis of this work. In Section 3.6.1, we investigate the extension of our online policies to the case of three customer classes. In Section 3.6.2, we focus on a call center working day. We assume that the objective ratio is achieved for every period of the day, and we investigate whether a balance is also reached for the whole day.

3.6.1 Extension to Three Customer Classes

In this section, we tackle the extension of the proposed scheduling policies to the case of three customer classes. We consider a generalization of our call center queueing model with three customer types A, B and C. In addition, the queueing model has three infinite queues denoted by queue 1, 2 and 3. Our basic goal here is to discuss the usefulness of our online policies

subject to reaching a proportionality on the service levels. We denote the stationary fractions of abandoning calls of type A, B and C by Q^A, Q^B and Q^C, respectively. In the case of three classes, we have to define two objective ratios. Without loss of generality, the first objective ratio, say c_1^*, is defined for types A and B by $c_1^* = Q^A/Q^B$. The second, say c_2^*, is defined for types B and C by $c_2^* = Q^B/Q^C$. As a consequence, the ratio Q^A/Q^C will be equal to $c_1^* \times c_2^*$, which we denote by c_3^*.

For workconserving non-preemptive policies, let S be the set of all feasible couples of ratios $(\frac{Q^A}{Q^B}, \frac{Q^B}{Q^C})$. Thus the following policy, say π_1', allows to reach any objective (c_1^*, c_2^*), such that $(c_1^*, c_2^*) \in S$.

Scheduling Policy π_1'

The scheduling policy π_1' is an extension of the policy π_1. For its initialization, it starts work identically as a strict non-preemptive policy giving the higher priority to customers A, then B, and customers C have the lower priority. After the epoch when at least one type B customer and one type C customer have finished their service, we apply the following assignment rule for any new arrival, denoted by the k^{th} arrival. Let D_k be the epoch of that arrival. Let Q_k^A, Q_k^B and Q_k^C be the achieved service levels from $t = 0$ until D_k for customers A, B and C, respectively. Let $c_{1,k}$, $c_{2,k}$ and $c_{3,k}$ be the achieved ratios from $t = 0$ until D_k, i.e., Q_k^A/Q_k^B, Q_k^B/Q_k^C and Q_k^A/Q_k^C, respectively.

- If $c_{1,k} \geq c_1^*$, then we give the priority to type A over type B. Otherwise, we give the priority to type B over type A.
- If $c_{2,k} \geq c_2^*$, then we give the priority to type B over type C. Otherwise, we give the priority to type C over type B.
- If $c_{3,k} \geq c_3^*$, then we give the priority to type A over type C. Otherwise, we give the priority to type C over type A.

From the previous tests, we therefore determine a strict priority level for each customer type. So, customers with the highest priority are assigned to queue 1. Those with the second higher priority are assigned to queue 2. Finally, those with the lower priority are assigned to queue 3. Thereafter based on its type, the new arrival is treated under this queue joining rule. The queue joining rule is updated for every new arrival. The scheduling policy π_1' is shown on Figure 3.5. For example, let a new arrival occurs. Based on the history of the process, assume that we have $c_{1,k} < c_1^*$, $c_{2,k} \geq c_2^*$ and $c_{3,k} < c_3^*$. Thus, starting from the highest priority, the priority levels are structured in turn for type B, C and A. Hence for the next arrival, types B, C and A have to be assigned to queues 1, 2 and 3, respectively.

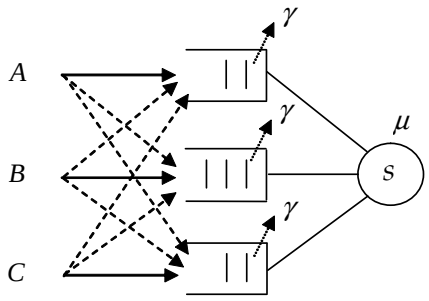

Figure 3.5: Scheduling policy π'_1

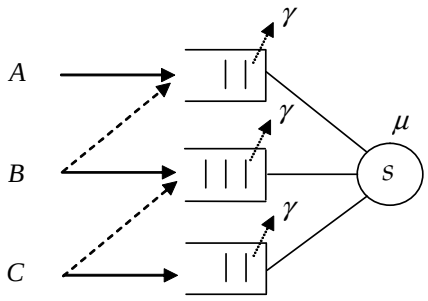

Figure 3.6: Scheduling policy π'_2

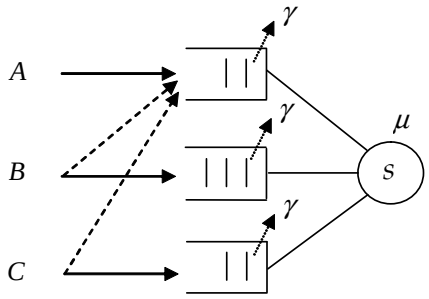

Figure 3.7: Scheduling policy π'_3

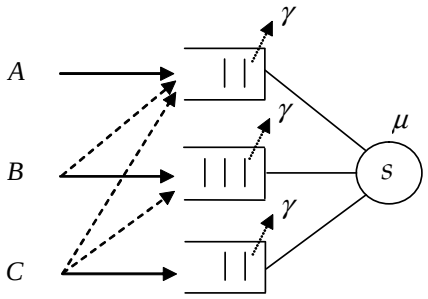

Figure 3.8: Scheduling policy π'_4

Let us remember that type A customers are more valuable for the company than type B ones, who are however more valuable than type C customers. Then, both objective ratios c_1^* and c_2^* should be in practice strictly lower than 1. In such a case, the scheduling policy π'_2, shown on Figure 3.6, should perform better with regard to the variance of the waiting time in queue. However, it could not achieve any objective in S ($c_1^* > 1$ or $c_2^* > 1$.) One may also propose additional alternative policies, namely π'_3 (Figure 3.7) and π'_4 (Figure 3.8.) The latter policies are different of π'_2 in the sense that they allow to achieve values beyond 1 for c_2^*. Note that it would be interesting to investigate similar policies for models with only two queues instead of three. It would be also interesting to compare the proposed policies in this section with regard to the variance of the waiting time and the region of feasible objectives. We leave this work for a future research.

3.6.2 Objective Ratio for the Whole Day

The origin of our optimization problem, i.e., satisfying a giving objective ratio, is as follows. In practice, a call center manager aims initially to reach some giving differentiated service levels, for example Q^{A*} and Q^{B*} for types A and B, respectively. As mentioned before, the staffing level or the service capacity are not a decision variables for our problem. As a consequence, the objective service levels may not be reached exactly. For that reason, the purpose of our call center manager becomes reaching a ratio of the achieved service levels which equals that of the

objective ones, $Q^{A*}/Q^{B*} = Q^A/Q^B$. The value of the objective ratio c^* defined in the beginning of this chapter is, indeed, the ratio Q^{A*}/Q^{B*}.

Let us now consider the general case of a call center with m customer types. For $1 \leq i \leq m$, Q^{i*} denotes the objective service level for type i customers, and under a giving scheduling policy, Q^i denotes the achieved one. Based on the previous remark, the constraint analyzed in this chapter, with regard to the ratio of the service levels is equivalent to

$$\frac{Q^i}{Q^{i*}} = \delta, \text{ for } i = 1..m, \tag{3.14}$$

where δ is the new objective ratio. The quantity δ translates an objective of proportionality between achieved and target service levels.

After this brief development, we are now ready to tackle the major question of this section. It is well known that in most call centers, the arrival rate is time varying (according to the period of day, day of the week, holidays, etc.) The number of agents is also fluctuating over the day. In practice, the change in these parameters are small enough, and are slow relative to the speed at which the call center reaches the steady state. It is a plausible assumption to consider constant parameters within each half-hour interval of time. We refer the reader to Gans et al. [40] and Garnett et al. [44] for more details. Note that the analysis in the core of this chapter focuses on a giving period of the day where the parameters of the system are assumed to be constant. In this context, it would be interesting to wonder whether there is a conservation of the proportionality during the whole day, knowing that we have reached that proportionality for each period of the day.

In mathematical terms, let us divide the day to n distinct periods. Let us assume that $\frac{Q_t^i}{Q_t^{i*}} = \delta_t$ for $i = 1..m$ and $t = 1..n$, where Q_t^i and Q_t^{i*} denote, respectively, achieved and objective service levels for type i customers during the period t. For simplicity, we assume that Q_t^{i*} is held constant for all periods, $Q_t^{i*} = Q^{i*}$. Thereafter, what is the condition under which we have for all customer types $\frac{Q^i}{Q^{i*}} = \delta$, where δ is a given constant? We give Proposition (3.1) to answer this question.

Proposition 3.1 *Assume that $\frac{Q_t^i}{Q^{i*}} = \delta_t$ for all $i, t \in [1..m] \times [1..n]$. Let λ_t^i be the mean arrival rate during the period t for type i customers, and λ_t that for all types. If $\frac{\lambda_t^i}{\lambda_t} = \beta^i$, where β^i is a given constant for $i = 1..m$. Then,*

$$\frac{Q^i}{Q^{i*}} = \delta, \text{ for all } i = 1..m, \tag{3.15}$$

where δ is a given constant.

Proof. Let $i,t \in [1..m] \times [1..n]$. Let $N_t^{tot,i}$ be the total number of type i arrivals during the period t. Let $N_t^{ab,i}$ be the number of type i abandoning customers during t. Then, one may write $Q_t^i = \frac{N_t^{ab,i}}{N_t^{tot,i}}$, and for the whole day, one has

$$Q^i = \frac{\sum_{t=1}^n N_t^{ab,i}}{\sum_{t=1}^n N_t^{tot,i}}. \tag{3.16}$$

Thereafter,

$$\begin{aligned}
\frac{Q^i}{Q^{i*}} &= \frac{N_1^{ab,i} + N_2^{ab,i} + ... + N_n^{ab,i}}{Q^{i*} \cdot \sum_{t=1}^n N_t^{tot,i}} \\
&= \frac{\prod_{t=1}^n N_t^{tot,i}}{\sum_{t=1}^n N_t^{tot,i}} \cdot \left(\frac{N_1^{ab,i}}{Q^{i*} \cdot \prod_{t=1}^n N_t^{tot,i}} + \frac{N_2^{ab,i}}{Q^{i*} \cdot \prod_{t=1}^n N_t^{tot,i}} + ... + \frac{N_n^{ab,i}}{Q^{i*} \cdot \prod_{t=1}^n N_t^{tot,i}} \right) \\
&= \frac{\prod_{t=1}^n N_t^{tot,i}}{\sum_{t=1}^n N_t^{tot,i}} \cdot \left(\frac{\delta_1}{\prod_{t=1,\, t\neq 1}^n N_t^{tot,i}} + \frac{\delta_2}{\prod_{t=1,\, t\neq 2}^n N_t^{tot,i}} + ... + \frac{\delta_n}{\prod_{t=1,\, t\neq n}^n N_t^{tot,i}} \right).
\end{aligned} \tag{3.17}$$

Assume that $\frac{\lambda_t^i}{\lambda_t} = \beta^i$, or equivalently, $\frac{N_t^{tot,i}}{N_t^{tot}} = \beta^i$, where N_t^{tot} denotes the total number of arrivals for all types during t. Hence, we get for $i,j \in [1..m]$

$$N_t^{tot,i} = \frac{\beta^i}{\beta^j} \cdot N_t^{tot,j}. \tag{3.18}$$

For $i,j \in [1..m]$, applying Equation (3.18) in Equation (3.17) leads to

$$\begin{aligned}
\frac{Q^i}{Q^{i*}} &= \frac{\prod_{t=1}^n \frac{\beta^i}{\beta^j} \cdot N_t^{tot,j}}{\frac{\beta^i}{\beta^j} \cdot \sum_{t=1}^n N_t^{tot,j}} \cdot \left(\frac{\delta_1}{\prod_{t=1,\, t\neq 1}^n \frac{\beta^i}{\beta^j} \cdot N_t^{tot,j}} + \frac{\delta_2}{\prod_{t=1,\, t\neq 2}^n \frac{\beta^i}{\beta^j} \cdot N_t^{tot,j}} + ... + \frac{\delta_n}{\prod_{t=1,\, t\neq n}^n \frac{\beta^i}{\beta^j} \cdot N_t^{tot,j}} \right) \\
&= \frac{\prod_{t=1}^n N_t^{tot,j}}{\sum_{t=1}^n N_t^{tot,j}} \cdot \left(\frac{\delta_1}{\prod_{t=1,\, t\neq 1}^n N_t^{tot,j}} + \frac{\delta_2}{\prod_{t=1,\, t\neq 2}^n N_t^{tot,j}} + ... + \frac{\delta_n}{\prod_{t=1,\, t\neq n}^n N_t^{tot,j}} \right).
\end{aligned} \tag{3.19}$$

Using Equation (3.17), Equation (3.19) implies the following relation

$$\frac{Q^i}{Q^{i*}} = \frac{Q^j}{Q^{j*}}, \text{ for any } i,j \in [1..m]. \tag{3.20}$$

Finally, we conclude that the proportionality of the achieved and the objective service levels is conserved for the whole day. This completes the proof of the proposition. □

In a real call center case, arrivals are divided into two types. The first type is first-attempt calls, also referred to as fresh calls. The second type represents retrial calls. In most cases, the required condition in Proposition (3.1) holds roughly for fresh calls and not necessarily for both

types, i.e., observed calls.

3.7 Conclusions and Further Research

We focused on a fundamental short-term problem for the management of call centers. We considered a two-class call center and developed real-time scheduling policies that determine the rule of assignment of customers, upon arrival, to waiting lines. We focused on service levels criteria related to the fraction of abandoning customers and the variance of the queueing delay. These policies are characterized to be relevant in practice. They are easy to understand for managers, predictable and easy to implement. Addressing exact analyzes for such problems is often too complex, and a challenging issue is to design dynamic scheduling policies that are simple, predictable and whose performance is good in an appropriate sense.

First, we gave some structural results in order to better understand the impact of scheduling policies on the performance measures of interest. Second, we proposed several dynamic scheduling policies allowing to meet a target ratio between the fractions of abandoning calls. Thereafter, we conducted a simulation study to compare the proposed policies with regard to the achieved variances of waiting times. Finally, we presented two possible extensions. In the first extension, we focused on a call center model with three customer types. In the second extension, we addressed one issue dealing with our objective for the whole call center day, and not only one period of the day.

An interesting subject for future research would be to investigate static scheduling policies analogous to those proposed in the core of this chapter. Let us give further details. The idea behind static policies comes from Aguir et al. [7]. In their work, the authors consider an identical model to that described in Section 3.3.1, whereas they do not allow customers to renege while waiting in queue. They characterize a class of optimal static policies subject to satisfying differentiated performances for customer classes. The proposed policies are based on two parameters p_A and p_B. The quantity p_A (p_B) represents the static probability to assign new type A (B) arrivals to the queue with the highest priority, i.e., queue 1. Hence a new type A arrival is routed to queue 2 with probability $1-p_A$, and a new type B arrival is routed to queue 2 with probability $1-p_B$. Unfortunately, such analysis is untractable when considering reneging. Closed-form expressions for the quantities Q^A and Q^B are not indeed not available. Even an exact numerical computation is no possible. Here we only give an alternative idea to tackle the problem and leave a rigorous analysis for future research.

Given an objective c^*, a possible method to get the probabilities p_A and p_B would be as follows. We simulate our system under a dynamic policy that achieves the objective ratio, as

those proposed in Section 3.4. At the end of the simulation run, we take the proportions of types A and B customers assigned to queue 1. These quantities are thereafter assigned to p_A and p_B, respectively. It is easy to see for example that for policy π_2, $p_A = 1$, and for policy π_3, $p_B = 0$.

One would expect that the static policies yield to higher variances of the waiting times than those achieved under their corresponding online policies. This may be due to a known property in queueing theory. The general idea is that dynamic policies achieve the lower variance of the waiting time, then cyclic policies, and finally static policies (based on random assignment) achieve the higher values. However, we should note that static policies are easier to implement in practice.

Chapter 4

Modeling Call Centers with Delays Information

In this chapter, we study the effect of informing customers about their anticipated delays in a call center with impatient customers. First, we consider a single class call center model. We propose a method for modeling the customer reaction with regard to delays information. Thereafter, we conduct a numerical comparison between performance measures of both models with and without information. The experiments show how the expected customer satisfaction in the model with information would tip the scales in favor of that model. Second, we extend the analysis to the case of a two-class call center with strict priority. Finally, some practical issues are discussed. In particular, we propose a method of delays announcement referred to as announcement by increments. We shown how this method would improve the system behavior through reducing errors approximations.

An extended version of this chapter is the working paper Jouini, Dallery and Akşin [68].

4.1 Introduction

As in Chapter 3, this chapter deals with a real-time issue related to operations management in call centers. We focus on analyzing call centers where the service provider communicates anticipated delays to customers upon their arrival. The main reason of informing customers about their queueing delays is to alleviate congestion and reduce customer dissatisfaction with waiting.

Information about anticipated delays is specially important in service systems with invisible queues (tele-queue) such as call centers. In such systems, the uncertainty involved in waiting is higher than that in systems with visible queues. Upon arrival and during their waiting, customers have no means to estimate queue lengths or progress rate. So, the feelings of frustration and anxiety increase over their sojourn in queue. We expect that delays information would avoid such situations, and make the waiting experience more acceptable. Zakay [144] stipulates that waiting information may distract customers attention from the passage of time. Hence, they may perceive the length of the wait as shorter. Furthermore, we point out a vicious circle in call centers. When a new arrival customer perceives that his anticipated delay is too long, he could balk upon arrival without joining the system. This feature would considerably reduce customers reneging in queue, which allows to make the system more stable in the sense that the variability of queueing delays is reduced. The latter would in turn improve the quality of delays information we give to customers, which even more reduces customers reneging, and so on. A further argument for predicting delays may be to help managers in reorganizing their facility. For instance in case of large predicted delays, the manager recognizes the need of increasing the staffing level.

Predicting delays for arrival customers is state of the system dependent. This is different from estimating stationary performances and usually makes the analysis untractable. In the context of prediction and announcement of delays, an extra layer of complexity should be noted. The analysis becomes more difficult since we have to take into account the description of the system in addition to the announcements given to each waiting customer in queue. Existing research projects often look for approximations, as announcing the stationary mean waiting time, or announcing the actual delay of the last customer (motivated by large systems in an overload regime, see Armony et al. [15].) In this work, we basically use exact methods. Paralleling to the relevant Whitt's [135] paper, we first consider a single call center model with impatient customers and working under the FCFS discipline. We derive our main insights from analyzing this simple model. Next, we turn to extend the results to a quite complex multiclass priority system where the anticipated delay for a given type of a new customer may be affected by

future arrivals of other types. The performance evaluation of this work is basically related to the transient analysis of birth-death processes. The analysis is somewhat straightforward, whereas we use some results derived from Chapter 5 when addressing the multiclass call center case. Note that the Markovian assumptions required for the use of birth-death processes are not necessarily valid for all call centers cases, especially for service times and times before reneging. However as already mentioned in Chapter 3, they provide a good approximation of the general model performances. These assumptions are in addition helpful to gain practical insights.

A central outcome of this work deals with the critical issue of the impact of delays information on customers behavior. This is at the same time interesting and difficult due to the attractive human element governing the call center environment. Starting from each model (single and multiclass), we detail and justify the quantitative building of the new model with delays information. In our models, customers has the opportunity to balk in response to their anticipated delay. Further to the balking reaction, it has been shown in a real experience that the reneging experience may also change in response to delays information, see Feigin [37]. We model that effect for the simple single class call center. We then extend the model of Whitt [135] by letting already informed customers renege even after having chosen to join the queue. We propose a method for approximating the new reneging experience by pertaining it to the quality of the delay information. To show the benefits of moving to a call center with delays information, we conduct a quantitative comparison between both models with and without announcement. We describe how balking in the second model may reduce customers reneging. In practice, this feature makes the second model preferable because reneging customers are the costliest. For example, a customers who balks has a higher probability to call back that that of a customer who reneges. A reneging customer leaves the system with frustration and loosing trust in the service provider. However, a balking customer leaves the system based on an information. This information would avoid to loose business because it is perceived by balking customers as an invitation to call back when the system will be able to serve them within a reasonable delay.

In this chapter, we try to be as near as possible to reality in order to get useful guidelines for practitioners. Once we get in hand the predicted waiting time distribution of a new arrival, we investigate how the service manager should profit from that information to make the announcement. For instance, he may decide to provide the mean or any other percentile of the distribution. However, we should be careful: From the one side, informing a short waiting time, which is likely to underestimate the actual waiting, might reduce the reliability of the service provider in the eyes of the customers. On the other side, informing large waiting times increases the number of balking customers while having a system that might allow to serve customers

within shorter and reasonable delays.

The remainder of this chapter is structured as follows. In Section 4.2, we give a literature review related to the present work. In Section 4.3, we develop the main analysis for a single class call center. In Sections 4.3.1 and 4.3.2, we describe the basic model and compute the performance measures of interest, respectively. In Section 4.3.3, we build the new call center model by incorporating delays announcement. The performance measures of the new model are thereafter derived in Section 4.3.4. Next, we conduct in Section 4.3.5 a numerical comparison of both models with and without delays information. In Section 4.4, we tackle the extension of the analysis to a two-class priority call center. We specify both models with and without delays announcement in Sections 4.4.1 and 4.4.2, respectively. To characterize the two-class model with announcement, we explicitly derive in Section 4.4.3 the first two moments of virtual delays for each customer type. In Section 4.5, we move to investigate some practical issues. In Section 4.5.1, we give a helpful approximation for virtual delays which may reduce numerical difficulties in practice. In Section 4.5.2, we propose a method for announcing delays, namely announcing delays by increments. In Section 4.6, we give some concluding remarks and discuss extensions.

4.2 Literature Review

The literature close to the work in this chapter spans two main areas. The first one deals with the prediction and announcement of delays seen under a queueing problem perspective. The second area is related to the psychology of waiting and the qualitative impact of announcing delays on the customer behavior.

As for the first area, we refer the reader to the relevant work of Whitt [137]. The author focuses on estimating state-dependent delays. He presents both accurate methods and approximations of waiting times in different situations. Models under consideration are queueing systems with different customer classes, exponential and non-exponential service times. Nakibly [103] reviews several classical results, and extends the analysis to some other complex models with priorities. As already mentioned, addressing such problems is known to be difficult. It is often related to the transient analysis of birth-death processes, and in general to Markov chains. The literature dealing with birth-death processes is extensive and growing. We refer the reader to Section 5.1 of Chapter 5 for further details. The problem of predicting and announcing delays have recently received a lot of attention in the field of call centers. Letting customers renege is a central feature that makes the study of value in practice. The reader is referred to Section 3.2 of Chapter 3 for a review of papers about queueing models, and in particular call centers models, with impatient customers. Let us now give some literature on call centers with delays

information. Whitt [135] models and quantifies the effect upon performance, in a Markovian call center model, of giving state information to customers. Taking general distributions for service times and times before reneging, the authors in Armony et al. [15] develop methods to study that effect. The methods are based on fluid approximations. Guo and Zipkin [51] consider a simple queueing model where three levels of information can be provided to customers, namely no information, partial information and full information (the exact waiting time). Under different assumptions of parameters distributions, the authors thereafter investigate how information about delays can enhance system performances. Armony and Maglaras [13] consider a slightly different model. Based on his anticipated delay information, one customer may balk, elect to wait, or leave a message. When a message is left, the service provider calls back the customer within a guaranteed time. The authors estimate the guaranteed time under heavy-traffic regimes. Further references on the subject include those by Ward and Whitt [129], Salah [5], Armony and Maglaras [13] and references therein. We should note that although the modeling approach in the literature differs from one work to another, the findings usually confirm the benefits of communicating delays to customers.

The second area of literature close to this chapter is related to psychology of waiting. The literature on customers influenced by delay information begins with Naor [104]. An overview of customer psychology in waiting situations, including the impact of uncertainty, can be found in Maister [94]. Taylor [124] considers the relationship between delays and evaluation of service. He shows how the delay may lead to lower evaluation of service by creating both feelings of anger and uncertainty. Providing information about anticipated delays affects both customer satisfaction and customer behavior. A survey on the relationship between information and customer satisfaction can be found in Hui and Tse [59]. In Katz et al. [73], the authors describes an empirical study conducted in a bank. They show that informing about anticipated delays results in shorter perceived waiting times. A similar result was also found in the work of Carmon and Kahenman [32]. Hui and Zhou [60] show in their experience that informed customers appear to maintain a sense of control during the wait, which in turn affects the quality of service evaluation through the perceived waiting duration.

4.3 Single Class Call Center

In this section, we address the analysis of a call center with a single group of agents, serving a single class of customers. In Section 4.3.1, we describe the basic model of the call center without delay information. In Section 4.3.2, we derive its performance measures expressions. Note that some performance measures of that basic model were already derived by Whitt [135]. In Section

4.3.3, we propose a new model to incorporate the change of customers behavior with respect to announcing delays. In particular, we let customers balk upon arrival based on the information we provide to them. In addition, we let them renege even they already elect to wait, and we propose a method to model that feature. The performance measures of the new model are thereafter derived in Section 4.3.4. Finally, we conduct in Section 4.3.5 a numerical comparison between both models with and without delays information.

4.3.1 Basic Model

Consider the queueing model of the call center with a single class of customers. The model consists of one infinite queue, and a set of s parallel, identical servers representing the set of agents. All agents are able to answer all customers. The call center is operated in such a way that at any time, any call can be addressed by any agent. So upon arrival, a call is addressed by one of the available agents, if any. If not, the call must join the queue. Customers waiting in queue are served in order of their arrival, i.e., under the FCFS discipline. Interarrival times as well as successive service times are assumed to be i.i.d. and exponentially distributed. The arrival rate is λ, and the service rate is μ. In the same manner as in Chapter 3, we let customers be impatient. Recall that abandonments make the system unconditionally ergodic. Times before reneging are assumed to be i.i.d. and exponentially distributed with rate γ. We refer the reader to Section 3.3.1 of Chapter 3 to explain our motivation of assuming identical distribution of patience for all waiting customers independently of their position in queue. Finally, retrials are ignored, and reneging is not allowed once one customer starts his service. Following similar arguments, the call center can be viewed as an $M/M/s+M$ queueing system. The model is referred to as Model 1.

4.3.2 Performance Measures without Announcement

In this section, we tackle the analysis of the original call center model presented in Section 4.3.1. Most of the quantitative analysis are inspired from Whitt [135]. Our approach is based on system states probabilities seen by a randomly chosen new arrival. From the PASTA property (Poisson Arrivals See Time Averages), these probabilities coincide with those seen by an outside random observer, that is simply the probabilities that the system is in a given state at a random instant. The PASTA property is based on the memoryless property of the Poisson process, which allows to generate a sequence of arrivals that take a random look at the system. We refer the reader to Kleinrock [81] for further explanation, and Wolff [141] for a rigorous proof.

We denote the system state by a random variable taking non-negative integer values rep-

resenting the number of customers in system at a given instant. Let $p_1(i)$ be the steady state probability that i customers are present in Model 1 at a random instant, $i \geq 0$. To compute the steady state probabilities, we define a birth-death model as shown on Figure 4.1. The birth rate is constant, it represents the arrival intensity λ. When the number of customers present in system is less than or equal to s, all departures are only service completions. Otherwise, departures may be service completions or abandonments. Thereby, the death rate is $i\mu$ for $0 \leq i \leq s$, and $s\mu + (i-s)\gamma$ for $i > s$.

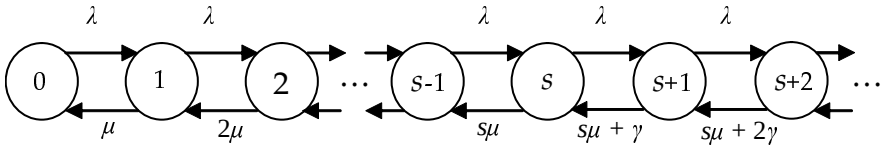

Figure 4.1: Birth-death process for Model 1

In a distant future, one has a set of infinite recursive relations relating the steady state probabilities. By adding the ergodicity condition which holds for any $\gamma > 0$, we go on to solve by iteration and get the following solutions

$$p_1(i) = \frac{\lambda^i}{i!\,\mu^i} \cdot p_1(0) \text{ for } 0 \leq i \leq s, \text{ and } p_1(i) = \frac{\lambda^i}{s!\mu^s \prod_{j=1}^{i-s}(s\,\mu + j\,\gamma)}\, p_1(0) \text{ for } i > s, \tag{4.1}$$

where $p_1(0)$ is the stationary probability to have no customers in system. It is given by

$$p_1(0) = \left(\sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i} + \frac{1}{s!\,\mu^s} \sum_{i=s+1}^{\infty} \frac{\lambda^i}{\prod_{j=1}^{i-s}(s\,\mu + j\,\gamma)} \right)^{-1}, \tag{4.2}$$

which enables us to compute all system states stationary probabilities. Thus, the probability of immediate service, say $P_{is,1}$, the mean number of customers in queue, say $L_{q,1}$, and the mean number of customers in Model 1, say $L_{s,1}$, can be calculated as

$$P_{is,1} = \sum_{n=0}^{s-1} p_1(n), \quad L_{q,1} = \sum_{n=1}^{\infty} n \cdot p_1(s+n), \quad \text{and} \quad L_{s,1} = \sum_{n=1}^{\infty} n \cdot p_1(n). \tag{4.3}$$

Let us proceed to compute the probability for a new arrival to enter service, the first and second moments of its conditional waiting time in queue given that service is completed. We denote by $P_1(S)$ the probability of being served. When a new customer finds less than s customers in system, he gets service immediately. This is equivalent to find at least one server idle, and it occurs with the probability $P_{is,1}$ given in Equation 4.3. Let us now consider the complementary

event, i.e., assume that a new customer finds all servers busy and n waiting customers in queue, $n \geq 0$. To analyze such situation, we define a pure-death process with a state-dependent death rates, see Figure 4.2. The process is derived from that in Figure 4.1, whereas we only consider states ranging from s to $s+n+1$ (the $n+s$ already existing customers plus the new arrival.) We do not consider birth rates because all future arrivals have no priority over the customer of interest (the discipline of service is FCFS.)

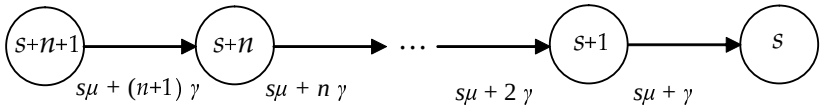

Figure 4.2: The customer $s+n+1$

The process starts from state $s+n+1$, i.e., all servers are busy and $n+1$ customers are waiting in queue (including the new arrival of interest.) The process moves from state $s+i$ to state $s+i-1$, $1 \leq i \leq n+1$, further to either a service completion with rate $s\mu$, or an abandonment with a rate equals to the number of waiting customers times the reneging rate, $i\gamma$. Consider the nth customer waiting in queue, and let $\psi_{n,1}$ be the probability that the process state moves down due to the reneging of that customer. One may see that

$$\psi_{n,1} = \frac{\gamma}{s\mu + n\gamma}. \tag{4.4}$$

Let $\Psi_{n,1}$ be the probability that the customer n in queue enters service. This event means that the customer does not renege in all positions in queue, starting from position n, until being served. Hence, the following holds

$$\Psi_{n,1} = \prod_{i=1}^{n} (1 - \psi_{i,1}). \tag{4.5}$$

Conditioning on a state seen by a new arrival and averaging thereafter over all possibilities, the probability of being served is given by

$$P_1(S) = \sum_{n=0}^{s-1} p_1(n) + \sum_{n=0}^{\infty} p_1(s+n) \cdot \Psi_{n+1,1}. \tag{4.6}$$

As for computing the probability of reneging, say $P_1(R)$, it suffices to take the complementary probability, $P_1(R) = 1 - P_1(S)$.

Let us now compute the first and second moments of the conditional waiting time given that service is completed. To do so, a direct method is quite complicated. We make a small

roundabout as in Whitt [135]. We define the random variable, say X_1, of the stationary waiting time in queue. We let X_1 be 0 if the customer reneges. The quantities we are looking for are thereafter $E(X_1 \mid S)$ and $E(X_1^2 \mid S)$. To compute the moments of X_1, we define a further random variable $X_{n,1}$ measuring the state-dependent duration to empty the queue of n waiting customers. Denoting by $E(X_{n,1})$ and $E(X_{n,1}^2)$ the respective first two moments of $X_{n,1}$, the mean of X_1 say $E(X_1)$, and its second moment say $E(X_1^2)$, are thereby given by $E(X_1) = \sum_{n=0}^{\infty} p_1(s+n) \times \Psi_{n+1,1} \times E(X_{n+1,1})$, and $E(X_1^2) = \sum_{n=0}^{\infty} p_1(s+n) \times \Psi_{n+1,1} \times E(X_{n+1,1}^2)$.

Consider a new arrival finding himself in the nth place in queue. The distribution of $X_{n,1}$ is the convolution of n independent exponential distributions with parameters $s\mu+\gamma$, $s\mu+2\gamma$, ..., and $s\mu+n\gamma$, which is an hypoexponential distribution. Hence, the first and second moments of $X_{n,1}$ are $\sum_{i=1}^{n} \frac{1}{s\mu+i\gamma}$ and $\sum_{i=1}^{n} \frac{2}{(s\mu+i\gamma)^2}$, respectively. We then deduce that

$$E(X_1) = \sum_{n=0}^{\infty} p_1(s+n) \cdot \Psi_{n+1,1} \cdot \left(\sum_{i=1}^{n+1} \frac{1}{s\mu+i\gamma} \right), \tag{4.7}$$

$$E(X_1^2) = \sum_{n=0}^{\infty} p_1(s+n) \cdot \Psi_{n+1,1} \cdot \left(\sum_{i=1}^{n+1} \frac{2}{(s\mu+i\gamma)^2} \right). \tag{4.8}$$

Now, we are ready to get the first and second moments of the conditional waiting time given that service is completed. Let $E(X_1 \mid R)$ and $E(X_1^2 \mid R)$ be the first and second moments of the conditional waiting time in queue given that service is not completed (customers renege), respectively. Since by construction of X_1 we have $E(X_1 \mid R) = E(X_1^2 \mid R) = 0$, we thereafter deduce that

$$E(X_1 \mid S) = \frac{E(X_1)}{P_1(S)}, \quad \text{and} \quad E(X_1^2 \mid S) = \frac{E(X_1^2)}{P_1(S)}. \tag{4.9}$$

As for its variance and standard deviation, they are given by $Var(X_1 \mid S) = E(X_1^2 \mid S) - E(X_1 \mid S)^2$ and $\sigma(X_1 \mid S) = \sqrt{Var(X_1 \mid S)}$, respectively.

The performance measures we derive in this section are used in Section 4.3.5 when addressing the comparison study between both models with and without announcement.

4.3.3 Impact of Announcing Delays

There is a modeling complexity when we give delay information to customers. The complexity comes from the change of customers behavior that may occur. In this section, we investigate the impact of announcing delays on the customer abandonment experience. When we inform one customer about his anticipated delay, he will decide from the beginning, either to hang up immediately because he estimates that his delay is too long, or to start waiting in queue. In the latter case, there are two further possibilities. The first is that all customers do never abandon

thereafter. The second possibility is that the customer patience will change, i.e., customers may abandon even if they had chosen to start waiting. It is natural to consider that customers would abandon under a different fashion of that in the original system (without announcement), depending on the information we provide to them. We refer the reader to Armony et al. [15] and Guo and Zipkin [51] for further details on the subject.

In concrete terms, we basically distinguish two new models that capture the change of customers behavior with respect to delays information. In the first model, a customer balks immediately upon arrival with probability $p_{bk}(n)$ depending on his anticipated delay or equivalently on the number n of customers ahead of him in queue. Once he decides to join the queue (with probability $1 - p_{bk}(n)$), he does never abandon thereafter. The second model is identical to the first one, however, we allow customers to renege even after having chosen to join the waiting line.

It is rather complicated to specify the appropriate model, because the information delay may have several different forms. So, we have to discern the human response for each possible kind of information. Whitt [135] considers the first model (no allowance for reneging in queue.) Assume a new arrival finding n customers in queue. He computes the quantity $p_{bk}(n)$ as the probability that the customer would abandon before a server becomes free for him (his virtual waiting time assuming he decides to join the queue.) Let T be the random variable measuring the random patience threshold of customers. If we denote by S_n the random variable of the virtual waiting time of the customer of interest, Whitt [135] stipulates that $p_{bk}(n) = P(T < S_n)$.

Recall that times before reneging are exponentially distributed with mean $1/\gamma$. Next, assuming that balking decisions of successive customers are independent, and denoting by $g_n(t)$ the probability density function (pdf) of S_n, we get

$$p_{bk}(n) = \int_0^\infty g_n(t) \cdot (1 - e^{-\gamma t})\, dt. \tag{4.10}$$

Since in the Whitt's Model there is no longer reneging in queue, then S_n has an $n+1$-Erlang distribution with parameter $s\mu$. Using the Laplace transform of that distribution with respect to the variable γ, Equation (4.9) becomes

$$p_{bk}(n) = 1 - \left(\frac{s\mu}{s\mu + \gamma}\right)^{n+1}. \tag{4.11}$$

Although the Whitt's model seems in the conceptual sense to be unfailing, it is at odds with reality. It is indeed not common to give one customer the distribution function of his anticipated delay so that he may decide to balk or not! For that reason, Salah et al. [6] and Whitt [135] propose (keeping no allowance for reneging) to communicate the expected delay $E(S_n)$, hence,

we would use the approximation $p_{bk}(n) \approx 1 - e^{-\gamma \cdot E(S_n)}$. From the Law of Large Numbers, this approximation should work well for large values of n. A further alternative is to communicate any other percentile, say β, of the distribution of S_n. Again, this is only exact when we stipulate that the customer acts as if the delay information was the actual delay, which is not the case. We should be careful when choosing the value of β. On the one hand, if β is too low, then arrivals tend to join the queue and it so happens that many of them must wait more than they were willing to do initially. On the other hand, if β is close to 100%, many customers balk while their delays do not exceed their patience threshold. As we shall see later in the numerical experiments, a value of β in the order of 90% or 95% should be a conceivable choice to preserve the reliability of the service provider.

To build an appropriate practical model, we should consider the second model in which we let customers abandon even they had elected to wait upon their arrival. In the following, we model that effect. In practice, the probability to renege should vary as a function of the announced delay. For a customer who joins the queue, the probability that he reneges is very low as long as his spending time in queue does not exceed the announced delay. However once the announced delay is no longer satisfied, customers become frustrated and loose their patience, which lead to a higher probability of reneging. For simplicity, we assume that times before reneging for customers who do not balk are i.i.d. and exponentially distributed with a new rate γ'. The system behaves as follows. Upon arrival, if less than s customers are in system, we do not do anything because the new customer gets service immediately. Otherwise, i.e., if all servers are busy and n waiting customers are ahead of the new customer, we derive the distribution of his virtual delay S_n, and we communicate him the delay which corresponds to a given coverage probability β. The distribution of S_n is no longer Erlang because of the reneging phenomenon. The time until a new arrival is scheduled to start service is the time it takes for the n customer ahead to leave the queue (either abandon or enter service) plus the time required for a service completion (when all servers are busy.) So, the virtual delay S_n can be characterized by a pure-death process as shown on Figure 4.3.

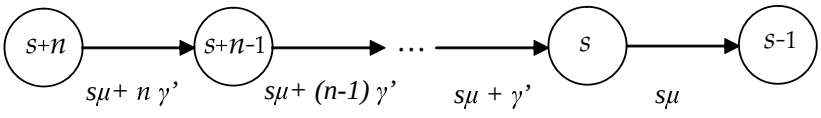

Figure 4.3: The random variable S_n

The random variable S_n represents the downcrossing time from state $n + s$ until absorption in state $s - 1$. Thus, the distribution of S_n is the convolution of $n + 1$ independent exponen-

tial distributions with parameters $s\mu$, $s\mu+\gamma'$, ..., and $s\mu+n\gamma'$, which is an hypoexponential distribution. Hence, the probability density function (pdf), $g_n(t)$, of S_n is given by

$$g_n(t) = \sum_{i=0}^{n} \left(\prod_{j=0,\, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \cdot \left(s\mu + i\gamma'\right) \cdot e^{-(s\mu+i\gamma')t}, \quad t > 0, \tag{4.12}$$

and its Probability Distribution Function (PDF), $G_n(t)$, is as follows

$$G_n(t) = \sum_{i=0}^{n} \left(\prod_{j=0,\, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \cdot \left(1 - e^{-(s\mu+i\gamma')t}\right), \quad t > 0. \tag{4.13}$$

Let D_n be the delay we communicate to the customer, $D_n = G_n^{-1}(\beta)$. As mentioned before, a reasonable value for β would be 95%. It means that the queueing delay of the new customer does not exceed D_n with 95% of chance. Whenever a new customer finds all servers busy and n waiting customers in queue, we stipulate that he decides to balk with the probability that his random patience threshold, T, would not exceed his anticipated delay D_n

$$p_{bk}(n) = P(T < D_n) = 1 - e^{-\gamma D_n}. \tag{4.14}$$

In practice, we should note that Relation (4.14) is an approximation of the customer behavior. It may happen that a new customer does not respect his patience threshold. For example, if he is willing to wait 1 min and we announce to him 1 min 2 sec, then he may join the queue because he estimates that 2 sec is a negligible duration. On the other hand, it may occur that if we announce to him 59 sec, he may be not logical so as he changes his mind and balks. Modeling the customer reaction is a difficult problem. We try here to be as near as possible to reality while still having tractable analysis so as we gain some useful guidelines.

Once the customer of interest elects to wait with probability $1-p_{bk}(n)$, he may renege within a random delay. As already assumed in Section 4.3.1, this random threshold is exponentially distributed with rate γ'. The resulting model referred to as Model 2 is shown on Figure 4.4

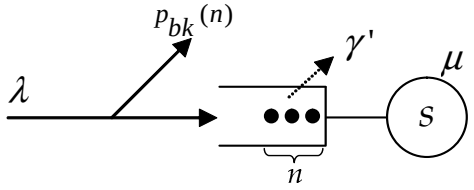

Figure 4.4: The new model incorporating delays announcement, Model 2

The remaining question now is how to compute the new reneging rate γ'. To answer this question, it is natural to relate the customer patience to the announced delay. To do so, we introduce the probability r_n for the customer in question. The quantity r_n is defined as the conditional probability that the queueing delay, S_n, exceeds the random patience, T, given that the customer joins the queue, i.e., given that the random patience exceeds the announced delay, D_n. For a given $\beta < 1$, r_n is seen as the non-zero probability that the customer of interest reneges.

$$r_n = P(T < S_n \mid T \geq D_n). \tag{4.15}$$

We present necessary details for calculating r_n in Section 4.3.4. Whenever we announce D_n based on a given coverage probability β, we shall make at worst a mistake with a chance of $1-\beta$. Let us give an explanation. Consider a new arrival finding n customers in queue. Assume that the customer is willing to wait T_n for service to begin. The duration T_n is a random realization of the random variable T. If $T_n > D_n$, then the customer joins the waiting line. After joining the queue, the probability that the time it takes for a server to become free (for the customer of interest) exceeds D_n is $1-\beta$. Since the customer is initially willing to wait up to T_n, hence knowing that $T_n > D_n$, the probability that the duration T_n passes before a server becomes free for our customer is less than $1-\beta$.

Assume we reach the stationary regime, and let λ^{ab} be the mean rate of abandoning customers. From the one hand, we have

$$\lambda^{ab} = \sum_{n=0}^{\infty} \lambda \cdot p_2(s+n) \cdot (1 - p_{bk}(n)) \cdot r_n, \tag{4.16}$$

where $p_2(i)$ is the stationary probability to have i customers present in Model 2 (in queue and in service), for $i \geq 0$. From the other hand, if we denote by $L_{q,2}$ the expected number of customers in queue, we can write

$$\lambda^{ab} = L_{q,2} \cdot \gamma'. \tag{4.17}$$

We shall give the expression of $L_{q,2}$ in Section 4.3.4. Next, combining Equations (4.16) and (4.17) would imply

$$\gamma' = \frac{\lambda}{L_{q,2}} \cdot \sum_{n=0}^{\infty} p_2(s+n) \cdot (1 - p_{bk}(n)) \cdot r_n. \tag{4.18}$$

The quantities $L_{q,2}$, $p_2(s+n)$ and r_n are functions of γ'. So, denoting the right hand side in Equation (4.17) by a continuous function f in γ', we may write $\gamma' = f(\gamma')$. As a consequence, we state that γ' is a point mapped to itself by the function f. In mathematical terms, γ' is said to be a fixed point of f.

There are numerous fixed point theorems in different parts of mathematics that describe the circumstances under which functions must have one or more fixed points. In this work, we do not discuss the existence of such solutions for f. We conjecture as it is often the case for similar problems of queueing theory that f has the necessary nice properties that guarantee the existence of a fixed point. This conjecture is experienced thereafter by several numerical experiments in Section 4.3.5. To compute numerically γ', we propose the following fixed point algorithm.

FIXED POINT ALGORITHM()

Initialization: $\gamma'^{(0)} \leftarrow \gamma,\ i \leftarrow 0,\ \epsilon$

Do

$i \leftarrow i+1$

$\lambda^{ab\,(i)} \leftarrow \lambda \times \sum_{n=0}^{\infty} p_2(s+n)(\gamma'^{(i-1)}) \times (1-p_{bk}(n)) \times r_n(\gamma'^{(i-1)})$

$L_{q,2}^{(i)} \leftarrow \gamma'^{(i-1)} \times \sum_{n=1}^{\infty} n \cdot p_2(s+n)(\gamma'^{(i-1)})$

$\gamma'^{(i)} \leftarrow \lambda^{ab\,(i)}/L_{q,2}^{(i)}$

While $\mid \gamma'^{(i)} - \gamma'^{(i-1)} \mid > \epsilon$

$\gamma' \leftarrow \gamma'^{(i)}$

END ALGORITHM.

4.3.4 Performance Measures with Announcement

In this section, we focus on deriving the performance measures for Model 2. To do so, we define a birth-death process as shown on Figure 4.5. Birth and death rates are both state-dependent. The new element here is that we have to take into account balking decisions when the process moves from state i to state $i+1$, for $i \geq s$, i.e., all servers are busy.

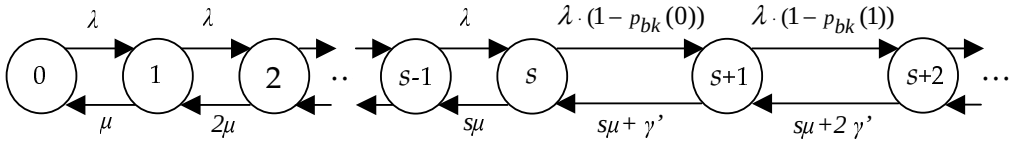

Figure 4.5: Birth-death process for Model 2

During the stationary regime, we get the following steady state probabilities

$$p_2(i) = \frac{\lambda^i}{i!\,\mu^i} \cdot p_2(0) \text{ for } 0 \leq i \leq s, \text{ and } p_2(i) = \left(\prod_{j=1}^{i-s} \frac{(1-p_{bk}(j-1))}{s\,\mu + j\,\gamma'} \right) \cdot \frac{\lambda^i}{s!\mu^s} \cdot p_2(0) \text{ for } i > s, \tag{4.19}$$

where $p_2(0)$ is the stationary probability to have no customers in system. It is given by

$$p_2(0) = \left(\sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i} + \frac{1}{s!\,\mu^s} \sum_{i=s+1}^{\infty} \left(\prod_{j=1}^{i-s} \frac{(1 - p_{bk}(j-1))}{s\,\mu + j\,\gamma'} \right) \cdot \lambda^i \right)^{-1}. \tag{4.20}$$

Hence, the probability of immediate service, $P_{is,2}$, the mean number of customers in queue, $L_{q,2}$, and the mean number of customers in system, $L_{s,2}$, can be calculated as provided by Equation (4.21).

$$P_{is,2} = \sum_{n=0}^{s-1} p_2(n), \quad L_{q,2} = \sum_{n=1}^{\infty} n \cdot p_2(s+n), \quad \text{and} \quad L_{s,2} = \sum_{n=1}^{\infty} n \cdot p_2(n). \tag{4.21}$$

Having the expression of $L_{q,2}$, it only remains for us to compute the quantity r_n so as we can apply the fixed point algorithm to get γ'. In what follows, we give a closed-form expression of r_n. By definition, Equation (4.15) may be rewritten as

$$r_n = \frac{P(D_n \leq T < S_n)}{P(T \geq D_n)}. \tag{4.22}$$

Since T is exponentially distributed with rate γ, the denominator in the right hand side of Equation (4.22) is simply

$$P(T \geq D_n) = e^{-\gamma D_n}. \tag{4.23}$$

As for the numerator, it is provided by

$$\begin{aligned} P(D_n \leq T < S_n) &= \int_{D_n}^{\infty} g_n(t) \cdot P(D_n \leq T < t)\, dt \\ &= \int_{D_n}^{\infty} g_n(t) \cdot \left(e^{-\gamma D_n} - e^{-\gamma t} \right)\, dt. \end{aligned} \tag{4.24}$$

Calculating further, we get

$$\begin{aligned} P(D_n \leq T < S_n) &= \left(\int_{D_n}^{\infty} g_n(t)\, dt \right) \cdot e^{-\gamma D_n} - \int_{D_n}^{\infty} g_n(t) \cdot e^{-\gamma t}\, dt \\ &= (1 - G_n(D_n)) \cdot e^{-\gamma D_n} - \int_{D_n}^{\infty} g_n(t) \cdot e^{-\gamma t}\, dt. \end{aligned} \tag{4.25}$$

Next, observing that

$$\begin{aligned} \int_{D_n}^{\infty} g_n(t) \cdot e^{-\gamma t}\, dt &= \sum_{i=0}^{n} \left(\prod_{j=0,\, j \neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \cdot \left(s\mu + i\gamma' \right) \cdot \int_{D_n}^{\infty} e^{-(s\mu+\gamma+i\gamma')t}\, dt \\ &= \sum_{i=0}^{n} \left(\prod_{j=0,\, j \neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \cdot \frac{s\mu + i\gamma'}{s\mu + \gamma + i\gamma'} \cdot e^{-(s\mu+\gamma+i\gamma')D_n}, \end{aligned} \tag{4.26}$$

and coming back to Equation (4.22), we deduce that

$$r_n = 1 - G_n(D_n) - \sum_{i=0}^{n} \left(\prod_{j=0,\, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \cdot \frac{s\mu + i\gamma'}{s\mu + \gamma + i\gamma'} \cdot e^{-(s\mu + i\gamma')D_n}. \tag{4.27}$$

Finally, using Equation (4.13) and simplifying further Equation (4.27) lead to

$$r_n = 1 - \sum_{i=0}^{n} \left(\prod_{j=0,\, j\neq i}^{n} \frac{s\mu + j\gamma'}{(j-i)\gamma'} \right) \cdot \left(1 - \frac{\gamma}{s\mu + \gamma + i\gamma'} \cdot e^{-(s\mu + i\gamma')D_n} \right). \tag{4.28}$$

Similarly to Section 4.3.2, the probability of being served, $P_2(S)$, is given by

$$P_2(S) = \sum_{n=0}^{s-1} p_2(n) + \sum_{n=0}^{\infty} p_2(s+n) \cdot (1 - p_{bk}(n)) \cdot \Psi_{n+1,2}, \tag{4.29}$$

where

$$\Psi_{n,2} = \prod_{i=1}^{n} \left(1 - \frac{\gamma'}{s\mu + i\gamma'} \right). \tag{4.30}$$

As for the probability of reneging, it is

$$P_2(R) = \sum_{n=0}^{\infty} p_2(s+n) \cdot (1 - p_{bk}(n)) \cdot (1 - \Psi_{n+1,2}). \tag{4.31}$$

Thereby, the probability of balking is simply

$$P_2(balking) = 1 - P_2(S) - P_2(R). \tag{4.32}$$

Now, we move on to compute the performance measures of interest. Following again a similar analysis as that in Section 4.3.2, we denote by X_2 the random variable of the stationary waiting time in queue. We let X_2 be 0 if the customer reneges. The first moment $E(X_2)$ and the second moment $E(X_2^2)$ of X_2 are given by

$$E(X_2) = \sum_{n=0}^{\infty} p_2(s+n) \cdot (1 - p_{bk}(n)) \cdot \Psi_{n+1,2} \cdot \left(\sum_{i=1}^{n+1} \frac{1}{s\mu + i\gamma'} \right), \tag{4.33}$$

$$E(X_2^2) = \sum_{n=0}^{\infty} p_2(s+n) \cdot (1 - p_{bk}(n)) \cdot \Psi_{n+1,2} \cdot \left(\sum_{i=1}^{n+1} \frac{2}{(s\mu + i\gamma')^2} \right). \tag{4.34}$$

Thus, the first moment, second moment, variance and standard deviation of the conditional

waiting time given that service is completed are

$$E(X_2 \mid S) = \frac{E(X_2)}{P(S)}, \quad E(X_2^2 \mid S) = \frac{E(X_2^2)}{P(S)},$$

$$Var(X_2 \mid S) = E(X_2^2 \mid S) - E(X_2 \mid S)^2, \quad \text{and} \quad \sigma(X_2 \mid S) = \sqrt{Var(X_2 \mid S)}, \text{ respectively.} \tag{4.35}$$

Note that the analysis reported in Sections 4.3.2, 4.3.3 and 4.3.4 can be easily extended to the case of a call center with a finite waiting line. A further easy extension is also a model in which a new customer may balk upon arrival whenever he has to wait (because all servers are busy.) For practical reasons, the first extension is not useful because queues capacities in call centers are usually large so as assuming infinite queues is a reasonable approximation. However, modeling the initial balking is of value in practice. It is recurrent in several call centers applications, and ignoring it may deviate the performance measures. In our work, we ignore that phenomenon because our purpose is specifically the investigation of advantages of delays information. So, including it would not affect our results in the qualitative sense.

4.3.5 Numerical Comparison

In this section, we conduct a comparison study for both models, with and without information about delays. We perform numerical experiments for various examples of call centers. The parameters of Model 1 are λ, s, μ and γ. Those of Model 2 are λ, s, μ, β and γ' (computed from γ).

It should be clear that a call center with delays information would improve customers satisfaction, and lead to concrete consequences. Hence, it would be incoherent to compare both models performances by ignoring the change on customers behavior (balking and reneging.) For instance, a high value of β (close to 1) may lead to a very good (low) waiting time in queue for Model 2 comparing to Model 1. The reason behind is not that Model 2 performs better, but it is due to the over-announcing in Model 2. The latter implies less accepted customers but more happy served customers. Thus, a comparison with regard to the throughput would on the contrary prefer Model 1 instead of Model 2. Alternatively, decreasing β should lead to less balking but more reneging. In the limit case (β close to 0), Model 2 would be the worst. Firstly, we will come back to the original drawback of Model 1, i.e., reneging due to frustration. Secondly, we will reduce the reliability of the company with regard to the correctness of the information it provides to customers.

To make a coherent comparison, we introduce penalty costs for lost customers. We define two different penalty costs C_1 and C_2. For each balking customer, the service provider pays C_1,

whereas a reneging customer does cost C_2. It is intuitively clear that $C_1 < C_2$. A customer who balks should have a higher probability to call back than that of a customer who reneges. A reneging customer leaves the system with frustration and loosing trust in the service provider. However, a balking customer leaves the system based on an information. This information would avoid to loose business because it is perceived by balking customers as an invitation to call back when the system will be able to serve them within a reasonable delay. Thereafter, introducing different costs for lost customers may be a way to translate this important issue. This method of comparison is a simplification of the reality so as we may illustrate the benefits of having a system with delays announcement. In practice, we may have for example a different cost for a customer who reneges in Model 1 from that in Model 2. We should note in addition that quantifying the costs parameters is a hard task.

We performed 20 numerical comparisons. We considered several sets of parameters to make sure of the robustness of the conclusions. For all cases, μ, γ, C_1 and C_2 are held constant. We let $\mu = 0.2$ and $\gamma = 1$. The abandoning cost is chosen to be twice greater than the balking cost, $C_1 = 1$ and $C_2 = 2$. To vary the staffing level, we consider models with $s = 5$, 10, 20, 50 and 100. The corresponding arrival rates are chosen so as the "service utilization", $\lambda/s\mu$, is 100%. Furthermore for each couple (λ, s), we vary the coverage probability β. We let $\beta = 50\%$, 70%, 90%, and 95%. Finally, the reneging rate for Model 2 is computed each time as we reported in Section 4.3.3 (as a function of β.) Numerical experiments for (λ, s) =(1,5), (2,10), (4,20), (10,50), and (20,100) are displayed in Tables 4.1, 4.2, 4.3 and 4.4, respectively. The line "Balking" is to indicate the stationary cost per unit of time (u.t) due to balking customers. It only concerns Model 2 (no balking in Model 1), and is calculated as $C_1 \times \lambda \times P_2(balking)$. The line "Reneging" is to indicate the cost per u.t for reneging customers. For Model 1 (Model 2), it is calculated as $C_2 \times \gamma \times L_{q,1}$ ($C_2 \times \gamma' \times L_{q,2}$).

	Model 1	Model 2			
		$\beta = 50\%$ $\gamma' = 0.271$	$\beta = 70\%$ $\gamma' = 0.128$	$\beta = 90\%$ $\gamma' = 0.033$	$\beta = 95\%$ $\gamma' = 0.015$
P_{is}	0.594	0.666	0.681	0.697	0.702
L_q	0.236	0.076	0.050	0.026	0.018
L_s	4.054	3.750	3.694	3.638	3.620
$P(S)$	0.764	0.735	0.729	0.722	0.720
$P(R)$	0.236	0.021	0.006	0.001	0.000
$P(balking)$	—	0.245	0.265	0.277	0.279
$E(X \mid S)$	0.124	0.048	0.033	0.018	0.013
$\sigma(X \mid S)$	0.301	0.210	0.178	0.131	0.112
Balking	—	0.245	0.265	0.277	0.279
Reneging	0.473	0.041	0.013	0.002	0.001
Total cost	0.473	0.286	0.278	0.278	0.280

Table 4.1: Numerical comparison for $s = 5$ and $\lambda = 1$

	Model 1	Model 2			
		$\beta = 50\%$ $\gamma' = 0.275$	$\beta = 70\%$ $\gamma' = 0.135$	$\beta = 90\%$ $\gamma' = 0.036$	$\beta = 95\%$ $\gamma' = 0.016$
P_{is}	0.625	0.689	0.708	0.731	0.740
L_q	0.342	0.159	0.121	0.079	0.064
L_s	8.634	8.276	8.187	8.081	8.041
$P(S)$	0.829	0.812	0.807	0.800	0.798
$P(R)$	0.171	0.022	0.008	0.001	0.001
$P(balking)$	—	0.166	0.185	0.198	0.202
$E(X \mid S)$	0.105	0.057	0.045	0.031	0.025
$\sigma(X \mid S)$	0.214	0.177	0.161	0.136	0.125
Balking	—	0.333	0.370	0.397	0.404
Reneging	0.683	0.088	0.033	0.006	0.002
Total cost	0.683	0.421	0.403	0.402	0.406

Table 4.2: Numerical comparison for $s = 10$ and $\lambda = 2$

	Model 1	Model 2			
		$\beta = 50\%$ $\gamma' = 0.256$	$\beta = 70\%$ $\gamma' = 0.127$	$\beta = 90\%$ $\gamma' = 0.034$	$\beta = 95\%$ $\gamma' = 0.015$
P_{is}	0.646	0.700	0.718	0.744	0.755
L_q	0.488	0.283	0.230	0.164	0.140
L_s	18.047	17.639	17.516	17.352	17.287
$P(S)$	0.878	0.868	0.864	0.859	0.857
$P(R)$	0.122	0.018	0.007	0.001	0.001
$P(balking)$	—	0.114	0.128	0.139	0.142
$E(X \mid S)$	0.085	0.054	0.045	0.034	0.029
$\sigma(X \mid S)$	0.140	0.125	0.118	0.105	0.099
Balking	—	0.456	0.514	0.557	0.568
Reneging	0.976	0.145	0.058	0.011	0.004
Total cost	0.976	0.601	0.572	0.568	0.573

Table 4.3: Numerical comparison for $s = 20$ and $\lambda = 4$

The end of this section is devoted to discuss the numerical results. We see that for each set of parameters, the probability of immediate service and that of reneging are better for Model 2. The reason is that Model 2, as it were, “refuses" entry for customers who potentially may renege. On the contrary, Model 1 does not care about them.

Some of the customers, who balk in Model 2, might wait in Model 1 until service begins without reneging. This implies that the probability of being served for a new arrival is better in Model 1, which agrees with the experiments. Less accepted customers in Model 2 makes the excepted numbers of customers in queue and in system larger for Model 1 than those for Model 2. In addition since we let reneging in Model 2, less customers enter service in the latter. As a consequence, the conditional mean waiting time given service is lower in Model 2. From experiments, we deduce that the conditional standard deviation of the waiting time given service is also lower in Model 2. One possible explanation may be related to the fact that realizations of waiting time values are the lower for Model 2. Note that the probability of being served also

	Model 1	Model 2			
		$\beta = 50\%$ $\gamma' = 0.220$	$\beta = 70\%$ $\gamma' = 0.109$	$\beta = 90\%$ $\gamma' = 0.029$	$\beta = 95\%$ $\gamma' = 0.013$
P_{is}	0.663	0.704	0.721	0.746	0.758
L_q	0.777	0.535	0.458	0.354	0.313
L_s	46.893	46.416	46.239	45.986	45.879
$P(S)$	0.922	0.918	0.916	0.913	0.911
$P(R)$	0.078	0.012	0.005	0.001	0.000
$P(balking)$	—	0.071	0.079	0.086	0.088
$E(X \mid S)$	0.060	0.044	0.038	0.030	0.027
$\sigma(X \mid S)$	0.072	0.069	0.067	0.063	0.060
Balking	—	0.706	0.794	0.863	0.883
Reneging	1.553	0.235	0.100	0.020	0.008
Total cost	1.553	0.941	0.894	0.884	0.891

Table 4.4: Numerical comparison for $s = 50$ and $\lambda = 10$

	Model 1	Model 2			
		$\beta = 50\%$ $\gamma' = 0.191$	$\beta = 70\%$ $\gamma' = 0.094$	$\beta = 90\%$ $\gamma' = 0.025$	$\beta = 95\%$ $\gamma' = 0.011$
P_{is}	0.672	0.704	0.719	0.743	0.754
L_q	1.101	0.828	0.729	0.587	0.527
L_s	95.598	95.062	94.837	94.501	94.353
$P(S)$	0.945	0.942	0.941	0.939	0.938
$P(R)$	0.055	0.008	0.003	0.001	0.000
$P(balking)$	—	0.050	0.055	0.060	0.061
$E(X \mid S)$	0.046	0.035	0.032	0.026	0.023
$\sigma(X \mid S)$	0.039	0.040	0.040	0.039	0.038
Balking	—	0.995	1.110	1.203	1.229
Reneging	2.201	0.316	0.137	0.029	0.012
Total cost	2.201	1.311	1.247	1.232	1.241

Table 4.5: Numerical comparison for $s = 100$ and $\lambda = 20$

indicates the throughput of both systems. Although more customers are lost in Model 2 than in Model 1, the difference is not significant, but only a bit worse for Model 2. From the experiments, the relative differences of the throughput are ranging from less than 1% up to 5%. Calculating thereafter the costs due to lost customers based on the parameters C_1 and C_2, makes Model 2 better than Model 1.

Consider now Model 2 by varying the coverage probability β. As we would expect, increasing β leads to more balking. Such change would better satisfy customers who elect to wait. The numerical results confirm that intuition; P_{is}, $P(R)$, $E(X \mid S)$ and $\sigma(X \mid S)$ improve, and $P(S)$ deteriorates as β increases. Increasing β leads to more lost customers because customers who balk in Model 2 may not renege when joining the queue (which happens for low values of β). The flow of lost customers does not differ much by varying β. Based on the parameters C_1 and C_2 we chosen, better costs are reached for $\beta = 90\%$.

Let us now focus on the comparison of both models as λ and s grow (keeping the load

constant.) We see that the conclusions are pretty much of the same quality. In addition, performances of both models improve. It is expected due to the idea of resource sharing discussed by Smith and Whitt [120]. We also notice from experiments that the relative decrease of cost, when moving from Model 1 to Model 2, increases in the system size.

4.4 Two-Class Call Center with Priority

In this section, we investigate the extension of the analysis to the case of a call center with two customer classes. We consider a call center model with two impatient customer classes, high and low priority classes. The priority rule is strict and non-preemptive. We believe that qualitative results should not differ from those derived in Section 4.3. Thereafter, we do not repeat the comparison analysis between models with and without announcement. However, we focus on building quantitatively the call center model when incorporating delays announcement. This step is particularly complex in the multiclass case because of the priority rule.

In Section 4.4.1, we present the original call center model without announcement. In Section 4.4.2, we specify the new model by including delays information. In a similar fashion to that in Section 4.3, we announce to each new arrival his virtual delay based on a given coverage probability β. The distribution of the waiting time for low customers is not straightforward. For simplicity issues, we assume no reneging once customers elect to wait. This simplification would work well when choosing high values of β (in the order of 90%.) In Section 4.4.3, we compute the first two moments of the virtual delays distributions for new arrivals. By doing so, we complete specificating the new model with announcement.

4.4.1 Two-Class Model without Announcement

Consider the queueing model of a call center with two classes of customers; important customers type A, and less important ones type B. The model consists of two infinite priority queues type A and B, and a set of s parallel, identical servers representing the set of agents. All agents are able to answer all types of customers. The call center is operated in such a way that at any time, any call can be addressed by any agent. So upon arrival, a call is addressed by one of the available agents, if any. If not, the call must join one of the queues. The scheduling policy of service assigns customers A (B) to queue A (B). Customers in queue A have priority over customers in queue B in the sense that agents are providing assistance to customers belonging to queue A first. The priority rule is non-preemptive, which simply means that an agent currently serving a customer pulled from queue B, while a new arrival joins queue A, will complete this service before turning to queue A customer. Within each queue, customers are served in FCFS

manner.

Arrival processes of type A and B customers follow a Poisson process with rates λ_A and λ_B, respectively. Let λ_T be the total arrival rate, $\lambda_T = \lambda_A + \lambda_B$. Successive service times are assumed to be i.i.d., and follow a common exponential distribution with rate μ for both types of customers. The motivation for considering a common service time distribution is similar to that reported for the models of Chapter 3. We let customers be impatient. Times before reneging for both types are assumed to be i.i.d., and exponentially distributed with a common rate γ for both customers types. Again, we note that abandonments make our system unconditionally stable. The resulting model, referred to as Model 3, is shown on Figure 4.6.

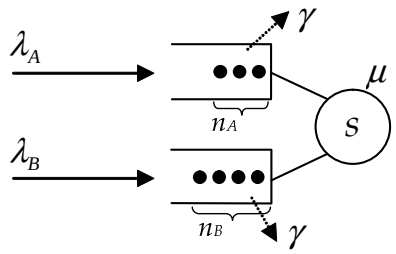

Figure 4.6: The original two-class model, Model 3

4.4.2 Two-Class Model with Announcement

Assume moving from the call center described in Section 4.4.1 to a call center with delays announcement. For each new arrival (type A or B), the service provider gives a percentile β of the distribution of the state-dependent virtual delay. The state-dependent virtual delay is the time it takes for a server to become free for the customer of interest. In other words, it is the time until all higher priority customers ahead of the arrival leave the queue plus the duration of a service completion. Paralleling to Section 4.3.3, we stipulate that a new customer elects to join the queue with the probability that a server becomes free for him before he would renege. For simplicity sake, we do not let customers renege once they join the waiting line. This is reasonable for high values of β provided that the estimation of the anticipated delay should be fairly accurate in that case, so that ignoring reneging would be valid. Assume that a new arrival finds n_A waiting type A customers in queue A, and n_B waiting type B customers in queue B. Note that implicitly we are focusing on new arrivals finding all servers busy. If the number seen by an arrival is less than s, then the new arrival does never balk and enters service immediately. Let us come back to a new arrival finding all servers busy. It should be clear that the probability of balking for a type A new arrival does depend only on n_A (due to the priority rule), $p_{bk}^A(n_A)$.

However, the probability of balking for a new type B arrival does depend on the couple (n_A, n_B), $p^B_{bk}(n_A, n_B)$. Furthermore, we should not fall in the confusion of only considering it as a function of $n_T = n_A + n_B$. Having different values of n_A and n_B, so as $n_T = n_A + n_B$ is held constant, would affect the virtual delay distribution of the customer of interest. The reason is that, under the model with announcement, the arrival rate of type A customers seen by our new type B customer, is state of queue A dependent. As a consequence, not considering the couple (n_A, n_B) to compute the balking probability of that customer would lead to a wrong result.

Let $Y^A_{n_A}$ be the random variable measuring the state-dependent virtual delay for a new type A arrival finding n_A waiting customers ahead of him. Let $Y^B_{(n_A,n_B)}$ be that for a new type B arrival finding n_A and n_B waiting customers ahead of him in queues A and B, respectively. Furthermore, let $G^A_{n_A}(t)$ and $G^B_{(n_A,n_B)}(t)$ for $t > 0$ be the PDF of $Y^A_{n_A}$ and $Y^B_{(n_A,n_B)}$, respectively. Then, the call center provides upon arrival the values $D^A_{n_A} = (G^A_{n_A})^{-1}(\beta)$ and $D^B_{(n_A,n_B)} = (G^B_{(n_A,n_B)})^{-1}(\beta)$ to type A and B customers, respectively. We calculate the balking probabilities as explained in Section 4.3.3. Hence, denoting again by T the random threshold patience for both types (exponentially distributed with rate γ), we have $p^A_{bk}(n_A) = P(T < D^A_{n_A})$ and $p^B_{bk}((n_A, n_B)) = P(T < D^B_{(n_A,n_B)})$. Next, the following holds

$$p^A_{bk}(n_A) = 1 - e^{-\gamma \cdot D^A_{n_A}}, \quad \text{and} \quad p^B_{bk}((n_A, n_B)) = 1 - e^{-\gamma \cdot D^B_{(n_A,n_B)}}. \tag{4.36}$$

The resulting model is shown on Figure 4.7, and is referred to as Model 4.

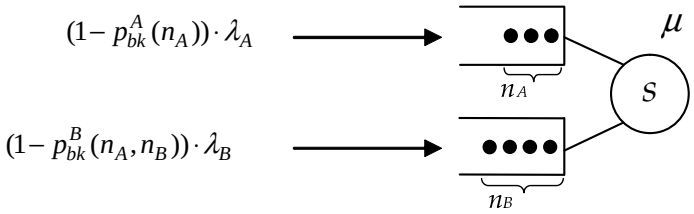

Figure 4.7: The new two-class model incorporating announcing, Model 4

What remains to be done in order to characterize Model 4 is to compute state-dependent arrival rates for each customer type, which in turn reduces to characterize the distribution functions of $Y^A_{n_A}$ and $Y^B_{(n_A,n_B)}$. In the next section, we give closed-form expressions for their first two moments. Based on these results, we thereafter propose in Section 4.5.1, a helpful approximation of their whole distributions.

4.4.3 Estimating Virtual Delays

In Sections 4.3.2 and 4.3.4, we assume that the equipment technology of our call center enables us to know when queues are empty, whether there is one available agent or not for an upcoming customer.

If less than s customers are present in system, the customer of interest gets service immediately. If not, he has to wait in his corresponding queue for service to begin. Knowing that all servers are busy, we focus on the random variables $Y^A_{n_A}$ and $Y^B_{(n_A,n_B)}$, $n_A, n_B \geq 0$. We separate the analysis depending on whether the arrival call is of type A or B. Type A customers observe a regular queue without priority. So, estimating their waiting time is easy to obtain. However that of type B customers is more complicated to compute because it is affected by future type A arrivals (with higher priority.)

Let us recall that we are calculating virtual delays which will be used within a second step in order to compute balking probabilities. In other words, we are calculating the time it takes until a server becomes free for the customer of interest in case he elects to wait (does not balk.)

Virtual Delays for Type A Customers

Consider a new type A arrival who finds all servers busy, n_A waiting customers in queue A and n_B waiting customers in queue B. Owing to his higher priority, the virtual delay of a new type A arrival does not depend on the number of type B customers already present in system, see Figure 4.8. The customer has to wait until the n_A waiting customers leave the queue plus the time it takes for a service completion (when all servers are busy.) By a customer who leaves the queue, we only mean a customer who enters service. In Model 4, there is no longer possibility for customers to renege once they join the queue.

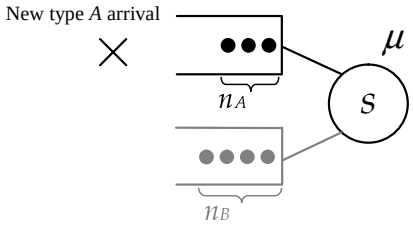

Figure 4.8: Virtual delay for a new type A arrival

Hence, the pdf of $Y^A_{n_A}$ is simply the convolution of the pdfs of n_A+1 i.i.d. exponential random variables each with parameter $s\mu$. So, $Y^A_{n_A}$ has an n_A+1-Erlang distribution with parameter $s\mu$.

The mean and variance of $Y^A_{n_A}$ are, respectively, given by

$$E(Y^A_{n_A}) = \frac{n_A + 1}{s\mu}, \quad \text{and} \quad Var(Y^A_{n_A}) = \frac{n_A + 1}{s^2\mu^2}. \tag{4.37}$$

Having in hand the PDF of $Y^A_{n_A}$, it only remains to come back to Equation (4.36) in order to compute the balking probability $p^A_{bk}(n_A)$. Define now the standard deviation of $Y^A_{n_A}$ by $\sigma(Y^A_{n_A}) = \sqrt{Var(Y^A_{n_A})}$, and the coefficient of variation by the ratio of the standard deviation over the mean, $cv(Y^A_{n_A}) = \sigma(Y^A_{n_A})/E(Y^A_{n_A})$. As shown on Equation (4.38), the ratio $cv(Y^A_{n_A})$ is characterized to have simple form independent of μ and s.

$$cv(Y^A_{n_A}) = \frac{1}{\sqrt{n_A + 1}} \tag{4.38}$$

From Equation (4.38), we deduce that for large values of n_A, the virtual delay of $Y^A_{n_A}$ is very concentrated about its mean. This implies that for large values of n_A, the mean value of $Y^A_{n_A}$ should provide a good approximation of the virtual delay.

Virtual Delays for Type B Customers

Knowing that all servers are busy, let n_A and n_B be the number of type A and B waiting customers seen by a new type B arrival, in queues A and B, respectively.

The random variable $Y^B_{(n_A,n_B)}$ is the time until the $n_T = n_A + n_B$ waiting customers start service, plus the time it takes for all future type A arrivals (during the waiting of the customer of interest) to enter service, plus the duration for a service completion (when all servers are busy), see Figure 4.9.

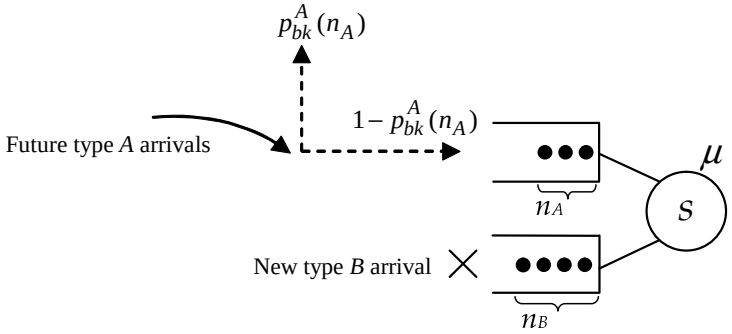

Figure 4.9: Virtual delay for a new type B arrival

To characterize $Y^B_{(n_A,n_B)}$, we ignore all future type B arrivals because the discipline of service within queue B is FCFS. However, all future type A arrivals have to be considered because of

their higher priority against the customer of interest. Recall that reneging is no longer possible. We only consider events of type A arrivals and service completions. Thereby, changes of queues states seen by our customer are as follows. As long as type A customers are waiting in queue, the number of type B waiting customers does not change, however that of type A customers increases by 1 further to a type A arrival or decreases by 1 further to a service completion. The number of type B waiting customer can not increase. It only decreases by 1 further to a service completion when no type A customers are waiting in queue. We should be careful to not forget that type A arrivals are state-dependent due to balking decisions of customers upon arrival.

Based on the above explanation, we move on to employ the following two-dimensional Markov chain. Let the system state at a given random instant be (m_A, m_B) where m_A (m_B) is the number of type A (B) customers in queue A (B), $m_A, m_B \geq 0$. In addition, the Markov chain has an absorbing state denoted by (-1). The system moves to (-1) further to a service completion when both queues are empty. Being in the latter state means that a server is available for the customer of interest. When m_A customers are waiting in queue A, we denote the state-dependent arrival rate of type A arrivals by $\lambda_A(m_A) = (1 - p^A_{bk}(m_A)) \times \lambda_A$. The non-zero transition rates are $q_{(m_A,m_B)(m_A+1,m_B)} = \lambda_A(m_A)$, for $m_A, m_B \geq 0$, $q_{(m_A,m_B)(m_A-1,m_B)} = s\mu$, for $m_A, m_B > 0$, $q_{(0,m_B)(0,m_B-1)} = s\mu$, for $m_B \geq 0$, and $q_{(0,0)(-1)} = s\mu$. As shown on Figure 4.10, measuring $Y^B_{(n_A,n_B)}$ may be formulated as to calculate the downcrossing time until absorption in state (-1), starting from state (n_A, n_B).

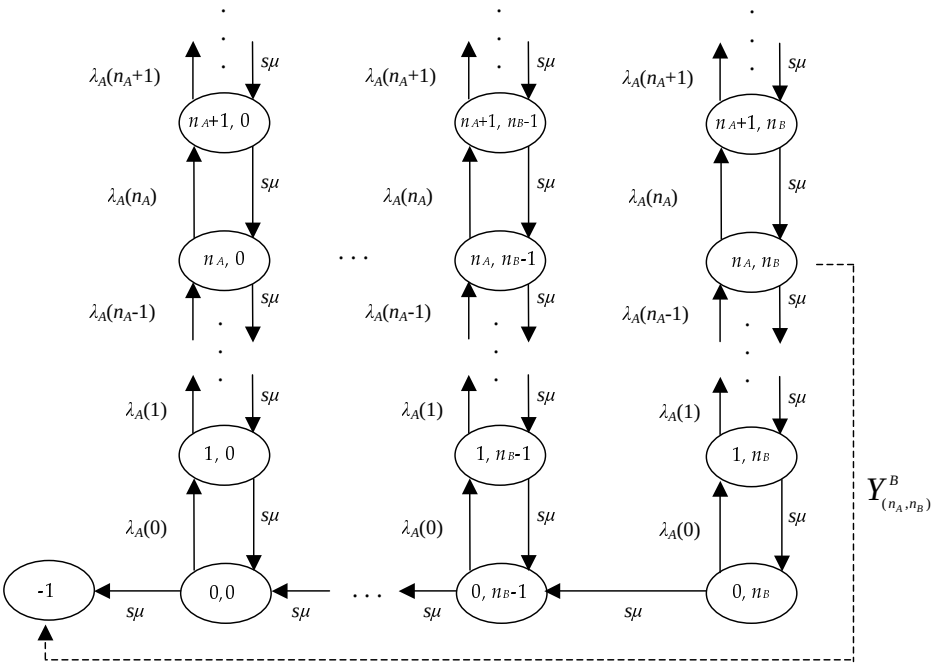

Figure 4.10: The random variable $Y^B_{(n_A,n_B)}$

The Markov chain we consider has a special structure allowing analytical solutions. Let us give an explanation. From Figure 4.10, the random variable $Y^B_{(n_A,n_B)}$ may be rewritten as

$$Y^B_{(n_A,n_B)} = U(n_A) + V_{n_B-1} + ... + V_0, \tag{4.39}$$

where $U(n_A)$ is the random variable measuring the downcrossing time until first passage at state $(0, n_B - 1)$ starting from state (n_A, n_B), V_i is the random variable measuring the downcrossing time until first passage time at state $(0, i-1)$ starting from state $(0, i)$ for $1 \leq i \leq n_B - 1$, and V_0 is the random variable measuring the downcrossing time until absorption in state (-1) starting from state $(0, 0)$.

The Markovian assumptions allow us to state that the random variables $U(n_A)$, V_0, ..., and V_{n_B-1} are independent. From Figure 4.10, we see that V_0, ..., and V_{n_B-1} are identically distributed. Let $E(Y^B_{(n_A,n_B)})$ and $Var(Y^B_{(n_A,n_B)})$ be the mean and variance of the random variable $Y^B_{(n_A,n_B)}$, respectively. Then, using the linearity property of expectations, we get

$$E(Y^B_{(n_A,n_B)}) = E(U(n_A)) + n_B \times E(V_0), \tag{4.40}$$

and from the independence, the following holds

$$Var(Y^B_{(n_A,n_B)}) = Var(U(n_A)) + n_B \times Var(V_0). \tag{4.41}$$

Let us now focus on computing means and variances of $U(n_A)$, V_0, ..., and V_{n_B-1}. To do so, we define an intermediate birth-death process with discrete state space taking non-negative integer values $\{0, 1, 2, 3, ...\}$. The transition rates of the process are denoted by

$$q_{0,1} = \lambda_A,\ q_{m,m+1} = \lambda_A(m-1) \text{ and } q_{m,m-1} = s\mu \text{ for } m \geq 1, \text{ and } q_{m,n} = 0 \text{ otherwise.} \tag{4.42}$$

The birth-death process is derived from the previous Markov chain and is shown on Figure 4.11. One may see that $U(n_A)$, V_0, ..., and V_{n_B-1} may be defined on the intermediate birth-death process. The random variable $U(n_A)$ is the downcrossing time until first passage at state 0, starting from state $n_A + 1$. As for the random variable V_i, $0 \leq i \leq n_B - 1$, it is only the first passage time at state 0, starting from state 1. Note that since we are calculating first passage times at state 0, the analysis is independent of the birth rate when the system is in that state, i.e., $q_{0,1} = \lambda_A$.

By considering a general birth-death process, the authors in Jouini and Dallery [65] give closed-form expressions for any moment of order $k \geq 1$ of several random variables related to

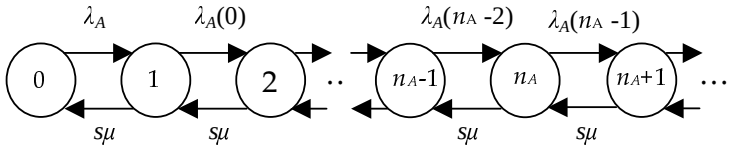

Figure 4.11: Intermediate birth-death process

first passage times. Further details about this analysis is given in Chapter 5. We use their results in our context here. To simplify the presentation, we introduce the quantities δ_m for $m \geq 0$. For $m = 0$ we let $\delta_0 = \lambda_A$, and for $m \geq 1$ we let $\delta_m = \lambda_A(m-1)$. Let us now define the potential coefficients of the intermediate birth-death process, say ϕ_m, as follows.

$$\phi_0 = 1, \text{ and } \phi_m = \frac{\prod_{j=0}^{m-1} \delta_j}{s^m \mu^m}, \text{ for } m \geq 1. \tag{4.43}$$

Thereafter, the mean $E(U(n_A))$ and variance $Var(U(n_A))$ of $U(n_A)$ are given by

$$E(U(n_A)) = \sum_{m=1}^{n_A+1} \frac{1}{\delta_{m-1}\phi_{m-1}} \sum_{j=m}^{\infty} \phi_j, \tag{4.44}$$

$$Var(U(n_A)) = \sum_{m=1}^{n_A+1} \frac{2}{\delta_{m-1}\phi_{m-1}} \sum_{j=m+1}^{\infty} \frac{1}{\delta_{j-1}\phi_{j-1}} \left(\sum_{l=j}^{\infty} \phi_l \right)^2 + \sum_{m=1}^{n_A+1} \frac{1}{\delta_{m-1}^2 \phi_{m-1}^2} \left(\sum_{j=m}^{\infty} \phi_j \right)^2. \tag{4.45}$$

It goes without saying that δ_0 (that is λ_A) could be eliminated from Equations (4.44) and (4.45), which agrees with our above claim. We only keep them here for presentation issues.

Concerning the mean $E(V_0)$ and variance $Var(V_0)$ of the random variable V_0, they are given by

$$E(V_0) = \frac{1}{\delta_0} \sum_{m=1}^{\infty} \phi_m, \tag{4.46}$$

$$Var(V_0) = \frac{2}{\delta_0} \sum_{m=2}^{\infty} \frac{1}{\delta_{m-1}\phi_{m-1}} \left(\sum_{j=m}^{\infty} \phi_j \right)^2 + \frac{1}{\delta_0^2} \left(\sum_{m=1}^{\infty} \phi_m \right)^2. \tag{4.47}$$

Substituting Equations (4.44), (4.45), (4.46) and (4.47) back into Equations (4.40) and (4.41) leads to the expressions of the mean and variance of the random variable $Y^B_{(n_A,n_B)}$. Finally, using the results of Section 4.4.3 to compute the balking probabilities for type A arrivals, the mean and variance of $Y^B_{(n_A,n_B)}$ are thereafter fully characterized.

Note that one may derive all higher order moments of the virtual delay for both customer types, which allows us to derive their full distributions. However, the analysis would be cum-

bersome and numerically time consumer. We thereafter content ourself with only the first two moments. By way of compensation, we propose a useful approximation of distributions as we shall explain later in Section 4.5.1.

4.5 Some Practical Issues

In this section, we investigate some practical issues for an eventual implementation of delays information. We distinguish two points that may help practitioners. The first is discussed in Section 4.5.1, and concerns an approximation for computing the anticipated delay we communicate to each new arrival. The second is discussed in Section 4.5.2, and is related to the way of communicating that virtual delay.

4.5.1 Normal Approximation of Virtual Delays

Given the system state upon each arrival and given a coverage probability β, the service provider has to compute the value of the anticipated delay. This numerical computation operation is characterized to be too heavy. In fact in the case of a single class call center, the virtual delay distribution (hypoexponential) has an alternate summation of terms with different signs. As for the low priority type in the two-class case, exact moments expressions for the virtual delay distribution involve infinite summations. This would imply several numerical difficulties especially that we are asked to conduct such real-time operations for each arrival!

From a practical point of view, a normal distribution provides a satisfactory approximation of virtual delays. Since the random variables of virtual delays we consider here deal with summations of independent random variables, the Normal approximation should works well, see Whitt [137] and Ward and Whitt [129]. This claim is supported by theoretical results based on the Law of Large Numbers and the Central Limit Theorem. The normal approximation should work well especially for new arrivals who find large number of waiting customers in queue.

We only need the mean and standard deviation of the state-dependent virtual delay in order to get its full distribution (approximately.) Thus, we propose to use the normal distribution by only picking up the means and variances we derived in Sections 4.3 and 4.4. We should however point out that for small values of β, such distributions may lead to negative values of anticipated delays. To be judicious, we may adapt a given normal distribution by truncated it so as it becomes no longer defined for negative values. For instance, let $h(t)$ and $H(t)$, $-\infty < t < +\infty$, be the pdf and PDF of the original normal distribution, respectively. Also, let $h_{tr}(t)$ and $H_{tr}(t)$, $0 < t < +\infty$, be those for the truncated normal distribution. The pdf of $h_{tr}(t)$ is calculated as $h_{tr}(t) = \frac{h(t)}{1-H(0)}$. By doing so, we even out the area of the negative region ($t < 0$) over that of the

positive region, so that we build a correct distribution. Note that this transformation should not really affect the original normal distribution. The reason is that the quantity $1 - H(0)$ is low for very small numbers of waiting customers in queue, and may be reasonably neglected otherwise.

4.5.2 Announcing About Anticipated Delays by Increments

Once the anticipated delay is computed, the next natural question is how we should profit from that information? In what follows, we briefly discuss some elements addressing this issue.

There is not a single best method for all circumstances. In practice, we do not think that it is interesting for the customer to get a high accuracy of his anticipated delay. Regarding to the customer reaction, we believe that there is no real difference for him between being informed about for example a delay of 1.64 min or 1.79 min. Furthermore, in most call centers cases, there is no need to tell customers large delays. If the anticipated delay is for example greater than 5 min, we should just inform the customer that he will wait more than 5 min for service to begin. A delay beyond 5 min seems to be excessive in most call centers cases. So, any value in this range will be perceived by customers in the same fashion.

The idea we propose here is to inform about delays by increments. For instance, we define the following increments: < 30 sec (we tell the customer that his delay will be less than 30 sec), < 1 min, < 2 min, < 3 min, < 4 min, < 5 min, and a final increment > 5 min. Hence for each new arrival we look, among the predefined intervals, for the increment in which the computed value of the virtual delay is belonging. The obtained increment is thereafter communicated to the customer so as he makes the decision to balk or not.

A major advantage of this method should be noted. In cases when the estimated delay and the real one are in the same interval of time, the error do not exist. This often happens since the coverage probability β is high. For example, if the first is 1.21 min and the second is 1.85 min, then we will announce that the delay is less than 2 min. So, the approximated delay and the real one coincide regarding to the increment we communicate to the customer. We expect that this effect would considerably reduce customers reneging.

An other idea would be to announce a lower bound and an upper bound of the estimated delay, instead of only an upper bound. The idea is of value when the PDF of the conditional waiting time is too close to zero for small times values. This case often happens for a new arrival call who finds a large number of customers in queues.

4.6 Conclusions and Further Extensions

We focused on a fundamental problem in the operations management of call centers, namely the issue of informing customers about their queueing delays. Predicting delays is especially important when customers do not have direct access to information about the state of the system. In such a case, it has been indeed recognized that customers become dissatisfied with the service provider when they are forced to wait for an unknown delay. Announcing delays would reduce the undesirable phenomenon of customers reneging in queue, which allows to decrease the variability of waiting times. As a consequence, we improve the quality of delays information we give to customers, which in turn reduces even more customers reneging, and so on.

A central feature of this work is the way we model a call center with impatient customers and incorporating delays information. We proposed an extension of the model in Whitt [135]. In the latter, the author assumes no reneging once a new arrival elects to join the queue. This may not be the case since the communicated delay based on which the customer makes his balking decision is estimated with a percentage error. We gave a method to estimate the new reneging behavior even if customers decide not to balk upon arrival.

In the first part of the chapter, we investigated the effect of giving anticipated delays to customers in a single class call center. We computed several performance measures for both models, with and without information about queueing delays. A numerical comparison is thereafter conducted in order to prove how customers satisfaction would make the model with delays information preferable. In the second part of the chapter, we tackled the extension of the analysis to an interesting case in practice, namely a two-class call center with priority. We focused for that case on building the new model incorporating delays information. We specifically characterized the distributions of new arrivals virtual delays. Finally, we discussed some practical issues that would help practitioners for the implementation step. We proposed a useful approximation for virtual delays. Next, we discussed a practical way for communicating anticipated delays. The method of announcement is referred to as announcement by increments. It has also been shown that this method would improve the system behavior through reducing errors approximations.

In a future study, it would be interesting to describe empirically customers reaction in response to delays announcement, in order to validate our stipulation here for modeling that reaction. It would also be of value to assess in a real call center case our claim regarding to the advantages of announcing delays by increments. An ambitious extension is to consider non-stationary arrivals which would be helpful in practice. A further direction is analyzing more complex systems: more than two customer classes, general distributions for service times and times before reneging, etc.

To make a coherent comparison between both models performances, we introduced different costs for different kinds of lost customers (due to balking or reneging.) Another way of value for comparison is to model the changes that may occur on system parameters. It should be clear that incorporating delays information would increase business. Thereafter, one may let lost customers call back. However, it is natural to consider different behaviors of retrials for different kinds of lost customers. Concretely, we may assume that balking customers have a higher probability to call back than that of reneging customers. We leave this work for a future research.

Chapter 5

Moments of First Passage Times in General Birth-Death Processes

The topic addressed in this chapter is related to the field of birth-death processes. We consider ordinary and conditional first passage times in a general birth-death process. Under existence conditions, we derive closed-form expressions for the kth order moment of the defined random variables, $k \geq 1$. We also give an explicit condition for a birth-death process to be ergodic degree 3. To the best of our knowledge, several results of this chapter are not given beforehand. Based on the obtained results, we next analyze some interesting applications for markovian queueing systems such as call centers models.

The paper version of this chapter, Jouini and Dallery [65], was submitted for publication.

5.1 Introduction

This chapter deals with the analysis of birth-death processes. It does not directly address some call centers issues as in the previous chapters. However, it provides useful applications for the performance analysis of call center queueing models. Some of these applications was already used for example in Chapter 4 in order to compute state-dependent queueing delays.

Birth-death processes, and in general Markov chains, are broadly used in the field of queueing theory. They are a rich and important class in modeling numerous phenomena in queueing systems. For instance, birth-death processes allow to account for customers balking and reneging. The analytical studies were intended to obtain useful information for the decision making process, basically related to the design, the control, and the measurement of effectiveness of the systems. Whitt [133] and Kelly [79] have used birth-death processes to study complex networks of service facilities. In the present chapter, we are dealing with the transient (time-dependent) analysis of birth-death processes. The characteristics of interest are the ordinary and conditional first passage times. These characteristics are known to be of value for the performance evaluation of several queueing systems. Analyzing either transient or stationary queueing delays and response times, for example, may be addressed by means of ordinary and conditional first passage times.

For the time-dependent solutions, advanced mathematical techniques are necessary. The well studied systems are the simple ones, namely, $M/M/1/K$, $M/M/1$, and $M/M/\infty$, see Kleinrock [81], and Gross and Harris [46]. Such solutions are due first to Bailey [20]'s work where the author has solved the partial differential equations governing the underlying birth-death process via generating functions. Another interesting approach, based on advanced combinatorial methods, was done by Champernowne [33]. In general, most of the popular procedures derive the transient expressions using a combination of generating probability functions and Laplace transforms, see Abate and Whitt [3], Parthasarathy [107], and Abate and Whitt [4] for an overview. In these papers, the numerical solutions are complex due to the use of the Bessel functions. Some other approaches were proposed as a method based on the Taylor series in Krinik and Sourouri [89], and a new method based on the uniformization technique and on generating functions proposed by Leguesdron et al. [92]. The last method is of interest in the sense that it leads to quite simple expressions for the transient probabilities. For the the $M/M/1/K$, Tarabia [123] gave an alternative simple approach to the procedure of Takâcs [122]. He showed that the measures of effectiveness such as the first and the second order moments of the queue length can be easily obtained in a new and elegant closed-form. The result was also derived for the $M/M/1$ case by taking the limit as $K \to \infty$.

The literature specifically related to birth-death processes is extensive and growing, see Karlin

and McGregor [71] and [72], Keilson [76] and [77], Sumita [121], Mao [99], Guillemin [49], and Coolen-Schrijner and van Doorn [35]. We refer the reader to Keilson [74] and [75], and Kijima [80] for an overview on the subject. In Guillemin and Pinchon [50], the authors revisited the resolution of the forward Chapman-Kolmogorov equations associated with a birth-death process through the spectral theory. Their work was based on the connection between probability theory and continued fractions addressed first by Karlin and McGregor [72]. They investigated, specifically, how the Laplace Transforms of different transient characteristics related to excursions in a general birth-death process can be expressed by means of the basic orthogonal polynomials system and the spectral measure. Flajolet and Guillemin [39] have developed a formal calculus of basic events described by lattice paths associated with birth-death processes. They expressed several basic events in terms of continued fractions and their associated orthogonal polynomials. An extension of the latter paper was developed in Ball and Stefanov [21], where the authors have used an approach based on viewing birth-death processes as exponential families.

In this chapter, we consider the transient behavior of general birth-death processes. We mean by “general" that transition rates are arbitrary and need not have some special structure. Using the associated Chapman-Kolmogorov equations, and via Laplace transforms we derive closed-form expressions for the moments of downcrossing and upcrossing times between pairs of states. Further interesting characteristics are sojourn times in states. This topic is out of the scope of this chapter. By addressing some applications, we show the equivalence between the analysis of various characteristics in queueing systems and that of hitting and return time random variables. Also, we recover in a simple way classical results such as the busy period and busy cycle durations in some basic Markovian systems.

The rest of the chapter is organized as follows. In Section 5.2, we specify the general birth-death process we consider. In Sections 5.2.1 and 5.2.2, we define two families of random variables; the ordinary first passage times and the conditional first passage times, respectively. In Section 5.3, we derive the moments for the defined ordinary random variables. In Section 5.4, we apply the same analysis for the conditional random variables. In Section 5.5, we investigate various applications of the analysis of the ordinary first passage times. In Section 5.6, we close the chapter by some concluding remarks and possible future research directions.

5.2 Model Description and Notations

We consider a continuous-time birth-death process $\{E(t), t \geq 0\}$ with discrete state space taking non-negative integer values $\{0, 1, 2, 3, ...\}$ defined on a probability space. The transition rates of

the process $\{E(t),\ t \geq 0\}$ are denoted by

$$q_{m,m+1} = \lambda_m > 0, q_{m,m-1} = \mu_m, q_{m,m} = -(\lambda_m + \mu_m) \text{ for } m \geq 0, \text{ and } q_{m,n} = 0 \text{ otherwise.} \quad (5.1)$$

The rate μ_0 is equal to 0, and $\mu_m > 0$ for $m > 0$. For $m \geq 0$, we define the quantities π_m by

$$\pi_0 = 1, \text{ and } \pi_m = \frac{\lambda_0 ... \lambda_{m-1}}{\mu_1 ... \mu_m} \text{ for } m \geq 1. \quad (5.2)$$

The quantities π_m are called the potential coefficients of the birth-death process $\{E(t), t \geq 0\}$. Starting from a given initial state, let the transient probabilities be $\{p_m(t),\ t \geq 0\}$, $m \geq 0$. The quantity $p_m(t)$ represents the probability that at an arbitrary time t, the system is in state m, $m \geq 0$. Under the ergodicity assumption, the stationary distribution of the process $\{E(t), t \geq 0\}$ defined for $m \geq 0$ by the quantities $p_m = \lim_{t\to\infty} p_m(t)$ can be easily solved through recursion. They are given by

$$p_0 = \frac{1}{\sum_{m=0}^{\infty} \pi_m} > 0, \text{ and } p_m = \frac{\pi_m}{\sum_{m=0}^{\infty} \pi_m} > 0, \text{ for } m \geq 1. \quad (5.3)$$

5.2.1 First Passage Times

In this section, we define the random variables associated with first passage times in birth-death processes. Let us consider the random variable θ_m representing the duration of an excursion by the process $\{E(t),\ t \geq 0\}$ above the level $m-1$, $m \geq 1$. In other words, θ_m represents the first passage time from state m to state $m-1$. We define θ_m by

$$\theta_m = Inf\{t > 0 :\ E(t) = m - 1 \ \mid \ E(0) = m\}. \quad (5.4)$$

Also, let τ_m be the first passage time from state $m-1$ to state m, defined by

$$\tau_m = Inf\{t > 0 :\ E(t) = m \ \mid \ E(0) = m - 1\}. \quad (5.5)$$

In general concerning the first passage time from a given state to another, we define the random variables $D_{m,m-i}$ and $U_{m,m+i}$ representing the downcrossing time from state m to state $m-i$, $1 \leq i \leq m$, and the upcrossing time from state m to state $m+i$, $i \geq 1$ and $m \geq 0$, respectively. These random variables are given by

$$D_{m,m-i} = Inf\{t > 0 :\ E(t) = m - i \ \mid \ E(0) = m\}, \quad (5.6)$$

$$\text{and, } U_{m,m+i} = Inf\{t > 0 :\ E(t) = m + i \mid E(0) = m\}. \tag{5.7}$$

0ne can easily see that

$$D_{m,m-i} \doteq \sum_{n=m-i+1}^{m} \theta_n, \text{ for } 1 \leq i \leq m, \tag{5.8}$$

$$\text{and, } U_{m,m+i} \doteq \sum_{n=m+1}^{m+i} \tau_n, \text{ for } i \geq 1, \tag{5.9}$$

where $\doteq$ denotes equality in terms of distributions. In particular, let us introduce the random variables D_m and U_m representing the downcrossing time from state m to the empty state 0 and the upcrossing time from state 0 to state m, respectively. Hence, $D_m \doteq D_{m,0} \doteq \sum_{n=1}^{m} \theta_n$ and $U_m \doteq U_{0,m} \doteq \sum_{n=1}^{m} \tau_n$. Finally, let C_m be the random variable denoting the time between two visits by the process $E(t)$ at state 0, giving that the process $E(t)$ hits the state m. Then, we state that $C_m \doteq U_m + D_m \doteq \sum_{n=1}^{m} \theta_n + \tau_n$.

The analysis of the random variables defined above is useful for various problems in Markovian queueing systems. For instance, it would be helpful for the busy period analysis of queueing systems with state-dependent arrival and service rates. Also, for computing the characteristics of the state-dependent waiting time distribution in complex service systems. We shall give further details of these applications in Section 5.5.

5.2.2 Conditional First Passage Times

The random variables of first passage times, we defined above, have great importance in Markovian queueing systems applications. An equally great interest is in the conditional first passage times. In what follows, we define their associated random variables.

Let ${}^{r}\theta_m$ be the first passage time of the process $\{E(t),\ t \geq 0\}$ from state m to state $m-1$ given that the process does not visit state r in between, $1 \leq m < r$, defined by

$${}^{r}\theta_m = Inf\{t > 0 :\ E(t) = m - 1 \mid E(0) = m \text{ and no visit to } r\}. \tag{5.10}$$

Similarly, let ${}^{r}\tau_m$ be the first passage time from state $m-1$ to state m given no visit to r, $0 \leq r < m - 1$, defined by

$${}^{r}\tau_m = Inf\{t > 0 :\ E(t) = m \mid E(0) = m - 1 \text{ and no visit to } r\}. \tag{5.11}$$

As above, we also consider the conditional first passage times from a given state to another. We define the random variables ${}^{r}D_{m,m-i}$ and ${}^{r}U_{m,m+i}$ to represent the duration of the downcrossing time from state m to state $m-i$ given no visit to state r, $1 \leq m < r$, and the upcrossing time

from state m to state $m+i$ given no visit to r, $0 \leq r < m-1$, respectively. One has

$$ {}^{r}D_{m,m-i} \doteq \sum_{n=m-i+1}^{m} {}^{r}\theta_n, \text{ for } 1 \leq i \leq m < r, \tag{5.12} $$

$$ \text{and, } {}^{r}U_{m,m+i} \doteq \sum_{n=m+1}^{m+i} {}^{r}\tau_n, \text{ for } 0 \leq r < m,\ i \geq 1. \tag{5.13} $$

An interesting application of the conditional first passage times would be as follows. Consider a birth-death process that models a Markovian queueing system with a limited system capacity, say K. In practice, It is useful to determine the time from a system with one customer until the system saturation given no idleness, namely the random variable ${}^{0}U_{1,K+1}$. A dual interesting random variable is the duration from a full system until idleness given no lost customers, namely the random variable ${}^{K+1}D_{K,0}$. We state from Sumita [121] that both of these random variables are identically distributed.

5.3 Moments of First Passage Times

In this section, we focus on calculating the kth order moment, $k \geq 1$, for the ordinary first passage times defined in Section 5.2.1. Before moving on to the moments computation, we should first discuss their conditions of existence. For the upcrossing times, τ_m, $U_{m,m+i}$, and U_m, it is clear that no specific conditions are required. However, it is not the case for the downcrossing times, θ_m, $D_{m,m-i}$, D_m, as well as for C_m. To guaranty the moments existence of these return times, we shall use the following set of conditions.

Condition C^k ($k \geq 1$): the birth-death process $\{E(t),\ t \geq 0\}$ has ergodic degree k.

Roughly speaking, the ergodic degree gives the number of finite moments possessed by the time of the first passage at a given state i starting from any state $j \neq i$. We refer the reader to Mao [99] for more details. In particular, Condition (C^1) simply reflects the classical ergodicity assumption. It is the necessary and sufficient condition for the mean of the first passage time from any state i to any state $j \neq i$ ($D_{i,j}$ and $U_{i,j}$) to be finite. From Karlin and McGregor [71], Condition (C^1) holds if and only if

$$ \sum_{m=0}^{\infty} \pi_m < \infty \text{ and } \sum_{m=0}^{\infty} \frac{1}{\lambda_m \pi_m} = \infty, \tag{5.14} $$

see also Keilson [75]. As for the special case of Condition (C^2), it means that the birth-death process has ergodic degree 2. From Theorem (6.1) in Coolen-Schrijner and van Doorn [35], this

condition holds if and only if

$$\left(\sum_{j=0}^{\infty}\pi_j\right)^{-1}\left(\sum_{m=0}^{\infty}\frac{1}{\lambda_m\pi_m}\left(\sum_{j=m+1}^{\infty}\pi_j\right)^2\right)<\infty. \tag{5.15}$$

An equivalent expression was also found in Theorem (4) of Karlin and McGregor [71]. To the best of our knowledge, no explicit expressions, for $k \geq 3$, exist in the literature. We give further details on Conditions (C^k) with higher orders at the end of this section. In addition, we derive a new result by giving Condition (C^3) in an explicit form.

In Theorem (5.1), we give the kth order moment expression of the random variable θ_m. Thereafter, we deduce the mean and the variance of θ_m in Corollaries (5.1) and (5.2), respectively. In Theorem (5.2), Corollaries (5.3) and (5.4), a similar analysis is given for the random variable τ_m. Note that the latter results are new except for the expressions of the expected values of θ_m and τ_m.

For the rest of the chapter, an empty sum is being interpreted as zero, and an empty product is being interpreted as one.

Theorem 5.1 *Under Condition (C^k), the kth order moment $E(\theta_m^k)$, $k \geq 1$, of the random variable θ_m, $m \geq 1$, is given by*

$$E(\theta_m^k)=\frac{1}{\lambda_{m-1}\pi_{m-1}}\sum_{n=m}^{\infty}\lambda_{n-1}\pi_{n-1}\,V_{n,k}, \tag{5.16}$$

where

$$V_{n,k}=\frac{k}{\mu_n}\,E(\theta_n^{k-1})+\frac{\lambda_n}{\mu_n}\sum_{j=1}^{k-1}C_k^j\,E(\theta_n^j)\,E(\theta_{n+1}^{k-j}),\ \textit{for } n\geq 1,\ k\geq 1,$$

and

$$C_k^j=\frac{k!}{j!\,(k-j)!},\ \textit{for } k\geq j\geq 1.$$

Proof. From the Strong Markov Property, we can write, for $m \geq 1$

$$\begin{cases}\theta_m \doteq \varepsilon_{\lambda_m+\mu_m} & \text{, with probability } \frac{\mu_m}{\lambda_m+\mu_m}\\ \theta_m \doteq \varepsilon_{\lambda_m+\mu_m}+\theta_{m+1}+\hat{\theta}_m & \text{, with probability } \frac{\lambda_m}{\lambda_m+\mu_m}.\end{cases} \tag{5.17}$$

where $\varepsilon_{\lambda_m+\mu_m}$ is a random variable exponentially distributed with parameter $\lambda_m+\mu_m$. The random variables θ_m, θ_{m+1} and $\hat{\theta}_m$ are independent. In addition, the random variables θ_m and $\hat{\theta}_m$ are identically distributed.

Let $\tilde{\theta}_m(s)$ be the Laplace transform of the random variable θ_m. Then, Equation (5.17) yields

$$(\lambda_m + \mu_m + s)\,\tilde{\theta}_m(s) = \mu_m + \lambda_m\,\tilde{\theta}_{m+1}(s)\,\tilde{\theta}_m(s), \text{ for } m \geq 1. \tag{5.18}$$

Let $\tilde{\theta}_m^{(k)}(s)$ be the kth derivative in s of $\tilde{\theta}_m(s)$. Taking the kth derivative in s of both sides in Equation (5.18) using the Leibnitz's differentiation formula, we get for $m \geq 1$, $k \geq 1$

$$(\lambda_m + \mu_m + s)\,\tilde{\theta}_m^{(k)}(s) + k\,\tilde{\theta}_m^{(k-1)}(s) = \lambda_m \sum_{j=0}^{k} C_k^j\,\tilde{\theta}_m^{(j)}(s)\,\tilde{\theta}_{m+1}^{(k-j)}(s), \tag{5.19}$$

or equivalently

$$\begin{aligned}(\lambda_m + \mu_m + s)\,\tilde{\theta}_m^{(k)}(s) + k\,\tilde{\theta}_m^{(k-1)}(s) = \lambda_m\,\tilde{\theta}_m(s)\,\tilde{\theta}_{m+1}^{(k)}(s) + \lambda_m \sum_{j=1}^{k-1} C_k^j\,\tilde{\theta}_m^{(j)}(s)\,\tilde{\theta}_{m+1}^{(k-j)}(s) \\ + \lambda_m\,\tilde{\theta}_m^{(k)}(s)\,\tilde{\theta}_{m+1}(s).\end{aligned} \tag{5.20}$$

For $m \geq 1$ and $j = 0$, $\tilde{\theta}_m^{(j)}(0) = 1$. For $m \geq 1$ and $j \geq 1$, $\tilde{\theta}_m^{(j)}(0) = (-1)^j E(\theta_m^j)$. Hence, Equation (5.20) becomes for $s = 0$

$$\begin{aligned}(\lambda_m + \mu_m)\,(-1)^k\,E(\theta_m^k) + k\,(-1)^{k-1}E(\theta_m^{k-1}) = \lambda_m\,(-1)^k\,E(\theta_{m+1}^k) + \lambda_m\,(-1)^k\,E(\theta_m^k) \\ + \lambda_m \sum_{j=1}^{k-1} C_k^j\,(-1)^k\,E(\theta_m^j)\,E(\theta_{m+1}^{k-j}).\end{aligned} \tag{5.21}$$

Simplifying by $(-1)^k$ and dividing by $(\lambda_m + \mu_m)$ leads to

$$E(\theta_m^k) = \frac{\lambda_m}{\mu_m}\,E(\theta_{m+1}^k) + \frac{k}{\mu_m}\,E(\theta_m^{k-1}) + \frac{\lambda_m}{\mu_m} \sum_{j=1}^{k-1} C_k^j\,E(\theta_m^j)\,E(\theta_{m+1}^{k-j}). \tag{5.22}$$

Let us introduce the sequence $V_{m,k}$, for $m \geq 1$ and $k \geq 1$, defined by

$$V_{m,k} = \frac{k}{\mu_m}\,E(\theta_m^{k-1}) + \frac{\lambda_m}{\mu_m} \sum_{j=1}^{k-1} C_k^j\,E(\theta_m^j)\,E(\theta_{m+1}^{k-j}). \tag{5.23}$$

Thus, we get the following recurrence relation

$$E(\theta_m^k) = \frac{\lambda_m}{\mu_m}\,E(\theta_{m+1}^k) + V_{m,k}, \;\; m \geq 1, \;\; k \geq 1. \tag{5.24}$$

With straightforward manipulations in Equation (5.24), we state that for $m \geq 1$, $i \geq 1$, and

$k \geq 1$

$$E(\theta_m^k) = V_{m,k} + \frac{\lambda_m}{\mu_m} V_{m+1,k} + ... + \left(\prod_{j=0}^{i-1} \frac{\lambda_{m+j}}{\mu_{m+j}} \right) V_{m+i,k} + \left(\prod_{j=0}^{i} \frac{\lambda_{m+j}}{\mu_{m+j}} \right) E(\theta_{m+i+1}^k). \quad (5.25)$$

For a given $k \geq 1$ and under the Condition (C^k), we deduce that $E(\theta_m^k)$ is bounded for $m \geq 1$. Moreover, the birth-death process has, in particular, ergodic degree 1. Then $\lim_{i\to\infty} \pi_i = 0$, also, $\lim_{i\to\infty} \prod_{j=0}^{i} \frac{\lambda_{m+j}}{\mu_{m+j}} = 0$, for $m \geq 1$. Hence, it follows that $\lim_{i\to\infty} \left(\prod_{j=0}^{i} \frac{\lambda_{m+j}}{\mu_{m+j}} \right) E(\theta_{m+i+1}^k) = 0$, for $m \geq 1$, $k \geq 1$.

Continuing forward manipulations in Equation (5.25) until i goes to ∞ implies

$$E(\theta_m^k) = \sum_{i=0}^{\infty} \left(\prod_{j=0}^{i-1} \frac{\lambda_{m+j}}{\mu_{m+j}} \right) V_{m+i,k}\ ,\ m \geq 1,\ k \geq 1. \quad (5.26)$$

Finally, observing that $\frac{1}{\lambda_n} \prod_{j=m}^{n} \frac{\lambda_j}{\mu_j} = \frac{\pi_n}{\lambda_{m-1}\pi_{m-1}}$, for $n \geq m \geq 1$, and through a change in the subscripts we get

$$E(\theta_m^k) = \frac{1}{\lambda_{m-1}\pi_{m-1}} \sum_{n=m}^{\infty} \lambda_{n-1}\pi_{n-1}\, V_{n,k},\ m \geq 1,\ k \geq 1, \quad (5.27)$$

which completes the proof. □

Corollary 5.1 *Under Condition (C^1) given in Expression (5.14), the mean value $\overline{\theta}_m$ of the random variable θ_m is given by*

$$\overline{\theta}_m = \frac{1}{\lambda_{m-1}\pi_{m-1}} \sum_{n=m}^{\infty} \pi_n,\ \textit{for } m \geq 1. \quad (5.28)$$

Proof. For $n \geq 1$ and $k = 1$, we have $V_{n,k} = \frac{1}{\mu_n}$. Then, applying Theorem (5.1) leads easily to the desired result. □

The result in Corollary (5.1) can be found in Keilson [75], Kijima [80], and in Lemma (1) of Guillemin and Pinchon [50]. Note however that in the latter, the authors have proved their result through a totally different approach based on counting processes and the Strong Law of Large Numbers.

Corollary 5.2 *Under Condition (C^2) given in expression (5.15), the variance $Var(\theta_m)$ of the random variable θ_m is given, for $m \geq 1$, by*

$$Var(\theta_m) = \frac{2}{\lambda_{m-1}\pi_{m-1}} \sum_{n=m+1}^{\infty} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\sum_{i=n}^{\infty} \pi_i \right)^2 + \frac{1}{\lambda_{m-1}^2\pi_{m-1}^2} \left(\sum_{i=m}^{\infty} \pi_i \right)^2. \quad (5.29)$$

Proof. The result here concerns the special case, $k = 2$, of Theorem (5.1). For $n \geq 1$ and $k = 2$, $V_{n,k} = 2\left(\overline{\theta}_n\right)^2$. Next, using Corollary (5.1), we get the second order moment $E(\theta_m^2)$ of the random variable θ_m as follows

$$E(\theta_m^2) = \frac{2}{\lambda_{m-1}\pi_{m-1}} \sum_{n=m}^{\infty} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\sum_{i=n}^{\infty} \pi_i\right)^2, \text{ for } m \geq 1. \tag{5.30}$$

Finally, knowing that $Var(\theta_m) = E(\theta_m^2) - (\overline{\theta}_m)^2$, the result holds immediately. □

Up to now, we computed the moments of the random variable θ_m. In what follows, we go on to find explicit expressions for the moments of the random variable τ_m. Recall that τ_m represents the upcrossing time from state $m-1$ to state m.

Theorem 5.2 *The kth order moment $E(\tau_m^k)$, $k \geq 1$, of the random variable τ_m, $m \geq 1$, is given by*

$$E(\tau_m^k) = \frac{1}{\lambda_{m-1}\pi_{m-1}} \sum_{n=1}^{m} \lambda_{n-1}\pi_{n-1}\, W_{n,k}, \tag{5.31}$$

where

$$W_{n,k} = \begin{cases} \frac{k!}{\lambda_0^k}, \text{ for } n = 1,\ k \geq 1, \\ \frac{k}{\lambda_{n-1}} E(\tau_n^{k-1}) + \frac{\mu_{n-1}}{\lambda_{n-1}} \sum_{j=1}^{k-1} C_k^j\, E(\tau_{n-1}^j)\, E(\tau_n^{k-j}), \text{ for } n \geq 2,\ k \geq 1, \end{cases}$$

and

$$C_k^j = \frac{k!}{j!\,(k-j)!}, \text{ for } k \geq j \geq 1.$$

Proof. From the Strong Markov Property, we can write, for $m \geq 2$

$$\begin{cases} \tau_m \doteq \varepsilon_{\lambda_{m-1}+\mu_{m-1}} & \text{, with probability } \frac{\lambda_{m-1}}{\lambda_{m-1}+\mu_{m-1}} \\ \tau_m \doteq \varepsilon_{\lambda_{m-1}+\mu_{m-1}} + \tau_{m-1} + \hat{\tau}_m & \text{, with probability } \frac{\mu_{m-1}}{\lambda_{m-1}+\mu_{m-1}}. \end{cases} \tag{5.32}$$

where $\varepsilon_{\lambda_{m-1}+\mu_{m-1}}$ is a random variable exponentially distributed with parameter $\lambda_{m-1} + \mu_{m-1}$. The random variables τ_m, τ_{m-1} and $\hat{\tau}_m$ are independent. In addition, the random variables τ_m and $\hat{\tau}_m$ are identically distributed.

Let $\tilde{\tau}_m(s)$ be the Laplace transform of the random variable τ_m. Then, Equation (5.32) yields

$$(\lambda_m + \mu_m + s)\, \tilde{\tau}_{m+1}(s) = \lambda_m + \mu_m\, \tilde{\tau}_m(s)\, \tilde{\tau}_{m+1}(s), \text{ for } m \geq 1. \tag{5.33}$$

Let $\tilde{\tau}_m^{(k)}(s)$ be the kth derivative in s of $\tilde{\tau}_m(s)$. Taking the kth derivative in s of both sides in

Equation (5.33) using the Leibnitz's differentiation formula, we get for $m \geq 1$, $k \geq 1$

$$(\lambda_m + \mu_m + s)\, \tilde{\tau}_{m+1}^{(k)}(s) + k\, \tilde{\tau}_{m+1}^{(k-1)}(s) = \mu_m \sum_{j=0}^{k} C_k^j\, \tilde{\tau}_m^{(j)}(s)\, \tilde{\tau}_{m+1}^{(k-j)}(s). \tag{5.34}$$

Let us introduce the sequence $W_{m,k}$, $m \geq 1$ and $k \geq 1$. For $m = 1$ and $k \geq 1$, $W_{m,k} = \frac{k!}{\lambda_0^k}$, and for $m \geq 2,\ k \geq 1$, it is defined by

$$W_{m,k} = \frac{k}{\lambda_{m-1}} E(\tau_m^{k-1}) + \frac{\mu_{m-1}}{\lambda_{m-1}} \sum_{j=1}^{k-1} C_k^j\, E(\tau_{m-1}^j)\, E(\tau_m^{k-j}). \tag{5.35}$$

For $m \geq 1$ and $j = 0$, $\tilde{\tau}_m^{(j)}(0) = 1$. For $m \geq 1$ and $j \geq 1$, $\tilde{\tau}_m^{(j)}(0) = (-1)^j E(\tau_m^j)$. Next, with some algebra, Equation (5.34) becomes for $s = 0$

$$E(\tau_{m+1}^k) = \frac{\mu_m}{\lambda_m}\, E(\tau_m^k) + W_{m+1,k},\ \ m \geq 1,\ \ k \geq 1. \tag{5.36}$$

Since the random variable τ_1 is exponentially distributed with rate λ_0, so $E(\tau_1^k) = \frac{k!}{\lambda_0^k}$. Then, $E(\tau_1^k) = W_{1,k}$, and backward manipulations in Relation (5.36) imply

$$E(\tau_m^k) = \sum_{i=1}^{m} \left(\prod_{j=i}^{m-1} \frac{\mu_j}{\lambda_j} \right) W_{i,k},\ \text{for } m \geq 1,\ \ k \geq 1. \tag{5.37}$$

Observing again that $\frac{1}{\mu_n} \prod_{j=n}^{m-1} \frac{\mu_j}{\lambda_j} = \frac{\pi_n}{\lambda_{m-1}\pi_{m-1}}$, for $m \geq n \geq 1$, we state finally that

$$E(\tau_m^k) = \frac{1}{\lambda_{m-1}\pi_{m-1}} \sum_{n=1}^{m} \lambda_{n-1}\pi_{n-1}\, W_{n,k},\ \text{for } m \geq 1,\ \ k \geq 1. \tag{5.38}$$

This completes the proof. □

Corollary 5.3 *The mean value $\overline{\tau}_m$ of the random variable τ_m is given by*

$$\overline{\tau}_m = \frac{1}{\lambda_{m-1}\pi_{m-1}} \sum_{n=0}^{m-1} \pi_n,\ \textit{for } m \geq 1. \tag{5.39}$$

Proof. For $n \geq 1$ and $k = 1$, we have $W_{n,k} = \frac{1}{\lambda_{n-1}}$. Then, applying Theorem (5.2) completes the proof. □

The result of Corollary (5.3) can be found in Keilson [75], Keilson [76], Sumita [121], and also in Lemma 1 of Guillemin and Pinchon [50].

Corollary 5.4 *The variance $Var(\tau_m)$ of the random variable τ_m is given by*

$$Var(\tau_m) = \frac{2}{\lambda_{m-1}\pi_{m-1}} \sum_{n=1}^{m-1} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\sum_{i=0}^{n-1} \pi_i \right)^2 + \frac{1}{\lambda_{m-1}^2 \pi_{m-1}^2} \left(\sum_{i=0}^{m-1} \pi_i \right)^2, \text{ for } m \geq 1. \quad (5.40)$$

Proof. For $n \geq 1$ and $k = 2$, we have $W_{n,k} = 2\,(\overline{\tau}_n)^2$. Using Corollary (5.3) for the expression of $\overline{\tau}_n$, $n \geq 1$, and then applying Theorem (5.2), we get the the second order moment $E(\tau_m^2)$ of τ_m as follows

$$E(\tau_m^2) = \frac{2}{\lambda_{m-1}\pi_{m-1}} \sum_{n=1}^{m} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\sum_{i=0}^{n-1} \pi_i \right)^2, \text{ for } m \geq 1. \quad (5.41)$$

Again, knowing that $Var(\tau_m) = E(\tau_m^2) - (\overline{\tau}_m)^2$, we complete the proof. □

In what follows, we use the results obtained above to get the moments for the remaining transient characteristics defined in Section 5.2.1. Using the independence between the random variables θ_i and θ_j for $i, j \geq 1$, and the Newton's binomial formula, we can get the closed-form expressions of the moments of $D_{m,m-i}$. With the same approach, we can compute all the moments of $U_{m,m+i}$ too. For presentation issues, we only explicitly derive the expectations and the variances of the random variables D_m and U_m. The expectation expressions can be found in Kijima [80], whereas the variance expressions we give below are to our knowledge new.

Under Condition (C^1), let $\overline{D}_m$ ($\overline{U}_m$) be the mean of the random variable D_m (U_m), and under Condition (C^2), let $Var(D_m)$ ($Var(U_m)$) be its variance. From Corollaries (5.1) and (5.3), we have respectively for $m \geq 1$,

$$\overline{D}_m = \sum_{l=1}^{m} \frac{1}{\lambda_{l-1}\pi_{l-1}} \sum_{n=l}^{\infty} \pi_n, \quad (5.42)$$

$$\overline{U}_m = \sum_{l=1}^{m} \frac{1}{\lambda_{l-1}\pi_{l-1}} \sum_{n=0}^{l-1} \pi_n. \quad (5.43)$$

Now, using the independence, for any $i, j \geq 1$, between the random variables θ_i and θ_j on the one hand, and τ_i and τ_j on the other hand, we deduce respectively that $Var(D_m) = \sum_{l=1}^{m} Var(\theta_l)$, and $Var(U_m) = \sum_{l=1}^{m} Var(\tau_l)$. So, from Corollaries (5.2) and (5.4), we state respectively that, for $m \geq 1$,

$$Var(D_m) = \sum_{l=1}^{m} \left(\frac{2}{\lambda_{l-1}\pi_{l-1}} \sum_{n=l+1}^{\infty} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\sum_{i=n}^{\infty} \pi_i \right)^2 + \frac{1}{\lambda_{l-1}^2 \pi_{l-1}^2} \left(\sum_{i=l}^{\infty} \pi_i \right)^2 \right), \quad (5.44)$$

$$Var(U_m) = \sum_{l=1}^{m} \left(\frac{2}{\lambda_{l-1}\pi_{l-1}} \sum_{n=1}^{l-1} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\sum_{i=0}^{n-1} \pi_i \right)^2 + \frac{1}{\lambda_{l-1}^2 \pi_{l-1}^2} \left(\sum_{i=0}^{l-1} \pi_i \right)^2 \right). \quad (5.45)$$

Finally with some algebra, the mean $\overline{C}_m$ under Condition (C^1) and the variance $Var(C_m)$ under Condition (C^2) of the random variable C_m, for $m \geq 1$, are given by

$$\overline{C}_m = \left(\sum_{n=0}^{\infty} \pi_n\right) \times \left(\sum_{l=1}^{m} \frac{1}{\lambda_{l-1}\pi_{l-1}}\right), \tag{5.46}$$

$$\begin{aligned} Var(C_m) = &\left(\sum_{l=1}^{m} \frac{2}{\lambda_{l-1}\pi_{l-1}}\right) \times \left(\sum_{n=1}^{\infty} \frac{1}{\lambda_{n-1}\pi_{n-1}} \left(\left(\sum_{i=0}^{n-1} \pi_i\right)^2 + \left(\sum_{i=n}^{\infty} \pi_i\right)^2\right)\right) \\ &+ \sum_{l=1}^{m} \frac{1}{\lambda_{l-1}^2\pi_{l-1}^2} \left(\left(\sum_{i=0}^{l-1} \pi_i\right)^2 + \left(\sum_{i=l}^{\infty} \pi_i\right)^2\right). \end{aligned} \tag{5.47}$$

Let us come back to investigate the condition under which a birth-death has ergodic degree k, $k \geq 1$. The quantities D_m, $m \geq 1$, play a key role to derive explicitly the Conditions (C^k) for higher order moments, $k \geq 3$, which to the best of our knowledge do not exist in the literature. Let the random variable $D_{e,j}$ be the first passage time from the ergodic distribution to state j, $j \geq 0$. In accordance with the notations in Section 5.2.1, the random variable D_e denotes the first passage time from the ergodic distribution to state 0. It is clear that the kth order moment of D_e, for $k \geq 1$, is given by

$$E(D_e^k) = \sum_{s=0}^{\infty} p_s \, E(D_s^k). \tag{5.48}$$

Recall that the quantities $\{p_s, s \geq 0\}$ are the stationary probabilities already given in Expression (5.3). Collecting thereafter some developments in Coolen-Schrijner and van Doorn [35], we state the following theorem.

Theorem 5.3 *Condition (C^k), $k \geq 1$, holds if and only if*

$$\sum_{s=0}^{\infty} p_s \, E(D_s^{k-1}) < \infty, \tag{5.49}$$

Proof. From Theorem (3.1) in Coolen-Schrijner and van Doorn [35], we state on the one hand that the kth order moment of the first passage time from any state i to any state j is finite, if and only if, the $(k-1)$th order moment of the first passage time from the ergodic distribution to any state j is finite. On the other hand, the latter condition is sufficient and necessary for the $(k-1)$th order moment of the first passage time from the ergodic distribution to some state j to be finite. Applying this statement to the particular case, $j = 0$, and using Equation (5.48) complete the proof.

As application, we give in Corollary (5.5) an explicit expression for Condition (C^3).

Corollary 5.5 *Condition (C^3) holds if and only if*

$$\sum_{s=0}^{\infty}\sum_{l=1}^{s}\frac{2\,\pi_s}{\lambda_{l-1}\pi_{l-1}\sum_{j=0}^{\infty}\pi_j}\sum_{n=l}^{\infty}\frac{1}{\lambda_{n-1}\pi_{n-1}}\left(\sum_{r=n}^{\infty}\pi_r\right)^2 + \sum_{s=0}^{\infty}\sum_{i,j=1,\,j>i}^{s}\frac{2\,\pi_s}{\lambda_{i-1}\lambda_{j-1}\pi_{i-1}\pi_{j-1}\sum_{j=0}^{\infty}\pi_j}\left(\sum_{n=i}^{\infty}\pi_n\right)\left(\sum_{n=j}^{\infty}\pi_n\right) < \infty. \tag{5.50}$$

Proof. From the independence between the random variables θ_i and θ_j, for $i, j \geq 1$ and $i \neq j$, we have

$$E(D_s^2) = \sum_{l=1}^{s} E(\theta_l^2) + 2\sum_{i=1}^{s}\sum_{j=1,\,j>i}^{s} E(\theta_i)E(\theta_j). \tag{5.51}$$

Using the above relation in Equation (5.48), and applying next Corollary (5.1) and Equation (5.30) lead to the desired result. □

We close the analysis for ordinary first passage times and turn to that of conditional first passage times in Section 5.4.

5.4 Moments of Conditional First Passage Times

In this section, we focus on calculating the kth order moment, $k \geq 1$, of the conditional first passage times defined in Section 5.2.2 , ${}^r\theta_m$, ${}^r\tau_m$, ${}^rD_{m,m-i}$ and ${}^rU_{m,m+i}$. The results we derive here has not been done before in the literature, except as we shall mention later, for a special case for ${}^r\tau_m$. Note that no existence conditions are required for the computation of their moments.

Before giving the results for the conditional random variables, we need to introduce some notations. These preliminaries are specifically related to the notion of ruin probabilities. Consider again the birth-death defined in Section 5.2. Let ${}^r\eta_m$ be the ruin probability that the particle, starting at m, reaches $m-1$ first before r, $1 \leq m < r$. It is clear that the ruin probability ${}^r\eta_{r-1}$ to reach $r-2$ starting at $r-1$, without visiting r, is given by $\frac{\mu_{r-1}}{\lambda_{r-1}+\mu_{r-1}}$. For a given m, $1 \leq m < r-1$, we define the event ${}^rA_m$ that the particle reaches first $m-1$ starting from m, without visiting r. Let us calculate now the probability that ${}^rA_m$ occurs, namely ${}^r\eta_m$. In state m, two events can occur: either the process goes down to $m-1$, say event ${}^rB_m$, or the process goes up to $m+1$ which is the complementary event of ${}^rB_m$, say ${}^rB_m^c$. Hence, we can write

$$Pr({}^rA_m) = Pr({}^rA_m \mid {}^rB_m) \times Pr({}^rB_m) + Pr({}^rA_m \mid {}^rB_m^c) \times Pr({}^rB_m^c). \tag{5.52}$$

The event ${}^rA_m \mid {}^rB_m$ is to reach $m-1$ starting from $m-1$ without visiting r, which obviously

occurs with probability 1 since the process is already in state $m-1$. The event ${}^{r}A_m \mid {}^{r}B_m^c$ is to reach $m-1$ first before r, starting at $m+1$, which is equivalent to the following: starting at $m+1$, the process reaches m without visiting r, then starting from m, it reaches $m-1$ without visiting r. So, $Pr({}^{r}A_m \mid {}^{r}B_m^c) = {}^{r}\eta_{m+1}\,{}^{r}\eta_m$. Furthermore, the event ${}^{r}B_m$ occurs with probability $\frac{\mu_m}{\lambda_m+\mu_m}$, and the event ${}^{r}B_m^c$ with probability $\frac{\lambda_m}{\lambda_m+\mu_m}$. These arguments lead to the following recursive relation

$$ {}^{r}\eta_m = \frac{\mu_m}{\lambda_m+\mu_m} + \frac{\lambda_m}{\lambda_m+\mu_m}\,{}^{r}\eta_{m+1}\,{}^{r}\eta_m, \text{ for } 1 \le m < r-1, \tag{5.53} $$

or equivalently

$$ {}^{r}\eta_m = \frac{\mu_m}{\mu_m + \lambda_m(1 - {}^{r}\eta_{m+1})}, \text{ for } 1 \le m < r-1, \tag{5.54} $$

starting with ${}^{r}\eta_{r-1} = \frac{\mu_{r-1}}{\lambda_{r-1}+\mu_{r-1}}$.

For $1 \le m < r-1$, we define the quantities δ_m by

$$ \delta_m = \mu_m + \lambda_m(1 - {}^{r}\eta_{m+1}), \tag{5.55} $$

and for $0 \le m < r-1$, we introduce the quantities χ_m as follows

$$ \chi_0 = 1, \text{ and } \chi_m = \frac{(\lambda_0\,{}^{r}\eta_1)\,(\lambda_1\,{}^{r}\eta_2)...(\lambda_{m-1}\,{}^{r}\eta_m)}{\delta_1\,\delta_2...\delta_m}, \text{ for } 1 \le m < r-1. \tag{5.56} $$

Theorem 5.4 *The kth order moment $E({}^{r}\theta_m^k)$, $k \ge 1$, of the random variable ${}^{r}\theta_m$, $1 \le m \le r-1$, is given by*

$$ E({}^{r}\theta_{r-1}^k) = \frac{k!}{(\lambda_{r-1}+\mu_{r-1})^k}, \tag{5.57} $$

and

$$ E({}^{r}\theta_m^k) = \frac{1}{\lambda_{m-1}\,{}^{r}\eta_m\,\chi_{m-1}} \sum_{n=m}^{r-1} \lambda_{n-1}\,{}^{r}\eta_n\,\chi_{n-1}\,{}^{r}V_{n,k}, \text{ for } 1 \le m < r-1, \tag{5.58} $$

where

$$ {}^{r}V_{n,k} = \frac{k}{\delta_n}\,E({}^{r}\theta_n^{k-1}) + \frac{\lambda_n\,{}^{r}\eta_{n+1}}{\delta_n} \sum_{j=1}^{k-1} C_k^j\,E({}^{r}\theta_n^j)\,E({}^{r}\theta_{n+1}^{k-j}), \text{ for } 1 \le m < r-1, $$

and

$$ C_k^j = \frac{k!}{j!\,(k-j)!}, \text{ for } k \ge j \ge 1. $$

Proof. One can easily see that the random variable ${}^{r}\theta_{r-1}$ is exponentially distributed with rate $\lambda_{r-1}+\mu_{r-1}$. Then, its kth order moment is given by $E({}^{r}\theta_{r-1}^k) = \frac{k!}{(\lambda_{r-1}+\mu_{r-1})^k}$, $k \ge 1$. For

$1 \leq m < r-1$, we can write using the Strong Markov Property

$$\begin{cases} {}^r\theta_m \doteq \varepsilon_{\lambda_m+\mu_m} & \text{, with probability } 1 - {}^r\omega_m \\ {}^r\theta_m \doteq \varepsilon_{\lambda_m+\mu_m} + {}^r\theta_{m+1} + {}^r\hat{\theta}_m & \text{, with probability } {}^r\omega_m. \end{cases} \tag{5.59}$$

where $\varepsilon_{\lambda_m+\mu_m}$ is a random variable exponentially distributed with parameter $\lambda_m + \mu_m$. The random variables ${}^r\theta_m$, ${}^r\theta_{m+1}$ and ${}^r\hat{\theta}_m$ are independent. In addition, the random variables ${}^r\theta_m$ and ${}^r\hat{\theta}_m$ are identically distributed. The quantity ${}^r\omega_m$ is the probability that the process goes up to state $m+1$ and subsequently comes back to m without visiting r, ${}^r\omega_m = \frac{\lambda_m}{\lambda_m+\mu_m}\,{}^r\eta_{m+1}$. Let ${}^r\tilde{\theta}_m(s)$ be the Laplace transform of the random variable ${}^r\theta_m$. Then, Equation (5.59) yields

$$(\lambda_m + \mu_m + s)\,{}^r\tilde{\theta}_m(s) = \delta_m + \lambda_m\,{}^r\eta_{m+1}\,{}^r\tilde{\theta}_{m+1}(s)\,{}^r\tilde{\theta}_m(s), \text{ for } 1 \leq m < r-1. \tag{5.60}$$

As in the proof of Theorem (5.1), using the Leibnitz's differentiation formula, we get the following recursive relation, for $1 \leq m < r-1$, $k \geq 1$

$$E({}^r\theta_m^k) = \frac{\lambda_m\,{}^r\eta_{m+1}}{\delta_m}\,E({}^r\theta_{m+1}^k) +^r V_{m,k}. \tag{5.61}$$

Finally, with straightforward manipulations we complete the proof. □

Corollary 5.6 *The mean value ${}^r\overline{\theta}_m$ of the random variable ${}^r\theta_m$ is given by*

$${}^r\overline{\theta}_{r-1} = \frac{1}{\lambda_{r-1}+\mu_{r-1}}, \text{ and, } {}^r\overline{\theta}_m = \frac{1}{\lambda_{m-1}\,{}^r\eta_m\,\chi_{m-1}}\sum_{n=m}^{r-1}\chi_n, \text{ for } 1 \leq m < r-1. \tag{5.62}$$

Proof. The first part of the corollary is immediately obtained from the special case, $k=1$, of Theorem (5.4). As for the second part, one has ${}^rV_{n,1} = \frac{1}{\delta_n}$ for $1 \leq n < r-1$, next observing that $\lambda_{n-1}{}^r\eta_n\chi_{n-1} = \delta_n\chi_n$ and again applying Theorem (5.4), for $k=1$, complete the proof. □

Corollary 5.7 *The variance $Var({}^r\theta_m)$ of the random variable ${}^r\theta_m$ is given by*

$$Var({}^r\theta_{r-1}) = \frac{1}{(\lambda_{r-1}+\mu_{r-1})^2}, \text{ and, for } 1 \leq m < r-1, \tag{5.63}$$

$$Var({}^r\theta_m) = \frac{2}{\lambda_{m-1}\,{}^r\eta_m\,\chi_{m-1}}\sum_{n=m+1}^{r-1}\frac{1}{\lambda_{n-1}\,{}^r\eta_n\,\chi_{n-1}}\left(\sum_{i=n}^{r-1}\chi_i\right)^2 + \frac{1}{\lambda_{m-1}^2\,{}^r\eta_m^2\,\chi_{m-1}^2}\left(\sum_{i=m}^{r-1}\chi_i\right)^2. \tag{5.64}$$

Proof. The first part of the corollary is a direct consequence (special case, $k=2$) of Theorem (5.4). For the second part, it suffices to see that ${}^rV_{n,2} = 2\left({}^r\overline{\theta}_n\right)^2$, $1 \leq m < r-1$. Next, by

simply applying Corollary (5.6) and again Theorem (5.4), for $k = 2$, we get the second order moment $E({}^r\theta_m^2)$ of ${}^r\theta_m$ as follows.

$$E({}^r\theta_m^2) = \frac{2}{\lambda_{m-1}\,{}^r\eta_m\,\chi_{m-1}} \sum_{n=m}^{r-1} \frac{1}{\lambda_{n-1}\,{}^r\eta_n\,\chi_{n-1}} \left(\sum_{i=n}^{r-1} \chi_i\right)^2, \text{ for } 1 \leq m < r-1. \tag{5.65}$$

Finally, the result holds using the definition, $Var({}^r\theta_m) = E({}^r\theta_m^2) - ({}^r\overline{\theta}_m)^2$. □

In what follows, we focus on the moments computation of the random variable ${}^r\tau_m$. As above, we first introduce some notations. Let ${}^r\nu_m$ be the ruin probability that the process, starting at $m-1$, reaches m first before r, $m \geq r+2$. It is clear that ${}^r\nu_{r+2} = \frac{\lambda_{r+1}}{\lambda_{r+1}+\mu_{r+1}}$. With a similar explanation as for the ruin probability ${}^r\eta_m$, we give the following recursive relation, for $m > r+2$,

$${}^r\nu_m = \frac{\lambda_{m-1}}{\lambda_{m-1} + \mu_{m-1}(1 - {}^r\nu_{m-1})}. \tag{5.66}$$

For $m \geq r+1$, we define the quantities β_m by

$$\beta_{r+1} = \lambda_{r+1} + \mu_{r+1}, \text{ and, } \beta_m = \lambda_m + \mu_m(1 - {}^r\nu_m), \text{ for } m > r+1, \tag{5.67}$$

and for $m \geq r+1$, we introduce the quantities ϕ_m as

$$\phi_{r+1} = 1, \text{ and } \phi_m = \frac{\beta_{r+1}\,\beta_{r+2}...\beta_{m-1}}{(\mu_{r+2}\,{}^r\nu_{r+2})\,(\mu_{r+3}\,{}^r\nu_{r+3})...(\mu_m\,{}^r\nu_m)}, \text{ for } m > r+1. \tag{5.68}$$

Theorem 5.5 *The kth order moment $E({}^r\tau_m^k)$, $k \geq 1$, of the random variable ${}^r\tau_m$, $m \geq r+2$, is given by*

$$E({}^r\tau_{r+2}^k) = \frac{k!}{(\lambda_{r+1} + \mu_{r+1})^k} \tag{5.69}$$

and

$$E({}^r\tau_m^k) = \frac{1}{\beta_{m-1}\phi_{m-1}} \sum_{n=r+2}^{m} \beta_{n-1}\phi_{n-1}\,{}^rW_{n,k}, \text{ for } m > r+2 \tag{5.70}$$

where

$$W_{n,k} = \frac{k}{\beta_{n-1}} E({}^r\tau_n^{k-1}) + \frac{\mu_{n-1}\,{}^r\nu_{n-1}}{\beta_{n-1}} \sum_{j=1}^{k-1} C_k^j\, E({}^r\tau_{n-1}^j)\, E({}^r\tau_n^{k-j}), \text{ for } n > r+2,\ k \geq 1,$$

and

$$C_k^j = \frac{k!}{j!\,(k-j)!}, \text{ for } k \geq j \geq 1.$$

Proof. It is easy to see that the kth order moment of the random variable ${}^r\tau_{r+2}$ is given by

$E({}^r\tau_{r+1}^k) = \frac{k!}{(\lambda_{r+1}+\mu_{r+1})^k}$, $k \geq 1$. For $m > r+2$, we can write from the Strong Markov Property

$$\begin{cases} {}^r\tau_m \doteq \varepsilon_{\lambda_{m-1}+\mu_{m-1}} & \text{, with probability } 1 - {}^r\alpha_{m-1} \\ {}^r\tau_m \doteq \varepsilon_{\lambda_{m-1}+\mu_{m-1}} + {}^r\tau_{m-1} + {}^r\hat{\tau}_m & \text{, with probability } {}^r\alpha_{m-1}. \end{cases} \tag{5.71}$$

where $\varepsilon_{\lambda_{m-1}+\mu_{m-1}}$ is a random variable exponentially distributed with parameter $\lambda_{m-1}+\mu_{m-1}$. The random variables ${}^r\tau_m$, ${}^r\tau_{m-1}$ and ${}^r\hat{\tau}_m$ are independent. In addition, the random variables ${}^r\tau_m$ and ${}^r\hat{\tau}_m$ are identically distributed. The quantity ${}^r\alpha_m$ is the probability that the process goes down from state m to state $m-1$ and subsequently comes back to m without visiting r, ${}^r\alpha_m = \frac{\mu_m}{\lambda_m+\mu_m}\,{}^r\nu_m$.

As above, we use the Laplace transform and the Leibnitz's differentiation formula to get

$$E({}^r\tau_{m+1}^k) = \frac{\mu_m\,{}^r\nu_m}{\beta_m}\,E({}^r\tau_m^k) + {}^rW_{m+1,k},\ \ m > r+2,\ \ k \geq 1. \tag{5.72}$$

Using the latter recursive relation, the result of the theorem follows. □

Note that the general recursive relation, given in Equation (5.72), can be found in Sumita [121] in the special cases, $k=1$ and $k=2$.

Corollary 5.8 *The mean value ${}^r\overline{\tau}_m$ of the random variable ${}^r\tau_m$ is given by*

$${}^r\overline{\tau}_m = \frac{1}{\beta_{m-1}\phi_{m-1}} \sum_{n=r+1}^{m-1} \phi_n,\ \textit{for } m \geq r+2. \tag{5.73}$$

Proof. Observing that ${}^rW_{n,1} = \frac{1}{\beta_{n-1}}$, for $n \geq r+2$, the result holds from the special case, $k=1$, of Theorem (5.5). □

Corollary 5.9 *The variance $Var({}^r\tau_m)$ of the random variable ${}^r\tau_m$ is given, for $m \geq r+2$, by*

$$Var({}^r\tau_m) = \frac{2}{\beta_{m-1}\phi_{m-1}} \sum_{n=r+2}^{m-1} \frac{1}{\beta_{n-1}\phi_{n-1}} \left(\sum_{i=r+1}^{n-1} \phi_i\right)^2 + \frac{1}{\beta_{m-1}^2\phi_{m-1}^2} \left(\sum_{i=r+1}^{m-1} \phi_i\right)^2. \tag{5.74}$$

Proof. For $n \geq r+2$ and $k=2$, we have ${}^rW_{n,k} = 2\,({}^r\overline{\tau}_n)^2$. Next, using Corollary (5.8) and applying Theorem (5.5) for the special case, $k=2$, give us the second order moment $E({}^r\tau_m^2)$ of ${}^r\tau_m$ as follows.

$$E({}^r\tau_m^2) = \frac{2}{\beta_{m-1}\phi_{m-1}} \sum_{n=r+2}^{m} \frac{1}{\beta_{n-1}\phi_{n-1}} \left(\sum_{i=r+1}^{n-1} \phi_i\right)^2,\ \text{for } m \geq r+2, \tag{5.75}$$

which leads to the desired result. □

From the above analysis of this section, we can obtain characteristics of several random variable of conditional first passage times. For instance, let ${}^{r}\overline{D}_{m,m-i}$ and $Var({}^{r}D_{m,m-i})$ be the mean and variance of the random variable ${}^{r}D_{m,m-i}$, respectively. From Corollaries (5.6) and (5.7), we have respectively, for $1 \leq i \leq m < r$,

$$ {}^{r}\overline{D}_{m,m-i} = \sum_{n=m-i+1}^{m} \frac{1}{\lambda_{n-1}\,{}^{r}\eta_{n}\chi_{n-1}} \sum_{j=n}^{r-1} \chi_{j}, \tag{5.76} $$

$$ \begin{aligned} Var({}^{r}D_{m,m-i}) = & \sum_{n=m-i+1}^{m} \frac{2}{\lambda_{n-1}\,{}^{r}\eta_{n}\chi_{n-1}} \sum_{j=n+1}^{r-1} \frac{1}{\lambda_{j-1}\,{}^{r}\eta_{j}\chi_{j-1}} \left(\sum_{l=j}^{r-1} \chi_{l} \right)^{2} \\ & + \sum_{n=m-i+1}^{m} \frac{1}{\lambda_{n-1}^{2}\,{}^{r}\eta_{n}^{2}\chi_{n-1}^{2}} \left(\sum_{l=n}^{r-1} \chi_{l} \right)^{2}. \end{aligned} \tag{5.77} $$

Similarly, we denote by ${}^{r}\overline{U}_{m,m+i}$ and $Var({}^{r}U_{m,m+i})$ the mean and variance of ${}^{r}U_{m,m+i}$, respectively. From Corollaries (5.8) and (5.9), we have respectively, for $0 \leq r < m$, $i \geq 1$,

$$ {}^{r}\overline{U}_{m,m+i} = \sum_{n=m+1}^{m+i} \frac{1}{\beta_{n-1}\phi_{n-1}} \sum_{j=r+1}^{n-1} \phi_{j}, \tag{5.78} $$

$$ Var({}^{r}U_{m,m+i}) = \sum_{n=m+1}^{m+i} \left(\frac{2}{\beta_{n-1}\phi_{n-1}} \sum_{j=r+2}^{n-1} \frac{1}{\beta_{j-1}\phi_{j-1}} \left(\sum_{l=r+1}^{j-1} \phi_{l} \right)^{2} + \frac{1}{\beta_{n-1}^{2}\phi_{n-1}^{2}} \left(\sum_{l=r+1}^{n-1} \phi_{l} \right)^{2} \right). \tag{5.79} $$

5.5 Applications

In this section, we give indications about some applications of the obtained theoretical results of this chapter. First, we revisit the important concepts of busy period and busy cycle in queueing systems. Second, we address another important application, which is the prediction of state-dependent queueing delays in non-standard queueing systems; systems with impatient customers (linear growth death rates), or state-dependent arrival rates, or in general, systems with state-dependent transition rates.

5.5.1 Busy Period Analysis for the $M/M/1$ and $M/M/s$ Queues

In this section, we apply some special cases of the results of Section 5.3 to retrieve known results for the simple $M/M/1$ and $M/M/s$ queues.

First, let us consider an $M/M/1$ queue. Customers arrive according to a Poisson process with rate λ. The time it takes to serve every customer is exponential with rate μ. Service times are assumed to be mutually independent and further independent from the interarrival times. When a customer enters an empty system his service starts at once. If the unique server is busy, a new customer joins the queue which has infinite capacity. When a service completion occurs, a customer from the queue (we do not need here to specify which one of the customers), if any, enters the service facility at once to start service. Let ρ be the server utilization, $\rho = \lambda/\mu$. Under stability conditions, $\rho < 1$. This is equivalent to Condition (C^1) defined in Section 5.3.

Let $E(t)$ be the number of customers in system at a random instant t. The process $\{E(t), t \geq 0\}$ is a particular case of the birth-death process we present in Section 5.2. The transition rates are constants. The birth rate is $\lambda_m = \lambda$ for $m \geq 0$, and the death rate is $\mu_m = \mu$ for $m \geq 1$. The busy period for the $M/M/1$ system is defined to begin with the arrival of a customer to an idle server and to end when the server next becomes idle. Hence, the busy period length of an $M/M/1$ queue is represented by the random variable measuring the first passage time from state 1 (one customer in system) to state 0 (no customers in system), namely the random variable θ_1 defined in Section 5.2.1. Let us now check the results obtained in Section 5.3 in the particular case we present here.

On the one hand, we use the expressions found in Section 5.3 to compute the first five order moments of the random variable θ_1. For our model, transition rates above one state do not depend on the state itself. Then, one should simplify the algebra using the fact that the random variables θ_i and θ_j are identically distributed, for $i, j \geq 1$. Next, by simply observing that $\sum_{i=1}^{\infty}(\frac{\lambda}{\mu})^i = \frac{\lambda}{\mu-\lambda}$ (for $\frac{\lambda}{\mu} < 1$), we deduce from Theorem (5.1) that the first five order moments are $\frac{1}{\mu-\lambda}$, $\frac{2\mu}{(\mu-\lambda)^3}$, $\frac{6\mu(\lambda+\mu)}{(\mu-\lambda)^5}$, $\frac{24\mu(\lambda\mu+(\lambda+\mu)^2)}{(\mu-\lambda)^7}$, and $\frac{120\mu(\lambda+\mu)(3\lambda\mu+(\lambda+\mu)^2)}{(\mu-\lambda)^9}$, respectively.

On the other hand, it is known from classical results, as in Gross and Harris [46], that the Laplace transform in t, $\tilde{\theta_1}(s)$, of the probability density function (pdf) of the busy period duration of the $M/M/1$ queue is given by $\tilde{\theta_1}(s) = \frac{2\mu}{\lambda+\mu+s+\sqrt{(\lambda+\mu+s)^2-4\lambda\mu}}$, for $s \geq 0$. Then, using the relation $E(\theta_1^k) = (-1)^k\tilde{\theta}_1^{(k)}(0)$, for $k \geq 1$, one can again find the expressions derived from our results.

The busy period results are of value when addressing the busy cycle analysis for the $M/M/1$ queue. A busy cycle is defined as the sum of a busy period and an adjacent idle period, or equivalently, the time between two successive departures leaving an empty system, or two successive

arrivals to an empty system. Since the arrivals here are assumed to follow a Poisson process, the probability density function (pdf) of the idle period is exponential with parameter λ; hence the pdf of the busy cycle for the $M/M/1$ queue is the convolution of this negative exponential with the pdf of the busy period itself. Following the notations in Section 5.2.1, the busy cycle duration is clearly $\tau_1 + \theta_1$, namely C_1. In particular, we deduce from Equation (5.46) that the busy cycle expectation is $\overline{C}_1 = \frac{\mu}{\lambda(\mu-\lambda)}$, which agrees with a classical result in queueing literature, see for example Gross and Harris [46].

Let us now address the previous analysis for an $M/M/s$ queue. We consider an $M/M/s$ queue with s identical and independent servers. We consider the same assumptions for the arrival and service processes as those for the $M/M/1$ queue. Again, we do not need to specify the service discipline, except to be non-idling. Finally, let ρ be the server utilization, $\rho = \lambda/s\mu$. Under stability condition, (C^1), we have $\rho < 1$.

The process $\{E(t),\ t \geq 0\}$ counting the number of customers in system is a birth-death process, and it is a particular case of the one we present in Section 5.2. The birth rate is $\lambda_m = \lambda$ for $m \geq 0$, and the death rate is $\mu_m = m\mu$ for $1 \leq m \leq s$, and $\mu_m = s\mu$ for $m > s$. The busy period of the $M/M/s$ system is defined as the time from an arrival of a customer to a system with only one idle server until the first time one of the servers becomes idle. Thus, it represents the duration of an excursion by the process $\{E(t),\ t \geq 0\}$ above the level $s-1$, namely θ_s. With a little thought it should be clear that the busy period pdf of the $M/M/s$ queue can be obtained by taking its expression in the case of an $M/M/1$ queue and substituting μ (capacity of service in the $M/M/1$ queue) with $s\mu$ (capacity of service in the $M/M/s$ queue.) One can easily validate this intuitive result by considering the state-transition-rate diagrams for both processes. In fact, transition rates, above any state $m \geq s-1$, of the birth-death process associated with the $M/M/s$ queue are constant. In addition, they reduce to the ones for the $M/M/1$ case if we substitute $s\mu$ by μ. In this configuration, both of the processes behaves equivalently when calculating an excursion duration from state m to state $m-1$, such that $m \geq s$, and in particular when calculating the busy period duration. Next, one may again check with some algebra that the expressions of the moments obtained from the results of Section 5.3 coincide with those already known from the literature.

5.5.2 Busy Period Analysis for the $M/M/1+M$ and $M/M/s+M$ Queues

In this section, we address the analysis of the busy period for some special cases of queueing systems with reneging. Incorporating reneging in queueing models is well known to be of interest. It has an important effect on the performance measures as we have shown in Chapters 3 and

4. For instance, reneging is of special interest in manufacturing systems dealing with perishable products, also in call centers where customers may hang up once they feel that their waiting time before getting service is too long, etc.

To the best of our knowledge, the results below are not given beforehand. First, let us consider an $M/M/1+M$ queue. The model is identical to the $M/M/1$ queue described in Section 5.5.1. However, the customers here are impatient (symbol M after the $+$.) Times before reneging are assumed to be mutually independent of each other and identically distributed with an exponential rate $\sigma > 0$. We consider a different notation for the reneging rate (σ instead of γ as Chapters 3 and 4) in order to avoid any confusion with the Incomplete Gamma Function we are using below. Again, we do not need for our analysis to specify the service discipline, except to be non-idling. Finally, recall that abandonments make the system unconditionally stable. In concrete terms, Condition (C^1) holds for any set of parameters such that $\sigma > 0$.

For our $M/M/1+M$ model, the corresponding infinitesimal transition rates in the generalized birth-death process are given by $\lambda_m = \lambda$ for $m \geq 0$, and $\mu_m = \mu + (m-1)\sigma$ for $m \geq 1$. The busy period duration is given by the random variable θ_1. To get any moment of order k, for $k \geq 1$, it suffices to use the obtained relations from Theorem (5.1). In what follows, we only give the mean $\overline{\theta}_1$ and variance $Var(\theta_1)$. From Corollary (5.1), one state that

$$\overline{\theta}_1 = \frac{1}{\lambda} \sum_{n=1}^{\infty} \frac{\lambda^n}{\prod_{j=0}^{n-1}(\mu + j\sigma)}. \tag{5.80}$$

Consider the Gamma Function $\Gamma(x)$ defined for $x \geq 0$, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. It is known that $\prod_{j=0}^{n-1} \mu + j\sigma = \frac{\sigma^n \Gamma(\mu/\sigma + n)}{\Gamma(\mu/\sigma)}$, see Ancker and Gafarian [10]. So, from the relation $\gamma(x,a) = e^{-x}x^a \sum_{n=0}^{\infty} \frac{\Gamma(a)}{\Gamma(a+n+1)}x^n$ where $\gamma(x,a) = \int_0^x t^{a-1}e^{-t}dt$ is the Incomplete Gamma Function defined for $a, x \geq 0$, we deduce with some algebra that

$$\overline{\theta}_1 = \frac{1}{\lambda} \cdot (\frac{\lambda}{\sigma})^{1-\frac{\mu}{\sigma}} \cdot e^{\frac{\lambda}{\sigma}} \cdot \gamma(\frac{\lambda}{\sigma}, \frac{\mu}{\sigma}). \tag{5.81}$$

Note that Equation (5.81) can be useful for numerical computation since the Incomplete Gamma Function is extensively tabulated. Concerning the variance $Var(\theta_1)$, we have not found a simpler expression. It is given by

$$Var(\theta_1) = (\overline{\theta}_1)^2 + \frac{2}{\lambda^2} \sum_{n=2}^{\infty} \frac{\prod_{j=0}^{n-2}(\mu + j\sigma)}{\lambda^{n-1}} \sum_{i=n}^{\infty} \frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu + j\sigma)}. \tag{5.82}$$

To get some numerical illustrations, we consider 3 cases by varying the system parameters. The parameters of the first $M/M/1+M$ system are $\lambda = 0.2$, $\mu = 0.3$ and $\sigma = 0.2$. Those for the

second system are $\lambda = 0.3$, $\mu = 0.5$ and $\sigma = 0.2$. Finally, we have for the third system $\lambda = 0.8$, $\mu = 0.5$ and $\sigma = 0.4$. The results are shown in Table 5.1.

	System 1		System 2		System 3	
k	Numer	Simu	Numer	Simu	Numer	Simu
1	5.15	5.15	3.24	3.24	5.66	5.66
2	65.03	65.04	27.63	27.63	88.05	88.04
3	1,319.64	1,319.51	392.12	392.14	2,157.46	2,157.20

Table 5.1: kth order moments of the busy period duration for the $M/M/1 + M$ queue, $k = 1..3$

As one would expect, the busy period duration for the special case without abandonments ($\sigma = 0$) gives an upper bound of that we consider here. The reason is that reneging leads to fewer customers in the last system. In Table 5.2, we give numerical examples for the first three order moments associated with the first two systems we consider above but without abandonments, $\sigma = 0$. We omit the computation for the third system because it becomes unstable when assuming no abandonments.

	$E(\theta_1^k)$	
k	System 1	System 2
1	10	3.33
2	600	37.03
3	90,000	864.19

Table 5.2: kth order moments of the busy period duration for the $M/M/1$ queue without abandonments, $k = 1..3$

We notice that the analysis above can be easily extended to the case when the reneging rate depend on the position of the customer in the queue. Also, as we have explained for the $M/M/s$ queue, the busy period moments for the $M/M/s + M$ queue can be obtained simply by taking those of the $M/M/1 + M$ queue and substituting the service capacity, μ, in the first model by that, $s\mu$, in the second model.

5.5.3 Estimating State-Dependent Waiting Times

In this section, we continue on showing the usefulness of the results of this chapter. We present an application related to the distribution of state-dependent queueing delays in a multiclass Markovian queueing system. The motivation of such application deals with the usefulness of the prediction of queueing delays, as we have discussed in Chapter 4. Many prediction methods have been done, see for example Whitt [137], Jouini and Dallery [64] and [67], Nakibly [103], Rosenlund [114], and Koole [84] where the author has developed one simple algorithm for calculating tail

probabilities of Cox distributions.

In Section 4.4 of the previous chapter, we considered a two-class Markovian queueing system with reneging. We focused on estimating virtual delays for a new arrival given that the call center provides to him information about anticipated delays. In particular for a new type B customer, we derived the analysis using the results of Section 5.3 of the present chapter. In what follows, we consider a slightly different model in the sense that we let customers renege, whereas we do not give delays information to them. We again show how the obtained results may help us to predict state-dependent delays for new arrivals.

We use the same notations as in Chapter 4, except for the common reneging rate (σ instead of γ) in order to be coherent with the notations of the present chapter. We consider a new arrival who finds all the servers busy, n_A type A waiting customers in queue A and n_B type B customers in queue B. We separate the study depending on whether the call of interest is of type A or B. Type A customers observe a regular queue without priority. Then, the conditional waiting time distribution of a new customer A is easy to derive. It follows an hypoexponential distribution, which is the convolution of $(n_A + 1)$ exponential distributions with parameters $s\mu$, $s\mu + \sigma$, $s\mu + 2\sigma$, ..., and $s\mu + n_A\sigma$.

As for the conditional waiting time for a new type B arrival, the analysis is more complicated because it is affected by future type A arrivals. In the following, we revisit Section 5.3 to address that issue. Consider a new type B arrival, and let n_T be the total number of customers in queues, $n_T = n_A + n_B$. We denote by $X^B_{n_T}$ the random variable representing his state-dependent virtual delay in queue. The latter is the time it takes for a server to become free for the customer of interest. In other words, it is the time until the n_A+n_B waiting customers leave the queue (either start service or abandon the queue), plus the time for future type A arrivals to either start service or abandon the queue, plus the duration for a service completion (when all servers are busy). On the contrary to the case with delays information, $X^B_{n_T}$ is no longer dependent on the couple (n_A, n_B) but only on n_T. The reason is that the rate of future type A arrivals is constant (λ_A) and does not depend on the number of type A waiting customers in queue. Furthermore, the discipline of service is workconserving, and type A and B are statistically identical with respect to the memoryless service times and times before reneging. Hence, varying the quantities n_A and n_B so as $n_A + n_B$ is held constant, do not affect the waiting time distribution of the customer of interest.

To characterize $X^B_{n_T}$, one may formulate the problem as to calculate the downcrossing time until the first passage at state 0, starting from state n_T, in a birth-death process with a constant birth rate, $\lambda_m = \lambda_A$ for $m \geq 0$, which represents future type A arrivals during the waiting of the

customer of interest, and with a death rate $\mu_m = s\mu + (m-1)\sigma$ for $m \geq 1$. We do not consider future type B arrivals because the discipline of service within queue B is FCFS. Thereafter using Equations (5.42) and (5.44), we give the expressions of the mean, $E(X^B_{n_T})$, and the variance, $Var(X^B_{n_T})$, of $X^B_{n_T}$ as follows

$$E(X^B_{n_T}) = \frac{1}{\lambda_A}\left(\sum_{i=1}^{n_T+1}\frac{1}{\pi_{i-1}}\sum_{j=i}^{\infty}\pi_j\right), \tag{5.83}$$

$$Var(X^B_{n_T}) = \frac{1}{\lambda_A^2}\left(2\sum_{i=1}^{n_T+1}\frac{1}{\pi_{i-1}}\sum_{j=i+1}^{\infty}\frac{1}{\pi_{j-1}}\left(\sum_{l=j}^{\infty}\pi_l\right)^2 + \sum_{i=1}^{n_T+1}\frac{1}{\pi^2_{i-1}}\left(\sum_{j=i}^{\infty}\pi_j\right)^2\right), \tag{5.84}$$

where the quantities π_i are defined as

$$\pi_0 = 1, \text{ and } \pi_i = \frac{\lambda_A^i}{\prod_{j=0}^{i-1}(s\mu + j\sigma)}, \text{ for } i \geq 1. \tag{5.85}$$

Also, the analysis can be extended to the case of more than 2 customer classes (types) with non-preemptive strict priority. Consider the previous model but with k customer classes, $k \geq 3$. Without loss of generality, we denote each class by the rank i, $1 \leq i \leq k$, according to its priority level (the lower rank for the higher priority.) To get the conditional waiting time characteristics for a new type k arrival ($k \geq 3$), we aggregate all the classes i having the priority over the one of interest (all i such that $1 \leq i < k$) into one class, next we pick up the same analysis as that for type B. The classes aggregation is justified by the fact that the waiting time distribution of the customer of interest is not affected by the order of service of the customers having higher priority.

5.6 Conclusions and Perspectives

We focused on the transient behavior analysis of a general birth-death process. We gave closed-form expressions for the moments of important state-dependent characteristics. The characteristics deal with the random variables of ordinary and conditional first passage times. We derived several new expressions of the moments of the defined hitting and return times. Furthermore, we retrieved some known results as special cases. We also discussed the condition under which a birth-death process is said to be ergodic degree k. In particular, we gave a new explicit expression for the condition of ergodicity degree 3. In the second part of the chapter, we investigated possible applications of the results for Markovian queueing models.

Several further applications could be also possible. For instance, deriving the stationary wait-

ing time moments for some Markovian model where the arrival rate depend on the system state. Concretely, for example in a system where a new customer has a state-dependent probability to join the queue, which is the case for many systems in practice. To do so, we may compute the state-dependent waiting times as shown in this work. Then, we compute the stationary probabilities of the system state using the associated birth-death process. Thereafter, we derive the desired stationary kth order moment of queueing delays, by summing the products of each stationary probability and its corresponding state-dependent kth order moment. Further investigations of these issues should be of value. It would be also interesting to investigate approximations or numerical methods for computing the different quantities. This would be helpful to avoid computation difficulties given that the closed-form expressions of interest are somewhat cumbersome.

Chapter 6

Monotonicity Properties for Multiserver Queues with Finite Waiting Lines

We focus on deriving monotonicity properties for queueing systems. The latter are known to be useful for the modeling and analysis of manufacturing and service systems such as call centers. We consider a markovian multiserver queue with a finite waiting line in which a customer may decide to leave and give up service if its waiting time in queue exceeds its random deadline. We focus on the performance measure in terms of the probability of being served under both transient and stationary regimes. We investigate monotonicity properties of first and second order of this performance with respect to the buffer size, say k. Under the stationary regime, we prove that our service level is strictly increasing and concave in k, whereas we prove under the transient regime that it is only increasing in k. Such results are helpful for optimal design issues.

The paper version of this chapter, Jouini and Dallery [66], is to appear in *Probability in the Engineering and Informational Sciences*.

6.1 Introduction

In this chapter, we derive some monotonicity properties in a queueing system with reneging. We do not address in a direct manner an issue related operations management of call centers. The issue here is different from that in Chapter 5, however in the same sense, we are investigating useful results in an upstream stage above the quantitative analysis of call centers.

Monotonicity properties of performance measures are useful for understanding and solving optimization problems of queueing systems. Optimization models are being used increasingly in the design of a variety of systems where queueing phenomena arise. Examples include flexible manufacturing systems, as well as service systems and telecommunications networks. For such problems, it is important to know the convexity properties of the performance measures with respect to the design variables. For call centers issues, the design variables on which the service provider could act are essentially the staffing level, the arrival rate and the buffer size. In some cases, it could be possible for him to act on processing times (for example by increasing or decreasing the training quality of the agents).

Monotonicity properties may enable us to reduce the performance optimization problem to a convex programming problem which is easier to solve. Using a convexity result, Yao and Shanthikumar [143] accelerate their computation procedure to design a loss queueing system subject to constraints on the loss probability. Koole and Pot [87] consider an optimization problem for an $M/M/s/K+M$ queue. The objective function is a profit function of the number of servers and the buffer size. They derive some monotonicity properties about the defined performance measure. Based on these properties, they develop a fast algorithm which avoids the research of all possible solutions to get the global optimum.

Several convexity properties about various performance measures have been investigated in the queueing literature. The major performance measures for delay systems are the average waiting time, the average queue length and the probability of delay. Those for pure loss systems include basically the probability for a new arrival to be lost. In general, the loss probability is related to systems involving finite buffers or systems with reneging. In this chapter, we consider a queueing system with impatient customers and finite waiting line. The performance measure of interest is the probability for a new arrival customer to enter service. Or equivalently, the probability to not be lost. We investigate first and second order monotonicity properties of our performance measure as a function of the queue size. Note that the design of the buffer size is an important issue in practice. Koole et al. [86] address this problem by investigating the maximum queue length during a busy period for an infinite buffer size.

Another central feature in many practical queueing systems is the reneging phenomenon, i.e.,

one customer may decide to leave the queue (abandons) before starting service. For instance, call abandonment is not negligible in call centers operations. A major drawback in many call center models is assuming customers to be infinitely patient. Garnett et al. [44] show using numerical examples that models with and without abandonment tend to give very different performance measures even if the abandonment rate is small. In this work, we analyze the simplest abandonment model, assuming that the customers patience is exponentially distributed. However, the model is still of interest in practice (especially for call centers) as we have explained in Chapters 3 and 4.

Here is how the rest of the chapter is organized. In Section 6.2, we review the literature close to our work. In Section 6.3, we present the framework of the work: Section 6.3.1 is devoted to formulate the queueing model, and Section 6.3.2 gives definitions and some preliminary results. In Section 6.4, we focus on the first order monotonicity results. In Section 6.4.1, we start by proving two helpful lemmae before proceeding to the main result. Next, we establish using coupling arguments that the transient and stationary probabilities of being served are increasing in the buffer size. In Section 6.4.2, we prove the result for the stationary performance measure using an analytical approach. In Section 6.5, we prove that the stationary probability of being served is strictly concave in the buffer size. Some numerical illustrations of the results are also presented. In Section 6.6, we conclude and propose some directions for future research.

6.2 Literature Review

In this section, we review the literature related to this chapter. We start by presenting some papers investigating monotonicity results for models without reneging. Second, we focus on those for models incorporating reneging.

We classify the results for models with infinitely patient customers into three classes: pure loss, limited buffer and infinite buffer models. For pure loss systems, Harel [53] proves that the throughput of an $M/G/s/s$ is concave in the arrival and service rates. He also characterizes the traffic intensity below which the Erlang loss formula is convex in the arrival rate, and above which it is concave. Furthermore, he shows that the Erlang loss formula is convex in the service rate. For the same model, Messerli [101] proves that the loss probability is a convex function of the number of servers. Additional properties of the loss probability are also discussed by Jagerman [61].

As for systems with limited buffer, Nagarajan and Towsley [102] investigate the convexity of the loss probability in the $M/M/1/K$ queue with respect to the traffic intensity and the service rate. They show that the loss probability is convex in the service rate. However, they prove that

there is a value of the traffic intensity which exactly delineates the convex and concave regions of the loss probability as a function of the traffic intensity. Pacheco [106] considers for his part a more general model with many servers, namely the $M/M/s/K$ queue. He proves that the loss probability is convex in the queue capacity. Meister and Shanthikumar [100] prove many convexity results for tandem queueing systems. Several interesting stochastic comparisons of various variants of multiserver queues with limited buffer are also derived by Berger and Whitt [25].

In what follows, we review some monotonicity results for models with infinite queue capacity. Tu and Kumin [126] prove that the expected number of customers in a $G/G/1$ queue is convex in the service rate. They also show that the result does not hold for a $GI/GI/2$ queue. Surprisingly, Harel [54] show that the expected number of customers in an $M/D/s$ queue is convex in both arrival and service rates. For the $M/M/s$ queue, Lee and Cohen [91] show that the average queue length and the probability of delay, are both convex in the arrival rate. For the same model, Harel and Zipkin [56] establish that the average sojourn time, as well as its standard deviation are convex in arrival and service rates. Again about the $M/M/s$ queue, Jagers and van Doorn [62] focus on the performance measure in terms of the probability for a customer to wait no longer than a given threshold. Note that this service level is widely used in call centers. The authors show that the probability of interest is concave as a function of the number of servers, if the latter is strictly greater than the offered load. We refer the reader for further convexity properties to Weber [131] and [132], Grassmann [45], Shanthikumar [119], Harel and Zipkin [55], Shaked and Shanthikumar [118] and Koole [83].

Let us now turn to the second area of literature related to this chapter. Queues with impatient customers have received a lot of attention in the queueing literature. The results focus especially on performance evaluation. We refer the reader to Chapter 3 for a review of the literature related to that subject. Concerning monotonicity properties, few results were derived. This is due to the mathematical complexities of such problems. Bhattacharya and Ephremides [26] consider multiserver queues with impatient customers. They show that the transient number of lost customers is a monotone function with respect to the arrival rate, the service rate, as well as the reneging rate. Armony et al. [14] consider a holding cost in an $M/M/s$ queue with impatient customers. They prove that this function is decreasing and convex in the service rate and the number of customers. Some sensitivity results for the Erlang-A model can also be found in Whitt [140].

In the present work, we consider an $M/M/s/K+M$ queue. The performance measures of interest are the transient and stationary probabilities of being served. We investigate the

monotonicity properties of first and second order of these service levels with respect to the buffer size. Under the stationary regime, we prove using an analytical approach that the service level is strictly increasing and concave in the buffer size, whereas under the transient regime, we prove that it is only an increasing function. Furthermore, we prove the latter intuitive result using coupling arguments for a more general model, namely the $GI/M/s/K+M$ queue.

6.3 Framework

This section is devoted to formulate the general framework of the research project. First, we describe the queueing system and detail the processes assumptions. Second, we define the performance measures of interest, namely, the fraction of customers who get service under both transient and stationary regimes. We next develop some preliminary results.

6.3.1 Model Formulation

Consider a multiserver queueing system with a single class of customers. The model consists of a set of s parallel, identical servers and a finite queue (waiting line.) There is a maximum number of customers that may be simultaneously present, we assume that the system can hold at most a total of K customers including those in service. Clearly $K \geq s$, and we denote the queue capacity by $k = K - s$, $k \geq 0$. The system is operated in such a way that at any time, any customer can be addressed by any server. So upon arrival, a customer is addressed by one of the available servers, if any. If not, the customer joins the queue if less than K customers are present in system. If not, the customer is refused entry and departs immediately without service. He is blocked and considered lost. In addition, we assume that customers are impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time, he will renege (abandon), and again considered to be lost. Finally, retrials are ignored, and reneging is not allowed once a customer starts his service.

The arrival of customers is assumed to follow a Poisson process. Interarrival times are i.i.d. and exponentially distributed with rate λ. Successive service times are assumed to be i.i.d., independent from the arrival process, and follow an exponential distribution with rate μ. Times before reneging are assumed to be i.i.d., and exponentially distributed with rate γ. Following similar arguments, the system can be modeled as an $M/M/s/K+M$ queue. Note that owing to reneging, the system is always ergodic even if the queue has infinite capacity. Also, ergodicity would always be assured for our system because of its limited capacity, even if the customers were assumed to be infinitely patient. In conclusion, the system we consider here is unconditionally ergodic.

6.3.2 Preliminaries

In this section, we focus on characterizing the performance measure of interest. It is defined in terms of the fraction of customers who get service, i.e., the fraction of customers who are not blocked and who do not renege.

Let us consider an interval of time $[0, t]$, $t > 0$. We initially assume that the system starts empty. Given that t units of time have elapsed, let $n(t)$, and $s(t)$ be the total number of arrivals (including blocked customers), and the number of those who enter service, respectively. We define the transient fraction of customers who enter service, $Q(t)$, during $[0, t]$ as $Q(t) = \frac{s(t)}{n(t)}$. The reader should not confuse the notation Q here with that of the probability of being lost analyzed in Chapter 3. Taking the limit as $t \to \infty$ of $Q(t)$ leads to the stationary fraction Q of successful departures, $Q = \lim_{t\to\infty} Q(t)$. Let $b(t)$ and $r(t)$ be the number of blocked customers, and that of those who renege, respectively. Since the quantities $s(t)$ and $n(t)-b(t)-r(t)$ coincide in the long-run, then Q can be rewritten as

$$Q = \lim_{t\to\infty} \frac{n(t) - b(t) - r(t)}{n(t)}. \tag{6.1}$$

In what follows, we derive a closed-form expression for Q. We denote the system state by a random variable taking non-negative integer values representing the total number of customers in system (including those in service.) The quantity Q represents the probability in the infinite horizon for a new arrival customer to enter service, which involves system states stationary probabilities seen by that arrival. From the PASTA property which holds for our system, it is equivalent to consider the system states stationary probabilities seen by an outside random observer (at a random instant.)

Let us now come back to Equation (6.1) by dividing both the numerator and the denominator in the right hand side over t. Computing Q may be reduced thereafter to computing separately the ratios $n(t)/t$, $b(t)/t$ and $r(t)/t$ as t goes to ∞. Recall that the mean number of customers per unit of time is λ. Hence in the long-run (as $t \to \infty$), the ratio $n(t)/t$ converges by construction to λ. As for the limit of $b(t)/t$ as $t \to \infty$, we may recognize it as the probability for a new arrival to be blocked times the mean arrival rate λ. So, it is the probability that a new arrival finds a full system times λ, namely the quantity $\lambda p(K)$. Let us now focus on the limit of $r(t)/t$ as t goes to infinity. One may recognize this quantity as the mean number of reneging per unit of time seen by a random outside observer. Since it takes in average $1/\gamma$ units of time for one customer waiting in queue to renege. Thus as $t \to \infty$, $r(t)/t$ converges to the mean number of customers in queue (in the distant future) times γ. Based on the previous analysis, Q can be rewritten as

follows.

$$Q = 1 - p(K) - \frac{\gamma}{\lambda} \sum_{i=s+1}^{K} (i-s)p(i). \tag{6.2}$$

To get explicitly the expression of Q, we move on to compute the stationary probabilities $p(i)$, for $0 \leq i \leq K$. In the usual way, we model our system as a finite continuous-time birth-death process with discrete state space taking non-negative integer values ranging from 0 to K and defined on a probability space. The birth rates are constant and equal to λ. The death rates are state-dependent; when moving from state i to state $i-1$, the death rate is $i\mu$ for $1 \leq i \leq s$, and it is $s\mu + (i-s)\gamma$ for $s < i \leq K$. Under the stationary regime, we easily get a set of K recursive equations relating $p(i)$ and $p(i+1)$ for $0 \leq i \leq K-1$. Proceeding to solve by iteration leads to

$$p(i) = \frac{\lambda^i}{i!\,\mu^i}\, p(0) \text{ for } 0 \leq i \leq s, \text{ and } p(i) = \frac{\lambda^i}{s!\mu^s \prod_{j=1}^{i-s}(s\,\mu + j\,\gamma)}\, p(0) \text{ for } s < i \leq K, \tag{6.3}$$

where $p(0)$ is the steady state probability to have no customers in system, and obviously, $p(i) = 0$ for $i > K$. Then, we couple the last set of equations with the probability conservation relation, i.e., $\sum_{i=0}^{\infty} p(i) = 1$, to get

$$p(0) = \left(\sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i} + \frac{\lambda^s}{s!\,\mu^s} \sum_{i=s+1}^{K} \frac{\lambda^{i-s}}{\prod_{j=1}^{i-s}(s\,\mu + j\,\gamma)} \right)^{-1}, \tag{6.4}$$

which determines all stationary probabilities. We still have to substitute them into Equations (6.2) to obtain Q.

6.4 Proof of First Order Monotonicity Property

One may intuitively state that the performance measures $Q(t)$ and Q increase with respect to the queue capacity k, keeping the parameters λ, μ, γ and s constant. The idea is that, although adding more places in the waiting line may increase abandonments, it is clear that it could not deteriorate the performances we consider here. On the contrary, it allows for more customers to enter service. If not, it will at worst achieve an equal fraction of successful departures comparing to a system with less queue capacity. In this section, we rigorously prove these results using two different approaches. In Section 6.4.1, we prove using coupling arguments that $Q(t)$ and Q increase in k for a more general case, namely for a $GI/M/s/K+M$ queue. In Section 6.4.2, we consider our original system (the $M/M/s/K+M$ queue) and prove using an analytical approach that Q increases in k.

6.4.1 Sample Path Approach

We start with a tangential development that will be of a great help to prove our main result. Let us relax some assumptions in our original system by considering a $GI/GI/s/K+M$ queue. We assume that interarrivals and service times are i.i.d., but we allow them to follow a general distribution. In Lemma (6.1), we present an interesting result about the relation between the performance measures of interest and the scheduling policy under which the system is working. For the rest of the chapter, we denote by Π the set of workconserving non-preemptive scheduling policies.

The following result in Lemma (6.1) can be seen as an extension of that in Theorem (3.2) of Chapter 3. In the latter, we proved a conservation result for the probability of being lost. Here, we prove the conservation result for the probability of being served by adding blocking (whenever the system is full).

Lemma 6.1 *Consider a $GI/GI/s/K+M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. Then, the probability of being served Q is constant for any workconserving non-preemptive scheduling policy.*

Proof. The result is trivial for a queue with no capacity or with capacity 1. In such cases, it is clear that the system behaves identically for any policy $\pi \in \Pi$, and as a consequence, Q remains constant. Otherwise, for $k \geq 2$, we prove the result by coupling arguments. Consider two $GI/GI/s/K+M$ models, say Model 1 and Model 2. We assume that Model 1 and Model 2 have identical parameters except for the scheduling policies. Model 1 and Model 2 are working under the policies π_1 and π_2, respectively, such that $\pi_1 \in \Pi$, $\pi_2 \in \Pi$, and $\pi_1 \neq \pi_2$. Our approach is based on a single sample path. In both models, we create identical successive arrival epochs, as well as identical successive service times. However, since times before reneging are exponentially distributed, the decision for one customer to abandon the queue is not affected by his elapsed waiting time. This is stochastically equivalent to create randomly, for our sample path, a new maximum time of patience for each customer in queue after each selection for service epoch (or equivalently after each successful departure epoch.) Assume that at time $t = 0$ both systems are empty, and let work begins.

Both models behave identically until a busy period starts and the following situation occurs: a server becomes idle (after a service completion) and more than one customer are waiting in queue. Let D_i be the epoch of that service completion (which occurs simultaneously in Model 1 and 2.) For both models, let n be the number of customers in queue just before D_i, $2 \leq n \leq k$. At D_i, the idle server in Model 1 selects one customer from the queue who can be different from

that selected by the same idle server in Model 2. However, the number of customers in queue goes down by 1 for both models, it becomes $n-1$. Recall that up to now, the number of blocked customers, as well as that of those who abandoned the queue, are identical for both models.

At the epoch D_i, we create for each customer, waiting in the queue of Model 1, a new patience time. Without altering distributions, since times before reneging are identically distributed, we create the same set of $n-1$ maximum patience times as in Model 1, and we assign them arbitrary to the customers waiting in Model 2. After D_i, three events are possible: one customer reneges, or a new arrival occurs, or a server becomes idle (service completion.) Recall that by construction of our single sample path, these events occur simultaneously in both models. Assume now that the first event occurs, then the number of customers who abandon the queue goes up by 1 in both models and as a consequence remains identical for them. It is still also identical if another customer abandons the queue. In general, it is the case as long as there are customers waiting in queue. If not, both models will behave identically, anyway. Assume now that a new arrival occurs. Note that the number of customers in queue is the same in both models. If the queues are currently full, i.e., k are waiting for service, then the new arrival will be blocked, and systems states remain unchanged. However, if at least there is one available space, hence, the number of customers in queue goes up by 1 in both models. Note that if another arrival occurs or that one customer abandons the queue, then, the number of customers in queue will increase by 1 (or remains unchanged if the system is full) or will decrease by 1, respectively. The main conclusion is that the number of blocked customers, as well as that of those who abandon the queue will vary identically in both models.

Assume now that one server becomes idle. If the number of customers in queue is currently less or equal to 1, it is obvious to see that policies π_1 and π_2 will select at the same time the unique available customer, if any. Otherwise, the busy period ends in both models, so both policies will again select identically new arrivals for service until the beginning of the next busy period. However, if the number of customers in queue is greater or equal to 2, the selected customer for service may be different from one model to another. As above, we create for the remaining set of waiting customers, the same set of patience time. Again, we can state that the number of blocked customers, as well as that of those who abandon the queue remains the same for both models.

Carrying on using the same arguments, we state that in a distant future, the number of blocked customers and that of those who renege in Model 1 coincide with those in Model 2. Since the number of arrivals are also equal for both models, the service level in terms of the fraction of successful departures is unchanged, $Q_{\pi 1} = Q_{\pi 2}$. This completes the proof. □

Although the probability of being served is independent of the scheduling policy, the mean waiting time in queue for the served customers does depend on the scheduling policy. We have proved the latter result in Theorem (3.3) of Chapter 3 by considering the particular case of a $GI/GI/s + M$ queue. We have also characterized the policies under which upper and lower bounds of the mean waiting time are achieved.

We should note however that the result in Lemma (6.1) does not hold if times before reneging are not i.i.d. and exponentially distributed, or if service times at any point during an arbitrary busy period are order of service dependent, we need to assume that no service needs are created or destroyed within the system: no renege in the midst of service, no forced idleness of servers, and so on.

In Lemma (6.2), we show that Q is still unchanged for any workconserving scheduling policy (with preemption or not) if we further assume that service times are i.i.d. and exponentially distributed.

Lemma 6.2 *Consider a $GI/M/s/K + M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. Then, the probability of being served Q is constant for any workconserving scheduling policy.*

Proof. We again show the result using coupling arguments. Based on a single sample path, we compare the quantity Q for two identical $GI/M/s/K + M$ models, say Model 1 and Model 2, working under two different scheduling policies π_1 and π_2, respectively. We assume that π_1 and π_2 are workconserving, and do not restrict them to be non-preemptive. We use a similar approach to that for Lemma (6.1). The only difference is only when an interruption of service occurs in one of the models. Note that just before the epoch of that event, both models are identical: all servers are busy, same number n of customers in queue, same remaining service times, and same set of remaining times before reneging for waiting customers in queue. Without loss of generality, assume that in Model 1, a new arrival interrupts the service of a customer currently in service. Since service times are assumed to be exponentially distributed, then the remaining time for a service completion is not affected by the elapsed time in service. This allows us to create randomly, for our sample path, a new set of remaining service times for the customers currently in service (s customers in both models), and also a new set of patience time for waiting customers in queue (n customers in both models.) Continuing the sample path comparison in the long run will subsequently show that Q coincides for both models. This completes the proof of the lemma. □

In Theorem (6.1), we show the main result of first order monotonicity for a $GI/M/s/K+M$ queue. The analysis resorts in part to the previous preliminary results of this section.

Theorem 6.1 *Consider a $GI/M/s/K+M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. Then, probability of being served Q is strictly increasing in the buffer size k.*

Proof. To prove the result of the theorem, it suffices to compare the achieved Q for the two following systems. The first, say Model 1, is a $GI/M/s/K+M$ queue with k waiting spaces. From Lemma (6.1), it does not restrict generality to assume that Model 1 works under the FCFS discipline of service. The second model is identical to the first in all parameters except that it has $k+1$ waiting spaces. From Lemma (6.2), the latter is equivalent, in terms of the achieved Q, to a $GI/M/s/K+M$ queue, say Model 2, with $k+1$ waiting spaces and working under any preemptive workconserving policy. In summary, it is left to establish that the stationary probability of being served in Model 2, say Q_2, is strictly greater than that in Model 1, say Q_1.

The proof follows the sample path approach. Before proceeding to the details, let us characterize a specific preemptive workconserving policy, say π, under which Model 2 is operated. We divide the queue in Model 2 (with capacity $k+1$) into two virtual queues. The first, say queue 1, has capacity k. The second, say queue 2, has the remaining capacity, i.e., 1. Upon arrival, a customer is addressed by one of the available servers, if any. If not, the customer must join one of the queues. We will specify the queue joining policy later. Customers in queue 1 have priority over customers in queue 2 in the sense that servers are handling customers belonging to queue 1 first. The priority rule is preemptive, which simply means that a server currently serving a customer pulled from queue 2, while a new arrival customer joins queue 1, will interrupt this service and turn to queue 1 customer. Within each queue, customers are served in order of their arrival, that is, under the FCFS discipline.

Let us now couple Model 1 and 2 and let work begins. Both models behaves identically until the situation where in Model 1 all servers are busy, k customers are waiting in queue and a new arrival occurs. Let us stop our clock temporarily. We denote that customer by the “low customer". Clearly, the “low customer" is blocked in Model 1 because the system is currently full, however, he joins the waiting line in Model 2. We assign him to queue 2 (with lower priority.) Recall that up to now the number of customers served is identical in both models. Let our clock resumes ticking: arrivals, blocking, abandonments, as well as service completions will occur at the same epochs in both models until the busy period in system 1 ends (which occurs with probability 1 due to the ergodicity condition.) We distingue two possible cases for the “low customer": either he has meanwhile abandoned, or he is still waiting in queue 2. In the first case,

both systems states become again identical. In the second case, i.e., if the “low customer" is still waiting, then we assign him to the server currently idle in Model 2. As long as the current idle period in Model 1 does not finish, we let the “low customer" stay in service. If the “low customer" finishes his service before that a new arrival occurs (at the same epoch in both models), therefore Model 2 will have one more service completion comparing to Model 1, and all events in both models become again identical. If not, that is if the idle period in Model 1 ends and the “low customer" has not successfully leaved Model 2, then we interrupt his service and we put him back in queue 2. The idea here from choosing the policy π is to ensure an identical behavior, in both models, with regard to all customers except for the “low customers". Such customers are blocked (lost) in Model 1, however, they join queue 2 in Model 2.

From the previous arguments, one may easily deduce that $Q_1 \leq Q_2$. Let us now proceed to establish that $Q_1 < Q_2$. It is clear that one “low customer" at most may be present in Model 2 at a given observation moment. Let us further define a particular cycle duration referred to as the “low cycle". The “low cycle" starts when a “low customer" enters Model 2, and terminates upon the arrival of the next “low customer". The latter allows the following “low cycle" to start, and so on. The duration of a “low cycle" is given by the time it takes so that the “low customer" who starts the cycle either reneges or successfully finishes his service plus the time it takes starting from that epoch until the next “low customer" arrival epoch. Since the systems we consider here are stable, hence, any busy period in Model 1 ends with probability 1, i.e., its duration is finite ($< \infty$). In addition, knowing that times before reneging are finite, we state that the “low cycle" duration is also finite. Furthermore, since interarrival times, times before reneging, as well as service times are i.i.d. and further independent of each others, it then follows that “low cycles" durations are also independent and identically distributed. Next, assuming the stationary regime and observing that there is a non-zero probability that a “low customer" finishes successfully his service within its corresponding “low cycle", it yields from the Law of Large Numbers that there is a non-zero proportion of “low customers" that will finish successfully their service. So, we state that the number of customers being served in Model 2 is strictly greater than that in Model 1. Finally, it is implied that the stationary probability of being served Q is strictly increasing in the buffer size k. This completes the proof. □

In a parallel to the proof of Theorem (6.1), we also state that the probability of being served under the transient regime, $Q(t)$, is an increasing function of k. Note that it is not necessarily strictly increasing in k as it is the case for the quantity Q.

6.4.2 Analytical Approach

In this section, we again consider our original $M/M/s/K+M$ queue described in Section 6.3.1. As shown in Section 6.3.2, a closed-form expression of the quantity Q may be derived. This allows us to again prove the result of Theorem (6.1) using an analytical approach. The analysis we address in this section is in particular useful for the proof of the convexity result in Section 6.5. Before giving the details of the proof of the monotonicity property, we begin with some preliminary results by means of Properties 6.1 and 6.2. For the rest of the chapter, an empty sum is being interpreted as zero, and an empty product is being interpreted as one.

Our objective is to show that Q is strictly increasing in k for an $M/M/s/K+M$ queue. To do so, we consider two models. The first is an $M/M/s/K+M$ queue with parameters λ, γ, μ, s, and k waiting spaces, $k \geq 0$. The second model is identical to the first however it has a larger buffer with $k+1$ waiting spaces. Recall that for our analysis, we do not need to specify the scheduling policy except that it is workconserving. Next, it suffices to show that the stationary probability of being served in the first model, say Q_k, is strictly lower than that in the second model, say Q_{k+1}. Or equivalently, if we introduce the sequence $\{U_k,\ k \geq 0\}$ defined as $U_k = Q_{k+1} - Q_k$, it remains for us to establish that $U_k > 0$ for all $k \geq 0$. From Equation (6.2), U_k can be rewritten as

$$U_k = p_k(k+s) - p_{k+1}(k+s+1) + \frac{\gamma}{\lambda}\left(\sum_{i=s+1}^{k+s}(i-s)p_k(i) - \sum_{i=s+1}^{k+s+1}(i-s)p_{k+1}(i)\right). \tag{6.5}$$

The stationary probabilities are given by Equations (6.3) and (6.4). The subscripts are to indicate to which system the stationary probabilities are corresponding, either for the one with queue capacity k, or for that with queue capacity $k+1$. In Property (6.1), we state a useful relation between U_k and U_{k+1} for any non-negative integer k.

Property 6.1 *For all $k \geq 0$, the following holds*

$$U_{k+1} = \frac{s!\mu^s \sum_{i=0}^{s}\phi_i + \sum_{i=s+1}^{k+s}\rho_i}{s!\mu^s \sum_{i=0}^{s}\phi_i + \sum_{i=s+1}^{k+s+2}\rho_i} \cdot \frac{\lambda}{s\mu + (k+2)\gamma} \cdot U_k, \tag{6.6}$$

where

$$\phi_i = \frac{\lambda^i}{i!\mu^i},\ \textit{for } i \geq 0,\ \textit{and, } \rho_i = \frac{\lambda^i}{\prod_{j=1}^{i-s}(s\mu + j\gamma)},\ \textit{for } i \geq s+1. \tag{6.7}$$

Proof. The proof is provided in Appendix C.1.

Note that proving Property (6.1) represents the "hard" part of the proof of the monotonicity result as well as that of the convexity result. One may verify that Equation (6.6) holds for different special cases. For instance, let us consider an infinite-server queue $M/M/s/K+M$

($s \to \infty$). Taking the limit in Relation (6.6) as s goes to ∞ implies that $U_k = 0$ for all non-negative integer $k \geq 1$ (in addition from Equation (C.4) for example, we have $U_0 = 0$), which obviously agrees with the classical queueing results. The result also holds for an $M/M/s/K+M$ queue with infinitely impatient customers ($\gamma = \infty$). In that case, the $M/M/s/K+M$ queue is equivalent to a loss system (without waiting space.) Thus, it is easy to see that the quantity Q_k does not depend on the buffer size k. So, $U_k = 0$ for any $k \geq 0$, which agrees with Equation (6.6).

Although we present in Property (6.2) an inequality that directly seems to be of independent interest, it is useful for forthcoming proofs of our results.

Property 6.2 *Let λ and μ be strictly positive reals and let $\{N_s,\, s \geq 1\}$ be a sequence defined as*

$$N_s = s\mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} - \lambda \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i}. \tag{6.8}$$

Then, $N_s > 0$ for all $s \geq 1$.

Proof. The inequality holds by induction. We have $N_1 = \mu > 0$, then Property (6.2) holds for $s = 1$. Assume now that $N_s > 0$ for a given $s \geq 1$, and let us show that $N_{s+1} > 0$. From Equation (6.8), N_{s+1} can be written as

$$\begin{aligned} N_{s+1} &= (s+1)\mu \sum_{i=0}^{s+1} \frac{\lambda^i}{i!\mu^i} - \lambda \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} \\ &= s\mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + s\mu \frac{\lambda^{s+1}}{(s+1)!\mu^{s+1}} + \mu \sum_{i=0}^{s+1} \frac{\lambda^i}{i!\mu^i} - \lambda \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i} - \frac{\lambda^{s+1}}{s!\mu^s} \\ &= N_s + \frac{s\,\lambda^{s+1}}{(s+1)!\mu^s} - \frac{\lambda^{s+1}}{s!\mu^s} + \mu \sum_{i=0}^{s+1} \frac{\lambda^i}{i!\mu^i} \\ &= N_s - \frac{\lambda^{s+1}}{(s+1)!\mu^s} + \mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \frac{\lambda^{s+1}}{(s+1)!\mu^s} \\ &= N_s + \mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i}. \end{aligned} \tag{6.9}$$

Using the induction assumption, it thus follows that $N_{s+1} > 0$. Finally, we conclude that $N_s > 0$ for all $s \geq 1$. This completes the proof. □

In Theorem 6.2, we state the main result of this section. Having Properties (6.1) and (6.2), we are now ready to establish the first order monotonicity property of the probability of being served, Q, with respect to the buffer size k.

Theorem 6.2 *Consider an $M/M/s/K+M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. Then, Q is strictly increasing in the buffer size k.*

Proof. As explained in the beginning of this section, proving the theorem is equivalent to proving that U_k is strictly positive for $k \geq 0$. Keeping the parameters λ, μ, s and γ constant, the result holds by induction on k.

Let us establish our claim for the first rank, $k = 0$. The quantity U_0 is given by $U_0 = Q_1 - Q_0$, where Q_0 and Q_1 are the probabilities of being served for the $M/M/s/s+M$ (no waiting space) and $M/M/s/s+1\ +M$ (single waiting space) systems, respectively. Using Equation (6.2), the probability Q_0 is given by $Q_0 = 1 - p_0(s)$, where $p_0(s)$ is the stationary probability to have s customers in the $M/M/s/s+M$ system. As for Q_1, it is given by $Q_1 = 1 - \frac{\lambda+\gamma}{\lambda}\, p_1(s+1)$, where $p_1(s+1)$ is the stationary probability to have $s+1$ customers in the $M/M/s/s+1+M$ system.

From Equations (6.3) and (6.4), we get

$$Q_0 = 1 - \frac{\frac{\lambda^s}{s!\,\mu^s}}{\sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i}}, \tag{6.10}$$

and,

$$Q_1 = 1 - \frac{\lambda+\gamma}{s\mu+\gamma} \frac{\frac{\lambda^s}{s!\,\mu^s}}{\sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i} + \frac{\lambda^s}{s!\,\mu^s}\,\frac{\lambda}{s\,\mu+\gamma}}. \tag{6.11}$$

Therefore,

$$U_0 = \frac{\lambda^s}{s!\,\mu^s\ \sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i}} - \frac{\lambda^{s+1} + \lambda^s\,\gamma}{s!\,\mu^s\,(s\,\mu+\gamma)\ \sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i} + \lambda^{s+1}}. \tag{6.12}$$

To prove that $U_0 > 0$, we consider U_0 as a real function of γ, for $\gamma \geq 0$, and we study the sign of $U_0(\gamma)$. It is clear that U_0 has the property to be continuous and derivable in γ. Taking the derivative, U_0', of U_0 in γ leads to

$$\begin{aligned}
U_0'(\gamma) &= -\frac{s!\mu^s s\mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + s!\mu^s\gamma \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1} - (\lambda+\gamma)s!\mu^s \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i}}{(s!\mu^s(s\mu+\gamma)\sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1})^2} \cdot \lambda^s \\
&= -\frac{s!\mu^s s\mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + s!\mu^s\gamma \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1} - (\lambda+\gamma)(s!\mu^s \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i} + \lambda^s)}{(s!\mu^s(s\mu+\gamma)\sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1})^2} \cdot \lambda^s \\
&= -\frac{s!\mu^s s\mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + s!\mu^s\gamma \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i} + \lambda^s\gamma + \lambda^{s+1}}{(s!\mu^s(s\mu+\gamma)\sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1})^2} \cdot \lambda^s \\
&\qquad + \frac{\lambda^{s+1} + \lambda s!\mu^s \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i} + \lambda^s\gamma + s!\mu^s\gamma \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i}}{(s!\mu^s(s\mu+\gamma)\sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1})^2} \cdot \lambda^s \\
&= -\lambda^s s!\mu^s \frac{s\mu \sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} - \lambda \sum_{i=0}^{s-1} \frac{\lambda^i}{i!\mu^i}}{(s!\mu^s(s\mu+\gamma)\sum_{i=0}^{s} \frac{\lambda^i}{i!\mu^i} + \lambda^{s+1})^2}.
\end{aligned} \tag{6.13}$$

Using the notation in Equation (6.8), $U_0'(\gamma)$ can be rewritten as

$$U_0'(\gamma) = \frac{-\lambda^s s! \mu^s}{(s!\mu^s(s\mu+\gamma)\sum_{i=0}^{s}\frac{\lambda^i}{i!\mu^i} + \lambda^{s+1})^2} \cdot N_s. \tag{6.14}$$

Applying now Property (6.2) for s strictly positive integer, and, for λ and μ strictly positive reals, we easily see that $U_0'(\gamma) < 0$. Then, U_0 is a strictly decreasing function in γ, for $\gamma \geq 0$. Hence, it follows that

$$U_0(\gamma) > \lim_{\gamma \to +\infty} U_0(\gamma), \text{ for } \gamma \geq 0. \tag{6.15}$$

Observing that

$$\lim_{\gamma \to +\infty} U_0(\gamma) = \frac{\lambda^s}{s!\,\mu^s\,\sum_{i=0}^{s}\frac{\lambda^i}{i!\,\mu^i}} - \frac{\lambda^{s+1} + \lambda^s\,\gamma}{s!\,\mu^s\,(s\,\mu+\gamma)\,\sum_{i=0}^{s}\frac{\lambda^i}{i!\,\mu^i} + \lambda^{s+1}} = 0, \tag{6.16}$$

we deduce that $U_0(\gamma) > 0$ for $\gamma \geq 0$. Thereafter, our claim is true for the first rank $k = 0$.

Let us consider $k \geq 0$ and assume that our claim is true for the rank k, i.e., $U_k > 0$. Let us now prove that our claim is true for the rank $k+1$. This is a direct consequence of Property 6.1. For $s \geq 1$, $\lambda, \mu > 0$ and $\gamma \geq 0$, we state using Property 6.1 that U_{k+1} is the product of U_k and a strictly positive real. So, $U_{k+1} > 0$. Finally, we conclude that $U_k > 0$ for $k \geq 0$, which completes the proof of the theorem. □

6.5 Proof of Second Order Monotonicity Property

In this section, we investigate the second order property of monotonicity (of the probability of being served) in the queue capacity. First, we prove using a simple counterexample that the transient probability of being served, $Q(t)$, is not concave in k. Second, we state our main result in Theorem (6.3) about the concavity property. Finally, we present some numerical illustrations of that result.

To prove the non-concavity of $Q(t)$ as a function of k, we consider three $M/M/1/K+M$ queues denoted by Model 1, Model 2 and Model 3. Assume the discipline of service to be FCFS. The models are identical in all parameters except for the buffer size. Specifically, Models 1, 2 and 3 contain 1, 2 and 3 waiting spaces, respectively. During an interval of time $[0, t]$, we denote the transient probability of being served for Model 1 by $Q_1(t)$. We denote those for Model 2 and 3 by $Q_2(t)$ and $Q_3(t)$, respectively. In what follows, we construct one possible sample path which shows that the transient probability of being served is not concave in k. In mathematical terms, it consists to find an instant t such that $Q_3(t) - Q_2(t) > Q_2(t) - Q_1(t)$.

Initially, the models are empty. Now, let work begins. All models behave identically until

the situation where in each model the unique available server is busy and there is one waiting customer in queue, say A_1. Thereafter, assume that one arrival, say A_2, occurs before a service completion or an abandonment. Note that this event occurs with a non-zero probability. The customer A_2 is blocked in Model 1, whereas he joins the queue in Models 2 and 3. Assume also that the next event is an arrival denoted by A_3. The customer A_3 is blocked in Models 1 and 2, however he joins the queue in Model 3. Next, assume that A_2 abandons the queue, which occurs simultaneously in Models 2 and 3. Then, assume that A_1 in all models and A_3 in Model 3 finish their service and successfully leave the systems. Let t_{A_3} be the epoch of the departure of A_3. So, we state that during $[0, t_{A_3}]$, the number of served customers in Model 1 is equal to that in Model 2. However, there is one served customer in more in Model 3 compared to the other models. In other words, $Q_3(t) > Q_2(t)$ and $Q_1(t) = Q_2(t)$, which leads to the inequality $Q_3(t) - Q_2(t) > Q_2(t) - Q_1(t)$ and closes the discussion.

Turning now to the concavity of the stationary quantity Q as a function of k, we present the following theorem.

Theorem 6.3 *Consider an $M/M/s/K+M$ queue. Times before reneging are assumed to be i.i.d. and exponentially distributed. Then, Q is a strictly concave function in the buffer size k.*

Proof. Let us again consider three $M/M/s/K+M$ queues denoted by Model 1, 2 and 3. All models are identical in all parameters except in the buffer size. In Model 1, there are k waiting spaces. However, Model 2 and Model 3 have $k+1$ and $k+2$ waiting spaces, respectively. We do not need here to specify the scheduling policy except that it is workconserving. For Model 1, 2 and 3, we denote by Q_k, Q_{k+1} and Q_{k+2} the stationary probabilities of being served, respectively. Following this introduction, one may easily see that proving our theorem is equivalent to proving that $U_k = Q_{k+1} - Q_k$ is strictly greater than $U_{k+1} = Q_{k+2} - Q_{k+1}$, for all $k \geq 0$. In other terms, it remains to prove that the sequence $\{U_k,\ k \geq 0\}$ is strictly decreasing. Knowing from Theorem (6.2) that $U_k > 0$, for $k \geq 0$, it suffices thereafter to show that $\frac{U_{k+1}}{U_k} < 1$, for $k \geq 0$. From Equation (6.6), we have

$$\frac{U_{k+1}}{U_k} = \frac{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+2} \rho_i} \cdot \frac{\lambda}{s\mu + (k+2)\gamma}, \text{ for } k \geq 0, \tag{6.17}$$

which may be rewritten, for $k \geq 0$, as

$$\frac{U_{k+1}}{U_k} = \frac{\left((s!\mu^s \sum_{i=0}^{s-1} \phi_i) \times \frac{\lambda}{s\mu+(k+2)\gamma}\right) + \left((\lambda^s + \sum_{i=s+1}^{k+s} \rho_i) \times \frac{\lambda}{s\mu+(k+2)\gamma}\right)}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+2} \rho_i}. \tag{6.18}$$

From the one hand, Property (6.2) leads to

$$\lambda \sum_{i=0}^{s-1} \phi_i < (s\mu + (k+1)\gamma) \sum_{i=0}^{s} \phi_i. \tag{6.19}$$

Hence,

$$(s!\mu^s \sum_{i=0}^{s-1} \phi_i) \times \frac{\lambda}{s\mu + (k+1)\gamma} < s!\mu^s \sum_{i=0}^{s} \phi_i. \tag{6.20}$$

From the other hand, we have for all i, such that $i < k + s + 1$,

$$\frac{\lambda}{s\mu + (k+2)\gamma} < \frac{\lambda}{s\mu + (i-s+1)\gamma}, \tag{6.21}$$

which implies

$$\frac{\lambda}{s\mu + (k+2)\gamma} \rho_i < \frac{\lambda}{s\mu + (i-s+1)\gamma} \rho_i = \rho_{i+1}. \tag{6.22}$$

Summing Equation (6.22) on all i, $s+1 \leq i \leq k+s$, we get

$$\frac{\lambda}{s\mu + (k+2)\gamma} \sum_{i=s+1}^{k+s} \rho_i < \sum_{i=s+1}^{k+s} \rho_{i+1} = \sum_{i=s+2}^{k+s+1} \rho_i. \tag{6.23}$$

Next, observing that

$$\frac{\lambda^{s+1}}{s\mu + (k+2)\gamma} < \frac{\lambda^{s+1}}{s\mu + \gamma} = \rho_{s+1}, \text{ for } k \geq 0, \tag{6.24}$$

the summation of both Inequalities (6.23) and (6.24) leads to

$$(\lambda^s + \sum_{i=s+1}^{k+s} \rho_i) \times \frac{\lambda}{s\mu + (k+2)\gamma} < \sum_{i=s+1}^{k+s+1} \rho_i < \sum_{i=s+1}^{k+s+2} \rho_i. \tag{6.25}$$

Finally, it remains to apply Relations (6.20) and (6.25) back into Relation (6.18), to state that $\frac{U_{k+1}}{U_k} < 1$. This completes the proof of the theorem. □

In simple words, one rule of thumb of the current chapter would be as follows. Consider a system that could be modeled as an $M/M/s/K+M$. Assume that the manager has to design the queue capacity subject to maximizing the throughput of the system. Then, there is no need to choose a very large queue capacity. Most of the benefits are, indeed, achieved with a small queue size.

To get some numerical illustrations of our results, we consider various $M/M/s/K+M$ models by taking a broad range of parameters values. The service rate is unchanged for all chosen examples, $\mu = 1$. The values of the reneging rate are 0.5, 1 and 2. The number of servers are 1,

2, 3, 5, 10, 15, 50, 70 and 100. To vary the "servers utilization" calculated as $\lambda/s\mu$, we consider $\lambda = 1.8$ for $s = 1$, 2 and 3; $\lambda = 8$ for $s = 5$, 10 and 15; $\lambda = 60$ for $s = 50$, 70 and 100. For each set of the previous values, the buffer size is ranging from 0 to 30. The detailed results are presented in Tables C.1, C.2 and C.3 of Appendix C.2.

From the numerical results, we underline the following comments. As expected from Theorems (6.2) and (6.3), Q is increasing and concave in k keeping all remaining parameters constant. One may see that there is no need to go beyond a buffer size around 10 to approximately reach the maximum of the probability of being served (reached within an infinite buffer size.) Starting from a system with no waiting space, most of the improvements are achieved by adding two places in the buffer. Obviously, we also see that Q is decreasing with respect to the abandonment rate γ. The reason is simply that the probability to abandon the queue is increasing in the abandonment rate. Furthermore, for a fixed server utilization, large systems allow to achieve higher service levels. This does not seem at odds with known results, it is due to the pooling effect. We refer the reader to Chapter 2 for further details on the subject.

6.6 Conclusions and Further Research

In this chapter, we considered a queueing system with reneging and finite buffer size. The model is of interest for the modeling in practice of several systems with impatient customers, such as call centers. We investigated monotonicity results of the probability of being served with respect to the buffer size. These results are helpful when addressing optimizations issues. We considered both transient and stationary quantities of the performance of interest. Under the transient regime, we proved that it is an increasing and non-concave function of the buffer size. Under the stationary regime, we proved that it is strictly increasing and concave in the buffer size.

As a topic for future research, it would be interesting to investigate in a similar fashion as here, the convexity properties of the performance measure as a function of other parameters such as the arrival rate, service rate, reneging rate, and number of servers. The interest on some design variables instead of others should depend on the application. For instance, a call center manager would be more interested by the analysis with respect to the arrival rate and the number of servers. In most practical cases, he could be able to increase or decrease the staffing level, and also to act on the arrival rate (overflows of customers). In a manufacturing application, the manager could be however able to act on the processing time of servers (facilities), which is kind of difficult for a call center manager.

Chapter 7

Conclusion and Perspectives

In this chapter, we give general concluding remarks and present directions for future research. For further details, we refer the reader to the concluding sections of the previous chapters.

7.1 Conclusions

A call center, or in general a contact center, is defined as a service system in which agents serve customers, over telephone, fax, email, etc. The call center industry has been steadily growing and it had been observed worldwide. In the past few years, call centers have been introduced with great success by many service-oriented companies such as banks and insurance companies. They become the main point of contact with the customer, and an integral part of the majority of corporations. The large-scale emergence of call centers has created a fertile source of management issues. In this thesis, we focused on various operations management issues of call centers. Our analyzes leaded to both qualitative and quantitative results for practical management. We used approaches that are based on stochastic models and in particular queueing models.

We investigated the impact of team-based organizations in call centers management. Agents of call centers are the interface between the company and the customers. Thus, managers have to support and motivate their employees, so that, the assistance they provide to customers is efficient. Partitioning agents into groups creates competition and makes agents more responsible, which motivates them to provide both rapid and improved responses. We developed queueing models that show that the benefits of the team based organization in providing more efficient answers to customers very often outweigh its drawback coming from the loss of pooling effects.

In addition, we focused on real-time issues of call centers. In the third chapter, We considered a two-class call center and developed real-time scheduling policies that determine the rule of assignment of new arrivals to the waiting lines. We focused on service levels criteria related to the fraction of abandoning customers and the variance of queueing delays. In the fourth chapter, we proposed a call center model in which we provide information about delays to customers, and we quantified its effect upon performance.

Next, we tackled the transient analysis of general birth-death processes. We computed several closed-form expressions of the moments of first passage times, and pointed out their applications for the quantitative analysis of some queueing systems, such as call centers. Finally, we derived monotonicity results for Markovian queueing systems with impatient customers.

7.2 Future Research

Worrying about accurate and practical results, much is left to be done. As detailed in the concluding sections of the previous chapters, several interesting areas of future research arise. In what follows, we point out some of these research directions.

One may continue our work by incorporating customers reneging in the team-based organiza-

tion analysis. A more ambitious extension would be to investigate the introduction of team-based organization in an SBR call center where agents have specific skills.

An interesting direction, for the real-time policies we developed, lies in investigating accurate analyzes in order to better understand the behavior of the variance of queueing delays with respect to these policies. In practice, it should be also of value to extend the research in case of different statistical behaviors for different customer types. In other words, service times as well as times before reneging are not equal for both customer types.

One important extension from a practical point of view is to describe empirically customers reaction in response to delays announcement. This would validate in a real call center case our claim regarding to the advantages of announcing delays. A further study to quantify the relation between costs of reneging and balking would be of great value.

We want to develop simple approximations or numerical methods for computing the moments of first passage times in birth-death processes. This would be helpful to avoid computation difficulties given that the closed-form expressions of interest are somewhat cumbersome. Further useful applications should be also pointed out.

As a topic for future research, one would investigate more convexity results by considering more complex systems with general distributions for service times and times before reneging. In our opinion, further results with regard to other design variables would be important for the call center industry.

Appendix A

Appendix of Chapter 2

This appendix deals with the analysis of Chapter 2. In Appendix A.1, we performed a more systematic analysis than that reported in Section 2.4 in order to confirm the robustness of our conclusions. Using simulation experiments, we validate in Appendix A.2 the approximations for the Portfolio Dedicated System already developed in Section 2.5.1. Finally, Appendix A.3 is devoted to the proof of a result used in Section 2.5.2.

A.1 Extension of the Quantitative Analysis

The numerical study in Section 2.4 was based on a set of basic data for the initial Pooled System: $\mu = 0.2$, $\alpha = 10\%$, $W(20sec) = 80\%$, and $s = 1000$. These basic data are representative of typical parameters encountered in the *Bouygues Telecom* call center. However, to make sure that the conclusions drawn from this set of data are robust, we have performed a large set of experiments, some of which are reported in this appendix. The study is divided into four steps. In each step, we first vary one parameter (μ, α, $W(20sec)$ or s), then we deduce λ^a and λ to get different initial pooled systems which cover many realistic call center cases. Next, we consider each case and we compute the required increase in the service rate or decrease in the call back proportion, in order to reach the same performance as in the fully pooled system, for different numbers of separated teams in the corresponding dedicated systems.

Varying the Service Rate μ

We consider four pooled systems: $s = 1000$, $\alpha = 10\%$, $W(20sec) = 80\%$, and $\mu = 0.1$, 0.2, 0.5, and 1, respectively. Then, $\lambda^a = 88.14$, 177.35, 446.52, and 896.12, and $\lambda = 97.93$, 197.06, 496.13, and 995.69, respectively. In Figure A.1, we plot the curves of the required service rate increase versus the number of pools in the dedicated systems. In Figure A.2, we plot the curves of the

required call back proportion decrease versus the number of pools. We vary n only from 1 to 10, so that, α_n stays positive.

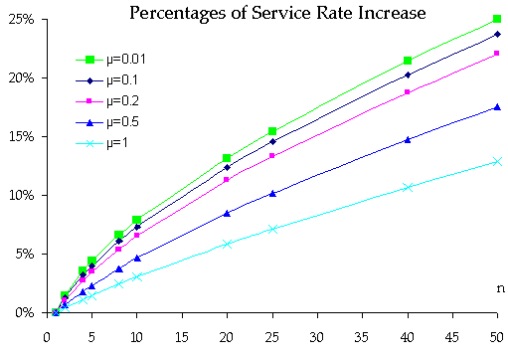

Figure A.1: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial service rates

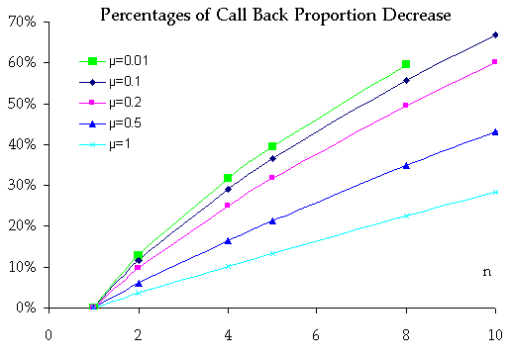

Figure A.2: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial service rates

For every value of μ, the results are pretty much of the same quality as in Section 2.4. Furthermore, an additional insight is that the costs of migrating to the team-based organization ($(\mu_n - \mu)/\mu$ or $(\alpha - \alpha_n)/\alpha$) are decreasing as the initial service rate is increasing. One intuitive explanation is as follows. Consider two pooled systems. The parameters of the first are s, $\lambda^{(1)}$, $\mu^{(1)}$, and $W(t)$. The parameters of the second are s, $\lambda^{(2)}$, $\mu^{(2)}$, and $W(t)$. We assume that $\mu^{(1)} < \mu^{(2)}$, then $\lambda^{(1)}$ must be less than $\lambda^{(2)}$ in order to match the same performance $W(t)$ in the two systems. Besides, since the servers are slower in the first system, the server utilization of the last is less than the one in the second system, else $W(t)$ will be higher in the second system. Hence, the second system has more pooling effect than the first one. Now, let us divide each system to n identical unpooled systems, so that, s is a multiple of n. The parameters of one of the first unpooled models are s/n, $\lambda^{(1)}/n$, and the service rate is $\mu_n^{(1)}$ such that $W_n(t) = W(t)$. The parameters of one of the second unpooled models are s/n, $\lambda^{(2)}/n$, and $\mu_n^{(2)}$ such that $W_n(t) = W(t)$. Thanks to the pooling effect that is more present in the second pooled system than in the first one, the second unpooled system will need an increase in the service rate regarding μ^2 being less than the one regarding $\mu^{(1)}$ in the first unpooled system, $(\mu_n^{(2)} - \mu^{(2)})/\mu^2 < (\mu_n^{(1)} - \mu^{(1)})/\mu^{(1)}$. An additional insight is that it appears that when μ decreases, the set of curves (for different values of μ) converges towards an asymptotic curve. Indeed, we have checked that the curves for $\mu = 0.001$ almost coincide with those for $\mu = 0.01$ in Figures A.1 and A.2.

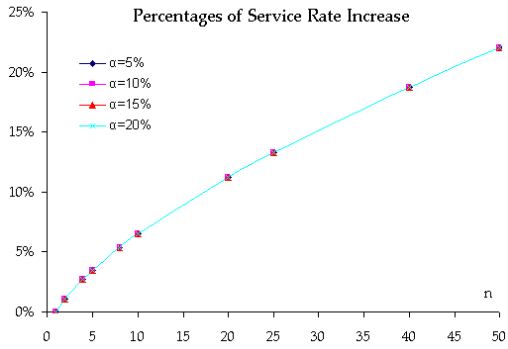

Figure A.3: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial call back proportions

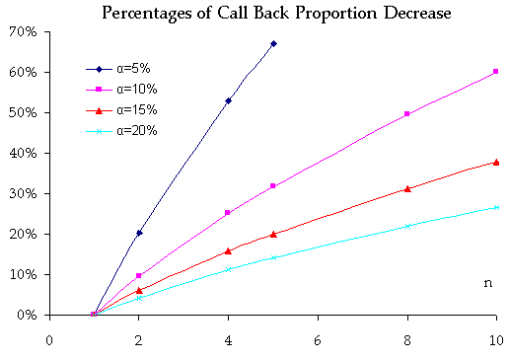

Figure A.4: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial call back proportions

Varying the Call Back Proportion α

Here, we vary the call back proportion with regard to the Pooled System of Section 2.4. We consider four pooled systems: $s = 1000$, $\mu = 0.2$, $W(20sec) = 80\%$, and $\alpha = 5\%$, 10%, 15%, and 20%, respectively. Then, $\lambda^a = 187.21$, 177.36, 167.5, and 157.65, respectively, and $\lambda = 197.06$, for the four systems. Figures A.3 and A.4 show, respectively, the required quantitative (service times) and qualitative (rate of calls successfully addressed) improvements according to the number of pools.

Again, we have the same qualitative results as in Section 2.4. In Figure A.3, the curves are identical. The explanation is as follows. Let us consider two pooled systems with the same parameters except for the dissatisfaction probability α. The required total arrival rates, to meet a given QoS, are identical in the two systems because they do not depend on α. Therefore, the two systems are equivalent to the same Erlang-C model. The required service rate μ_n and increase in the service rate $(\mu_n - \mu)/\mu$, for the corresponding dedicated systems, do not change for a fixed number of pools n.

Figure A.4 shows that the required improvement in the dissatisfaction probability $(\alpha - \alpha_n)/\alpha$ is decreasing with the initial call back proportion α. The proof of this result is as follows. Consider two pooled systems with the same parameters s, λ, μ, and $W(t)$. The arrival rate of first-attempt calls and the call back proportion for the first system are $\lambda^{a,1}$ and $\alpha^{(1)}$, respectively. The ones for the second system are $\lambda^{a,2}$ and $\alpha^{(2)}$, respectively. We assume that $\alpha^{(1)} < \alpha^{(2)}$. Now, let us divide each pooled system to n identical unpooled systems, while leaving unchanged the total number of servers s, the service rate μ, and the quality of service $W_n(t) = W(t)$. So, the total arrival rate, the number of servers, and the service rate for each type of unpooled system are

λ/n, s/n, and μ, respectively. The arrival rate of first-attempt calls and the call back proportion for the first unpooled systems are $\lambda_n^{a,1} = \lambda^{a,1}/n$ and $\alpha_n^{(1)}$, respectively. The ones for the second unpooled systems are $\lambda_n^{a,2} = \lambda^{a,2}/n$ and $\alpha_n^{(2)}$, respectively. Clearly, we have $\alpha_n^{(1)} < \alpha^{(1)}$ and $\alpha_n^{(2)} < \alpha^{(2)}$ because of the loss of the pooling effect. From the pooled systems, we deduce that $\lambda = \lambda^{a,1}/(1-\alpha^{(1)}) = \lambda^{a,2}/(1-\alpha^{(2)})$, and from the unpooled systems, we deduce that $\lambda/n = \lambda_n^{a,1}/(1-\alpha_n^{(1)}) = \lambda_n^{a,2}/(1-\alpha_n^{(2)})$. The two last relations give Equation (A.1) below.

$$\frac{1-\alpha_n^{(1)}}{1-\alpha^{(1)}} = \frac{1-\alpha_n^{(2)}}{1-\alpha^{(2)}}. \tag{A.1}$$

Since $\alpha^{(1)} < \alpha^{(2)}$, then $\lambda^{a,1} > \lambda^{a,2}$, and equivalently $\lambda_n^{a,1} > \lambda_n^{a,2}$, so $\alpha_n^{(1)} < \alpha_n^{(2)}$. Moreover, $\alpha_n^{(1)} < \alpha^{(1)}$, we deduce then from Equation (A.1) that $\alpha_n^{(1)}/\alpha^{(1)} < \alpha_n^{(2)}/\alpha^{(2)}$. Hence, $1-(\alpha_n^{(1)}/\alpha^{(1)}) > 1-(\alpha_n^{(2)}/\alpha^{(2)})$, and finally $(\alpha^{(1)}-\alpha_n^{(1)})/\alpha^{(1)} > (\alpha^{(2)}-\alpha_n^{(2)})/\alpha^{(2)}$.

Varying the Quality of Service $W(20sec)$

Now, we vary the quality of service $W(20sec)$ with regard to the Pooled System of Section 2.4. We consider five pooled systems: $s = 1000$, $\mu = 0.2$, $\alpha = 10\%$, and $W(20sec) = 60\%$, 80%, 90%, 95%, and 99%, respectively. Then, $\lambda^a = 178.47$, 177.36, 176.27, 175.22, and 172.90, and $\lambda = 198.30$, 197.06, 195.86, 194.69, and 192.11, respectively. Figures A.5 and A.6 show, respectively, the required quantitative and qualitative improvements according to the number of pools.

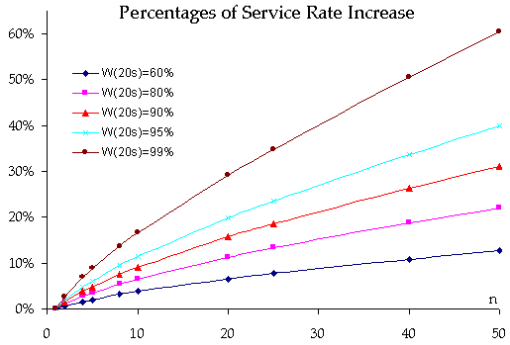

Figure A.5: Percentages of service rate increase according to number of pools n in a Dedicated System in order to achieve the same $W_n(20sec)$ as in the Pooled System, for a different values of $W_n(20sec)$

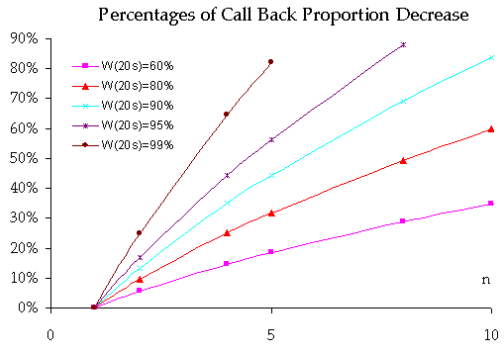

Figure A.6: Percentages of call back proportion decrease according to number of pools n in a Dedicated System in order to achieve the same $W_n(20sec)$ as in the Pooled System, for a different values of $W_n(20sec)$

Once again, we underline the qualitative similarity of the results as in Section 2.4. The additional insight here is that the costs of partitioning the big call center $((\mu_n-\mu)/\mu$ or $(\alpha-\alpha_n)/\alpha)$ are increasing as the chosen quality of service is increasing. For instance, let us partition

two pooled systems into n identical unpooled systems. The two pooled systems have the same number of servers and the same service rate. However, the first pooled system has a quality of service lower than the one in the second. Both of the unpooled systems will need an increase in the service rate, because of the loss of the pooling effect. Moreover, since we have to reach a higher QoS in the second unpooled systems, then we will need for them a higher increase in the service rate. We notice again from the curves that the costs of migrating do not roughly increase with the chosen quality of service.

Varying the Number of Servers s

Up to now, all our analyzes were performed as a function of n, the number of dedicated pools. As stated in Chapter 2, as long as the total number of servers s is fixed, the results obtained can alternatively be reinterpreted in terms of s/n. Indeed, specifying n is equivalent to specifying s/n. In this section however, we want to perform our analyzes for different values of s. In that case, it seems more consistent to compare configurations having the same number of servers in each pool. Therefore, the analyzes will be performed as a function of the number of dedicated servers in each pool, s/n, for different values of the total number of servers, s.

Consider now five pooled systems: $\mu = 0.2$, $\alpha = 10\%$, $W(20sec) = 80\%$, and $s = 100$, 200, 500, 1000, and 5000, respectively. Then, $\lambda^a = 16.63$, 34.26, 87.74, 177.36, and 896.60, and $\lambda = 18.48$, 38.07, 97.49, 197.06, and 996.22, respectively. In Figure A.7, we plot the curves of the required service rate improvement, when we partition the pooled systems chosen here, according to the size of the generated teams s/n. We notice from Figure A.7 that the costs, for a fixed size of pools, are increasing with the initial number of servers. One explanation may be as follows. Consider once again two pooled systems. The parameters of the first are $s^{(1)}$, $\lambda^{(1)}$, μ, and $W(t)$. Those of the second are $s^{(2)}$, $\lambda^{(2)}$, μ, and $W(t)$. We assume that $s^{(1)} < s^{(2)}$. Let us now migrate to the corresponding unpooled systems such that the size of each type of unpooled system is $s^{(p)}$, $s^{(1)} = n_1 s^{(p)}$ and $s^{(2)} = n_2 s^{(p)}$. It goes without saying that $n_1 < n_2$. The parameters of the first unpooled systems are $s^{(p)}$, $\lambda^{(1)}/n_1$, and $\mu_n^{(1)}$ such that the quality of service is $W_n(t) = W(t)$. The ones of the second unpooled systems are $s^{(p)}$, $\lambda^{(2)}/n_2$, and $\mu_n^{(2)}$ such that the quality of service is $W_n(t) = W(t)$ too. Due to the pooling effect that is more present in the second pooled system than in the first, $\lambda^{(2)}/n_2$ is larger than $\lambda^{(1)}/n_1$. Else, the first pooled system will match a quality of service that is lower than the second. To do a summary for the unpooled systems, we have the same number of servers $s^{(p)}$, the same quality of service $W(t)$, and a larger arrival rate for the second unpooled systems. Thus, we easily deduce that the servers in the latter cases must be faster so as to match the same performance in both types of dedicated systems, $\mu_n^{(1)} < \mu_n^{(2)}$.

Finally, $(\mu_n^{(1)} - \mu)/\mu < (\mu_n^{(2)} - \mu)/\mu$.

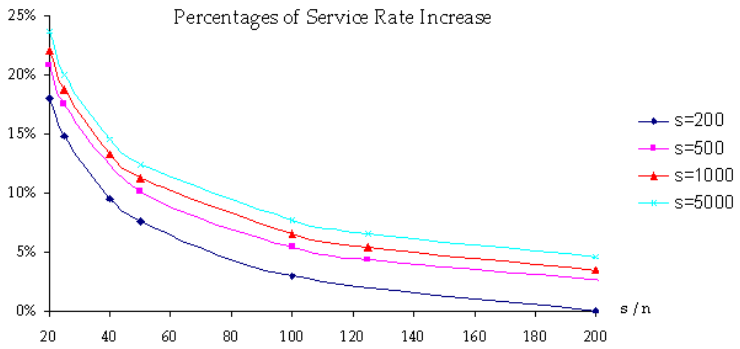

Figure A.7: Percentages of service rate increase according to size of pools s/n in a Dedicated System in order to achieve $W_n(20sec) = 80\%$, for a different initial number of servers

In addition, we see from Figure A.7 that the gap between the curves is decreasing when s increases. Then, we can deduce that the unpooling of two pooled systems with different large number of servers, namely greater than 500, will need quite the same increase in the service rates. This is due to the fact that a "large" Pooled System does not gain too much in pooling effect by adding more servers.

A.2 Validation of the Approximation Models

The analysis of the Portfolio Dedicated System is to be used to design our call center; calculating staffing level, required total arrival rate (or required call back proportion), or required service rate in order to achieve a given QoS. To examine the accuracy of the approximation models, we propose two different formulations:

- QoS, $s \Rightarrow \lambda$: formulation 1 consists of calculating the required total arrival rate λ given a fixed QoS and a fixed staffing level s.

- QoS, $\lambda \Rightarrow s$: formulation 2 consists of calculating the required staffing level s given a fixed QoS and a fixed total arrival rate λ.

We compare performances given from pessimistic models with those from simulation. We simulated 30 cases: the number of pools is $n = 10$, the OPTF customers proportion is $p = 5\%$ or 10%, and for each p, n $s_n = 250$, 350 or 500, and for each p and s we chose 5 values of λ^a (in order to vary the server utilization). The mean service time, and the call back proportion are kept constant ($1/\mu_n = 5$ min, and $\alpha_n = 10\%$).

Deviations between performance measures given by pessimistic models and those given by simulation are presented in Table A.1. For each pessimistic model (PTF or OPTF), deviations for one parameter are calculated as $\frac{performance(model)-performance(simulation)}{performance(simulation)}$.

	Total Arrival Rate, $\lambda^a/(1-\alpha_n)$		Total Staffing Level, $n\ s_n$	
	$W_n(20sec)$	W_n	$W_n(20sec)$	W_n
PTF Pessimistic Model	-2.84%	-2.92%	4.00%	4.00%
OPTF Pessimistic Model	-5.65%	-5.61%	4.43%	4.43%

Table A.1: Deviations between pessimistic models and simulation

A.3 Proof of the Result: W^{global} does not depend on p

First, consider an $M/M/s$ queue with a single class of customers. The arrival rate is λ, the number of servers s, and the service rate is μ. Hence, the stationary mean waiting time in queue is given by

$$W = \frac{P_D}{s\mu - \lambda}, \tag{A.2}$$

where P_D is the probability of delay, that is, the probability that an incoming customer waits for service. Second, consider a non-preemptive priority $M/M/s$ queue with two types of customers, say A and B. Type A customers have priority over type B ones. The total arrival rate is λ, the number of servers is s, and the service rate to handle any type of customers is μ, as in the first $M/M/s$ model. The arrival rate of type A customers is λ^A, and that of type B customers is λ^B, $\lambda^A + \lambda^B = \lambda$. Let p be the proportion of type B customers. Then, $\lambda^A = (1-p)\lambda$. As in Kella and Yechiali [78], the average waiting times in queue of customers A (W^A) and customers B (W^B) are respectively given as follows

$$W^A = \frac{P_D}{s\mu - \lambda^A}, \text{ and } W^B = \frac{P_D}{(s\mu - \lambda^A)(1 - \frac{\lambda}{s\mu})}, \tag{A.3}$$

where the probability of delay P_D is identical to that in the first model. For any proportion $p \in [0,1]$, we have

$$\begin{aligned} W^{global} &= (1-p)\, W^A + p\, W^B \\ &= (1-p)\, \frac{P_D}{s\mu - \lambda^A} + p\, \frac{P_D}{(s\mu - \lambda^A)(1-\frac{\lambda}{s\mu})} \\ &= \frac{P_D}{s\mu - (1-p)\lambda} \times \frac{s\mu - (1-p)\lambda}{s\mu - \lambda} \\ &= \frac{P_D}{s\mu - \lambda} \\ &= W, \end{aligned} \tag{A.4}$$

which completes the proof. □

Appendix B

Appendix of Chapter 3

In this appendix, we present supporting simulation experiments for the analysis of Section 3.5 of Chapter 3. We consider the systems already chosen in Section 3.5, and we simulate them working under scheduling policies π_A, π_1, π_2 and π_3 . In Tables B.1 and B.2, we show the results for the target ratios $c^* = 0.5$ and $c^* = 0.9$, respectively.

		π_A	π_1	π_2	π_3
System 1: $\lambda_A = \lambda_B = 4.5$	c	0.305	0.500	0.500	0.500
	Q^A	1.211%	2.291%	1.727%	2.559%
	Q^B	3.966%	4.582%	3.453%	5.118%
	Q	2.589%	3.436%	2.590%	3.838%
	W^A	0.035	0.062	0.049	0.072
	W^B	0.102	0.119	0.090	0.135
	W	0.068	0.090	0.069	0.103
	σ^A	0.098	0.192	0.141	0.184
	σ^B	0.311	0.334	0.280	0.343
	σ	0.232	0.273	0.222	0.276
System 2: $\lambda_A = \lambda_B = 4.75$	c	0.282	0.500	0.500	0.500
	Q^A	1.838%	2.917%	2.786%	2.919%
	Q^B	6.526%	5.834%	5.573%	5.839%
	Q	4.182%	4.376%	4.179%	4.379%
	W^A	0.053	0.079	0.079	0.082
	W^B	0.170	0.152	0.145	0.155
	W	0.110	0.115	0.111	0.118
	σ^A	0.119	0.221	0.182	0.194
	σ^B	0.407	0.379	0.363	0.361
	σ	0.303	0.311	0.288	0.291

		π_A	π_1	π_2	π_3
System 3: $\lambda_A = \lambda_B = 4.9$	c	0.258	0.500	0.500	0.500
	Q^A	2.737%	4.036%	4.794%	3.757%
	Q^B	10.605%	8.073%	9.587%	7.514%
	Q	6.671%	6.054%	7.190%	5.636%
	W^A	0.080	0.109	0.137	0.106
	W^B	0.279	0.212	0.250	0.202
	W	0.175	0.159	0.192	0.153
	σ^A	0.144	0.268	0.245	0.223
	σ^B	0.537	0.454	0.494	0.411
	σ	0.399	0.375	0.391	0.332
System 4: $\lambda_A = \lambda_B = 4.95$	c	0.258	0.500	0.500	0.500
	Q^A	2.692%	4.303%	5.542%	4.819%
	Q^B	10.427%	8.606%	11.083%	9.639%
	Q	6.556%	6.455%	8.313%	7.229%
	W^A	0.078	0.117	0.158	0.136
	W^B	0.275	0.226	0.288	0.261
	W	0.173	0.170	0.221	0.197
	σ^A	0.142	0.279	0.267	0.257
	σ^B	0.528	0.471	0.540	0.472
	σ	0.392	0.389	0.427	0.383
System 5: $\lambda_A = \lambda_B = 5$	c	0.253	0.500	0.500	0.500
	Q^A	2.889%	4.349%	5.010%	4.270%
	Q^B	11.414%	8.699%	10.020%	8.541%
	Q	7.152%	6.524%	7.515%	6.406%
	W^A	0.084	0.118	0.143	0.121
	W^B	0.303	0.229	0.263	0.231
	W	0.188	0.172	0.201	0.174
	σ^A	0.147	0.280	0.248	0.238
	σ^B	0.556	0.472	0.504	0.439
	σ	0.413	0.390	0.398	0.355
System 6: $\lambda_A = \lambda_B = 6$	c	0.180	0.500	0.500	0.500
	Q^A	6.655%	12.281%	11.750%	11.973%
	Q^B	36.871%	24.562%	23.501%	23.946%
	Q	21.763%	18.421%	17.626%	17.960%
	W^A	0.199	0.327	0.350	0.348
	W^B	1.149	0.693	0.639	0.716
	W	0.582	0.496	0.484	0.518
	σ^A	0.210	0.557	0.376	0.431
	σ^B	1.085	0.872	0.816	0.739
	σ	0.848	0.742	0.637	0.621

Table B.1: Simulation experiments for $c^* = 0.5$

		π_A	π_1	π_2	π_3
System 1: $\lambda_A = \lambda_B = 4.5$	c	0.296	0.900	0.900	0.900
	Q^A	1.553%	2.876%	2.895%	2.555%
	Q^B	5.247%	3.195%	3.217%	2.839%
	Q	3.400%	3.036%	3.056%	2.697%
	W^A	0.045	0.076	0.081	0.071
	W^B	0.136	0.084	0.088	0.078
	W	0.089	0.080	0.084	0.075
	σ^A	0.112	0.246	0.205	0.192
	σ^B	0.365	0.266	0.232	0.214
	σ	0.271	0.256	0.219	0.203
System 2: $\lambda_A = \lambda_B = 4.75$	c	0.270	0.900	0.900	0.900
	Q^A	2.331%	4.643%	4.521%	4.889%
	Q^B	8.643%	5.159%	5.023%	5.433%
	Q	5.487%	4.901%	4.772%	5.161%
	W^A	0.068	0.122	0.127	0.137
	W^B	0.225	0.135	0.138	0.151
	W	0.144	0.129	0.133	0.144
	σ^A	0.135	0.319	0.256	0.270
	σ^B	0.482	0.345	0.301	0.300
	σ	0.358	0.332	0.278	0.286
System 3: $\lambda_A = \lambda_B = 4.9$	c	0.253	0.900	0.900	0.900
	Q^A	2.997%	5.931%	5.141%	7.381%
	Q^B	11.846%	6.590%	5.713%	8.201%
	Q	7.422%	6.261%	5.427%	7.791%
	W^A	0.087	0.156	0.145	0.210
	W^B	0.312	0.173	0.158	0.232
	W	0.194	0.165	0.151	0.221
	σ^A	0.152	0.368	0.272	0.333
	σ^B	0.575	0.397	0.311	0.369
	σ	0.427	0.383	0.292	0.352
System 4: $\lambda_A = \lambda_B = 4.95$	c	0.261	0.900	0.900	0.900
	Q^A	2.551%	6.008%	7.779%	7.189%
	Q^B	9.774%	6.675%	8.644%	7.987%
	Q	6.163%	6.312%	8.211%	7.588%
	W^A	0.074	0.159	0.222	0.204
	W^B	0.258	0.175	0.241	0.225
	W	0.162	0.167	0.232	0.215
	σ^A	0.138	0.370	0.338	0.330
	σ^B	0.508	0.399	0.388	0.365
	σ	0.378	0.385	0.364	0.348
System 5: $\lambda_A = \lambda_B = 5$	c	0.257	0.900	0.900	0.900
	Q^A	2.702%	7.363%	7.291%	7.533%
	Q^B	10.529%	8.181%	8.101%	8.370%
	Q	6.616%	7.772%	7.696%	7.952%
	W^A	0.079	0.194	0.208	0.215
	W^B	0.279	0.216	0.226	0.237
	W	0.175	0.205	0.217	0.226
	σ^A	0.141	0.417	0.325	0.337
	σ^B	0.529	0.450	0.374	0.374
	σ	0.393	0.434	0.350	0.356

		π_A	π_1	π_2	π_3
System 6: $\lambda_A = \lambda_B = 6$	c	0.188	0.900	0.900	0.900
	Q^A	5.686%	16.697%	18.299%	19.256%
	Q^B	30.241%	18.552%	20.332%	21.396%
	Q	17.964%	17.624%	19.315%	20.326%
	W^A	0.170	0.448	0.566	0.597
	W^B	0.910	0.503	0.609	0.665
	W	0.484	0.475	0.587	0.630
	σ^A	0.189	0.694	0.480	0.512
	σ^B	0.941	0.741	0.563	0.563
	σ	0.729	0.718	0.523	0.539

Table B.2: Simulation experiments for $c^* = 0.9$

Appendix C

Appendix of Chapter 6

This appendix deals with the analysis of Chapter 6. In Appendix C.1, we give the proof of the result in Property 6.1. In Appendix C.2, we present some numerical experiments in order to illustrate the concavity results already derived in Chapter 6.

C.1 Proof of Property 6.1

Using the notations in Equation (6.7), the stationary probabilities for an $M/M/s/s+k \ \ +M$ queue (k extra waiting lines) given in Equations (6.3) and (6.4) may be rewritten as

$$p_k(i) = \phi_i \times p(0), \quad \text{for } 0 \leq i \leq s, \tag{C.1}$$

$$p_k(i) = \rho_i \times p(0), \quad \text{for } s < i \leq s+k, \tag{C.2}$$

and

$$p_k(0) = \frac{s!\,\mu^s}{s!\,\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{s+k} \rho_i}. \tag{C.3}$$

Substituting them into Equation (6.5) yields, for $k \geq 0$, to

$$\begin{aligned} U_k = &\left(\frac{\rho_{k+s}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i} - \frac{\rho_{k+s+1}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i} \right) \\ &+ \frac{\gamma}{\lambda} \left(\frac{\sum_{i=s+1}^{k+s} (i-s)\rho_i}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i} - \frac{\sum_{i=s+1}^{k+s+1} (i-s)\rho_i}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i} \right). \end{aligned} \tag{C.4}$$

Or equivalently with some algebra

$$
\begin{aligned}
U_k(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i) &= \rho_{k+s}(1 + \frac{\rho_{k+s+1}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i}) - \rho_{k+s+1} \\
&+ \frac{\gamma}{\lambda}\left(\sum_{i=s+1}^{k+s} (i-s)\rho_i(1 + \frac{\rho_{k+s+1}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i}) - \sum_{i=s+1}^{k+s+1} (i-s)\rho_i \right) \\
&= \rho_{k+s} + \frac{\rho_{k+s}\quad \rho_{k+s+1}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i} - \rho_{k+s+1} \\
&+ \frac{\gamma}{\lambda}\left(-(k+1)\rho_{k+s+1} + \frac{\rho_{k+s+1}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i} \sum_{i=s+1}^{k+s} (i-s)\rho_i \right) \\
&= \rho_{k+s} + \rho_{k+s+1}(-1 + \frac{\rho_{k+s}}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i} - \frac{(k+1)\gamma}{\lambda} + \frac{\gamma}{\lambda} \frac{\sum_{i=s+1}^{k+s}(i-s)\rho_i}{s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i}).
\end{aligned}
\tag{C.5}
$$

Calculating further gives

$$
\begin{aligned}
U_k(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i)(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) &= \rho_{k+s}(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) \\
&+ \rho_{k+s+1}\left(\rho_{k+s} - (s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) - \frac{(k+1)\gamma}{\lambda}(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) + \frac{\gamma}{\lambda} \sum_{i=s+1}^{k+s} (i-s)\rho_i \right) \\
&= \rho_{k+s}(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) + \rho_{k+s+1}\left(\rho_{k+s} - (1 + \frac{(k+1)\gamma}{\lambda})(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) + \frac{\gamma}{\lambda} \sum_{i=s+1}^{k+s} (i-s)\rho_i \right).
\end{aligned}
\tag{C.6}
$$

Observing that $\rho_{k+s+1} = \frac{\lambda}{s\mu+(k+1)\gamma}\, \rho_{k+s}$, for $k \geq 0$, we may write

$$
\begin{aligned}
&U_k(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i)(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i)\frac{1}{\rho_{k+s}} \\
&= (s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i + \rho_{k+s+1}) - \frac{\lambda + (k+1)\gamma}{s\mu + (k+1)\gamma}(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) + \frac{\gamma}{s\mu + (k+1)\gamma} \sum_{i=s+1}^{k+s} (i-s)\rho_i \\
&= (s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i) - (s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) \\
&\qquad - \frac{\lambda - s\mu}{s\mu + (k+1)\gamma}(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) + \frac{\gamma}{s\mu + (k+1)\gamma} \sum_{i=s+1}^{k+s} (i-s)\rho_i.
\end{aligned}
\tag{C.7}
$$

Simplifying Equation (C.7) implies the following relation

$$
\begin{aligned}
&U_k(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s+1} \rho_i)(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i)\frac{1}{\rho_{k+s}} \\
&\qquad = \rho_{k+s+1} - \frac{\lambda - s\mu}{s\mu + (k+1)\gamma}(s!\mu^s \sum_{i=0}^{s} \phi_i + \sum_{i=s+1}^{k+s} \rho_i) + \frac{\gamma}{s\mu + (k+1)\gamma} \sum_{i=s+1}^{k+s} (i-s)\rho_i.
\end{aligned}
\tag{C.8}
$$

For the rank $k+1$, Relation (C.8) becomes

$$\begin{aligned} U_{k+1}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+2}\rho_i)(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+1}\rho_i)\frac{1}{\rho_{k+s+1}} \qquad (C.9)\\ =\rho_{k+s+2}-\frac{\lambda-s\mu}{s\mu+(k+2)\gamma}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+1}\rho_i)+\frac{\gamma}{s\mu+(k+2)\gamma}\sum_{i=s+1}^{k+s+1}(i-s)\rho_i\\ =\frac{\lambda}{s\mu+(k+2)\gamma}\rho_{k+s+1}-\frac{\lambda-s\mu}{s\mu+(k+2)\gamma}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s}\rho_i+\rho_{k+s+1})\\ +\frac{\gamma}{s\mu+(k+2)\gamma}\sum_{i=s+1}^{k+s}(i-s)\rho_i+\frac{(k+1)\gamma}{s\mu+(k+2)\gamma}\rho_{k+s+1}. \end{aligned}$$

Hence,

$$\begin{aligned} U_{k+1}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+2}\rho_i)(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+1}\rho_i)\frac{1}{\rho_{k+s+1}} \qquad (C.10)\\ =\frac{s\mu+(k+1)\gamma}{s\mu+(k+2)\gamma}\rho_{k+s+1}-\frac{\lambda-s\mu}{s\mu+(k+2)\gamma}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s}\rho_i)+\frac{\gamma}{s\mu+(k+2)\gamma}\sum_{i=s+1}^{k+s}(i-s)\rho_i. \end{aligned}$$

Multiplying both sides in Equation (C.10) by $\frac{s\mu+(k+2)\gamma}{s\mu+(k+1)\gamma}$ implies

$$\begin{aligned} U_{k+1}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+2}\rho_i)(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+1}\rho_i)\frac{1}{\rho_{k+s+1}}\cdot\frac{s\mu+(k+2)\gamma}{s\mu+(k+1)\gamma} \qquad (C.11)\\ =\rho_{k+s+1}-\frac{\lambda-s\mu}{s\mu+(k+1)\gamma}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s}\rho_i)+\frac{\gamma}{s\mu+(k+1)\gamma}\sum_{i=s+1}^{k+s}(i-s)\rho_i. \end{aligned}$$

From Equations (C.8) and (C.11), we next deduce that

$$\begin{aligned} U_{k+1}(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+2}\rho_i)(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+1}\rho_i)\frac{1}{\rho_{k+s+1}}\cdot\frac{s\mu+(k+2)\gamma}{s\mu+(k+1)\gamma} \qquad (C.12)\\ =U_k(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+1}\rho_i)(s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s}\rho_i)\frac{1}{\rho_{k+s}}. \end{aligned}$$

Finally, simplifying Equation (C.12) and again observing that $\frac{\rho_{k+s+1}}{\rho_{k+s}}=\frac{\lambda}{s\mu+(k+1)\gamma}$, we get for all $k\geq 0$,

$$U_{k+1}=\frac{s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s}\rho_i}{s!\mu^s\sum_{i=0}^{s}\phi_i+\sum_{i=s+1}^{k+s+2}\rho_i}\cdot\frac{\lambda}{s\mu+(k+2)\gamma}\cdot U_k, \qquad (C.13)$$

which completes the proof of the property. □

C.2 Numerical illustrations

In this appendix, we present numerical examples to illustrate the convexity results. We compute the probability of being served as a function of the queue capacity for several systems chosen so as to cover a broad range of parameters values. Systems parameters are presented in Section 6.5.

	$\lambda = 1.8$			$\lambda = 8$			$\lambda = 60$		
k	$s = 1$	$s = 2$	$s = 3$	$s = 5$	$s = 10$	$s = 15$	$s = 50$	$s = 70$	$s = 100$
0	35.7143	63.3484	81.9733	52.0992	87.8339	99.0899	78.3881	97.6256	99.9999
1	44.3548	73.3209	89.1589	56.3701	90.9867	99.5033	79.3985	98.0027	100.0000
2	47.5087	77.5281	92.0093	58.5389	92.9632	99.7085	80.1610	98.3100	100.0000
3	48.8651	79.3467	93.0916	59.7405	94.2080	99.8076	80.7453	98.5601	100.0000
4	49.4796	80.1024	93.4736	60.4509	94.9854	99.8542	81.1989	98.7634	100.0000
5	49.7536	80.3947	93.5977	60.8924	95.4627	99.8755	81.5551	98.9283	100.0000
6	49.8691	80.4986	93.6349	61.1773	95.7490	99.8850	81.8377	99.0615	100.0000
7	49.9140	80.5324	93.6452	61.3664	95.9160	99.8890	82.0639	99.1688	100.0000
8	49.9299	80.5425	93.6478	61.4942	96.0106	99.8908	82.2466	99.2548	100.0000
9	49.9352	80.5454	93.6485	61.5815	96.0625	99.8915	82.3952	99.3235	100.0000
10	49.9367	80.5461	93.6486	61.6411	96.0901	99.8918	82.5169	99.3780	100.0000
11	49.9372	80.5463	93.6487	61.6815	96.1044	99.8919	82.6174	99.4211	100.0000
12	49.9373	80.5463	93.6487	61.7086	96.1115	99.8919	82.7008	99.4550	100.0000
13	49.9373	80.5463	93.6487	61.7264	96.1149	99.8919	82.7704	99.4815	100.0000
14	49.9373	80.5463	93.6487	61.7379	96.1165	99.8919	82.8289	99.5021	100.0000
15	49.9373	80.5463	93.6487	61.7450	96.1173	99.8919	82.8782	99.5180	100.0000
16	49.9373	80.5463	93.6487	61.7494	96.1176	99.8919	82.9200	99.5302	100.0000
17	49.9373	80.5463	93.6487	61.7519	96.1178	99.8919	82.9556	99.5395	100.0000
18	49.9373	80.5463	93.6487	61.7534	96.1178	99.8919	82.9860	99.5466	100.0000
19	49.9373	80.5463	93.6487	61.7542	96.1178	99.8919	83.0121	99.5519	100.0000
20	49.9373	80.5463	93.6487	61.7546	96.1179	99.8919	83.0345	99.5559	100.0000
21	49.9373	80.5463	93.6487	61.7548	96.1179	99.8919	83.0539	99.5589	100.0000
22	49.9373	80.5463	93.6487	61.7549	96.1179	99.8919	83.0706	99.5611	100.0000
23	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.0851	99.5627	100.0000
24	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.0977	99.5639	100.0000
25	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.1087	99.5648	100.0000
26	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.1182	99.5654	100.0000
27	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.1265	99.5658	100.0000
28	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.1337	99.5662	100.0000
29	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.1399	99.5664	100.0000
30	49.9373	80.5463	93.6487	61.7550	96.1179	99.8919	83.1454	99.5665	100.0000

Table C.1: Values of Q_k (in %) for $\gamma = 0.5$

	$\lambda = 1.8$			$\lambda = 8$			$\lambda = 60$		
k	$s = 1$	$s = 2$	$s = 3$	$s = 5$	$s = 10$	$s = 15$	$s = 50$	$s = 70$	$s = 100$
0	35.7143	63.3484	81.9733	52.0992	87.8339	99.0899	78.3881	97.6256	99.9999
1	42.9864	71.9585	88.3281	56.1529	90.8551	99.4904	79.3906	98.0001	100.0000
2	45.2522	74.8962	90.3886	58.1089	92.6103	99.6776	80.1380	98.3012	100.0000
3	46.0253	75.8487	90.9857	59.1433	93.6035	99.7606	80.7026	98.5414	100.0000
4	46.2760	76.1248	91.1378	59.7245	94.1442	99.7954	81.1343	98.7317	100.0000
5	46.3486	76.1951	91.1719	60.0621	94.4250	99.8094	81.4682	98.8810	100.0000
6	46.3671	76.2108	91.1787	60.2603	94.5635	99.8147	81.7291	98.9972	100.0000
7	46.3713	76.2140	91.1799	60.3754	94.6283	99.8166	81.9351	99.0867	100.0000
8	46.3721	76.2146	91.1801	60.4406	94.6570	99.8173	82.0990	99.1549	100.0000
9	46.3723	76.2147	91.1802	60.4761	94.6690	99.8175	82.2305	99.2063	100.0000
10	46.3723	76.2147	91.1802	60.4945	94.6739	99.8176	82.3367	99.2447	100.0000
11	46.3723	76.2147	91.1802	60.5036	94.6757	99.8176	82.4229	99.2730	100.0000
12	46.3723	76.2147	91.1802	60.5078	94.6764	99.8176	82.4932	99.2937	100.0000
13	46.3723	76.2147	91.1802	60.5097	94.6766	99.8176	82.5508	99.3086	100.0000
14	46.3723	76.2147	91.1802	60.5105	94.6767	99.8176	82.5980	99.3192	100.0000
15	46.3723	76.2147	91.1802	60.5108	94.6767	99.8176	82.6368	99.3267	100.0000
16	46.3723	76.2147	91.1802	60.5109	94.6767	99.8176	82.6687	99.3320	100.0000
17	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.6949	99.3356	100.0000
18	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7163	99.3380	100.0000
19	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7338	99.3397	100.0000
20	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7481	99.3408	100.0000
21	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7596	99.3415	100.0000
22	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7688	99.3420	100.0000
23	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7762	99.3423	100.0000
24	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7821	99.3425	100.0000
25	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7867	99.3426	100.0000
26	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7902	99.3427	100.0000
27	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7930	99.3427	100.0000
28	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7951	99.3428	100.0000
29	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7966	99.3428	100.0000
30	46.3723	76.2147	91.1802	60.5110	94.6767	99.8176	82.7978	99.3428	100.0000

Table C.2: Values of Q_k (in %) for $\gamma = 1$

	$\lambda = 1.8$			$\lambda = 8$			$\lambda = 60$		
k	$s = 1$	$s = 2$	$s = 3$	$s = 5$	$s = 10$	$s = 15$	$s = 50$	$s = 70$	$s = 100$
0	35.7143	63.3484	81.9733	52.0992	87.8339	99.0899	78.3881	97.6256	99.9999
1	41.2371	70.1107	87.1346	55.7787	90.6222	99.4669	79.3752	97.9950	100.0000
2	42.5412	71.7812	88.3617	57.3866	92.0354	99.6247	80.0936	98.2842	100.0000
3	42.8391	72.1385	88.6026	58.1445	92.6993	99.6847	80.6211	98.5063	100.0000
4	42.8970	72.2021	88.6419	58.5085	92.9858	99.7055	81.0121	98.6735	100.0000
5	42.9064	72.2116	88.6473	58.6788	93.0991	99.7122	81.3045	98.7969	100.0000
6	42.9077	72.2129	88.6480	58.7540	93.1400	99.7142	81.5250	98.8861	100.0000
7	42.9079	72.2130	88.6480	58.7848	93.1537	99.7147	81.6922	98.9493	100.0000
8	42.9079	72.2130	88.6480	58.7964	93.1578	99.7148	81.8195	98.9931	100.0000
9	42.9079	72.2130	88.6480	58.8004	93.1590	99.7149	81.9164	99.0228	100.0000
10	42.9079	72.2130	88.6480	58.8017	93.1594	99.7149	81.9901	99.0426	100.0000
11	42.9079	72.2130	88.6480	58.8021	93.1594	99.7149	82.0460	99.0555	100.0000
12	42.9079	72.2130	88.6480	58.8022	93.1595	99.7149	82.0879	99.0637	100.0000
13	42.9079	72.2130	88.6480	58.8022	93.1595	99.7149	82.1192	99.0688	100.0000
14	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1422	99.0720	100.0000
15	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1588	99.0739	100.0000
16	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1708	99.0750	100.0000
17	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1791	99.0756	100.0000
18	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1849	99.0760	100.0000
19	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1888	99.0762	100.0000
20	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1914	99.0763	100.0000
21	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1930	99.0763	100.0000
22	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1941	99.0764	100.0000
23	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1948	99.0764	100.0000
24	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1952	99.0764	100.0000
25	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1954	99.0764	100.0000
26	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1956	99.0764	100.0000
27	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1956	99.0764	100.0000
28	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1957	99.0764	100.0000
29	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1957	99.0764	100.0000
30	42.9079	72.2130	88.6480	58.8023	93.1595	99.7149	82.1957	99.0764	100.0000

Table C.3: Values of Q_k (in %) for $\gamma = 2$

Bibliography

[1] McKinsey Quarterly: The Online Journal of McKinsey & Co. Available at www.mckinseyquarterly.com, 2006.

[2] North American Call Center Summit (NACCS), Call Center Statistics, Call Center Summit on Strategic Outsourcing. Available at www.callcenternews.comresourcesstatistics.shtml, 1999.

[3] J. Abate and W. Whitt. Transient Behavior of the M/M/1 Queue via Laplace Transforms. *Advances in Applied Probability*, 20:145–178, 1988.

[4] J. Abate and W. Whitt. Calculating Time-Dependent Performance Measures for the M/M/1 Queue. *IEEE Transactions on Communications*, 37:1102–1104, 1989.

[5] M. S. Aguir. *Modèles Stochastiques pour l'Aide à la Décision dans les Centres d'Appels*. 2004. Ph.D. Thesis, Ecole Centrale Paris.

[6] M. S. Aguir, F. Karaesmen, O.Z. Akşin, and F. Chauvet. The impact of Retrials on Call Center Performance. *OR Spectrum*, 26:353–376, 2004.

[7] M.S. Aguir, F. Karaesmen, and Y. Dallery. Head-Of-The-Line Priority with Random Assignment. 2006. Working paper, Ecole Centrale Paris, France.

[8] O. Z. Akşin and F. Karaesmen. Designing Flexibility: Characterizing the value of Cross-Training Practices. 2002. Working paper, INSEAD, Fontainebleau, France.

[9] O. Z. Akşin, F. Karaesmen, and E. L. Örmeci. A Review of Workforce Cross-Training in Call Centers from an Operations Management Perspective. 2005. Working paper, Koç University.

[10] C. J. Ancker and A. Gafarian. Queueing with Impatient Customers Who Leave at Random. *Journal of Industrial Engineering*, 13:84–90, 1962.

[11] M. Armony. Dynamic Routing in Large-Scale Service systems with Heterogeneous Servers. *Queueing Systems: Theory and Applications (QUESTA)*, 51:287–329, 2005.

[12] M. Armony and C. Maglaras. Contact Centers with a Call-Back Option and Real-Time Delay Information. *Operations Research*, 52:527–545, 2004.

[13] M. Armony and C. Maglaras. On Customers Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules, and System Design. *Operations Research*, 52:271–292, 2004.

[14] M. Armony, E. Plambeck, and S. Seshadri. Convexity Properties and Comparative Statics for an M/M/S Queue with Impatient Customers: Why You Shouldn't Shout at the DMV. 2005. Submitted for publication.

[15] M. Armony, N. Shimkin, and W. Whitt. The Impact of Delay Announcements in Many-Server Queues with Abandonment. 2005. Working paper, New York University.

[16] R. Atar, A. Mandelbaum, and I.M. Reiman. Scheduling a Multi Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy Traffic. *The Annals of Applied Probability*, 14:1084–1134, 2004.

[17] B. Avi-Itzhak and H. Levy. On Measuring Fairness in Queues. *Advances In Applied Probability*, 36:919–936, 2004.

[18] A.N. Avramidis, A. Deslauriers, and P. l'Ecuyer. Modeling Daily Arrivals to a Telephone Call Center. *Management Science*, 50:896–908, 2004.

[19] F. Baccelli and G. Hebuterne. On Queues With Impatient Customers. *Performance'81 North-Holland Publishing Company*, pages 159–179, 1981.

[20] N. T. J. Bailey. A Continuous Time Treatment of a Single Queue Using Generating Functions. *J. Roy. Statist. Soc. Ser.*, 16:288–291, 1954.

[21] F. Ball and V.T. Stefanov. Further Approaches to Computing Fondamental Characteristics of Birth-Death Processes. *Journal of Applied Probability*, 38:995–1005, 2001.

[22] F. Baskett, K.M. Chandy, R.R. Muntz, and F. Palacios-Gomez. Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22:248–260, 1975.

[23] A. Bassamboo, J.M. Harrison, and A. Zeevi. Dynamic Routing and Admission Control in High-Volume Service Systems: Asymptotic via Multi-Scale Fluid Models. 2004. Submitted for publication.

[24] S. Benjaafar. Performance Bounds for the Effectiveness of Pooling in Multi-Processing Systems. *European Journal of Operational Research*, 87:375–388, 1995.

[25] A.W. Berger and W. Whitt. Comparisons of Multi-Server Queues with Finite Waiting Rooms. *Stochastic Models*, 8:719–732, 1992.

[26] P.P. Bhattacharya and A. Ephremides. Stochastic Monotonicity Properties of Multiserver Queues with Impatient Customers. *Journal of Applied Probability*, 28:673–682, 1991.

[27] S. Bhulai and G. Koole. A Queueing Model for Call Blending in Call Centers. *IEEE Transactions on Automatic Control*, 48:1434–1438, 2003.

[28] S. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning Large Call Centers. *Operations Research*, 52:17–34, 2004.

[29] J. Boudreau. Organizational Behavior, Strategy, Performance, and Design in Management Science. *Management Science*, 50:1463–1476, 2004.

[30] J. Boudreau, W. Hopp, J.O. McClain, and L.J Thomas. On the Interface between Operations and Human Ressources Management. *Manufacturing & Service Operations Management*, 5:179–202, 2003.

[31] A. Brandt and M. Brandt. Assymptotic Results and a Markovian Approximation for the M(n)/M(n)/C + GI System. *Queueing Systems: Theory and Applications (QUESTA)*, 41:73–94, 2002.

[32] Z. Carmon and D. Kahenman. The Experienced Utility of Queueing: Experience Profiles and Retrospective Evaluatoins of Simulated Queues. pre-print.

[33] D. G. Champernowne. An Elementary Method of Solution of the Queueing Problem with a Single Server and a Constant Parameter. *J. Roy. Statist. Soc. Ser.*, 18:125–128, 1956.

[34] P. Chevalier, R.A. Shumsky, and N. Tabordon. Routing and Staffing in Large Call Centers with Specialized and Fully Flexible Servers. 2004. Université catholique de Louvain, University of Rochester and Belgacom Mobile/Proximus. Working paper.

[35] P. Coolen-Schrijner and E.A. van Doorn. The Deviation Matrix of a Continuous-Time Markov Chain. *Probability in the Engineering and Informational Sciences*, 16:351–366, 2002.

[36] F. de Véricourt and Y.-P. Zhou. Managing Response Time in a Call Routing Problem with Service Failure. *Operations Research*, 53:968–981, 2005.

[37] P. Feigin. Analysis of Customer Patience in a Bank Call Center. 2005. Working Paper, The Technion.

[38] M. Fischer, D. Garbin, A. Gharakhanian, and D. Masi. Traffic Engineering of Distributed Call Centers: Not as Straight Forward as it May Seem. 1999. Mitretek Systems.

[39] P. Flajolet and F. Guillemin. The Formal Teory of Birth-and-Death Processes, Lattice Path Combinatorics and Continued Fractions. *Advances in Applied Probability*, 32:750–778, 2000.

[40] N. Gans, G. Koole, and A. Mandelbaum. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5:73–141, 2003.

[41] N. Gans and G. van Ryzin. Optimal Dynamic Scheduling of a General Class of Parallel-Processing Queueing Systems. *Advances in Applied Probability*, 30:1130–1156, 1998.

[42] N. Gans and Y.P. Zhou. A Call-Routing Problem with Service-Level Constraints. *Operations Research*, 51:255–271, 2003.

[43] O. Garnett and A. Mandelbaum. An Introduction to Skills-Based Routing and its Operational Complexities. 2001. Teaching notes, Technion.

[44] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.

[45] W. Grassmann. The Convexity of the Mean Queue Size of the M/M/C Queue with Respect to the Traffic Intensity. *Journal of Applied Probability*, 20:916–919, 1987.

[46] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. Wiley series in probability and mathematical statistics, 1985. 2rd edition.

[47] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. Wiley series in probability and mathematical statistics, 1998. 3rd edition.

[48] R. Guérin. Queueing-Blocking System with two Arrival Streams and Guard Channels. *IEEE Transactions on Communications*, 36:153–163, 1998.

[49] F. Guillemin. Spectral Analysis of Birth and Death Processes. 2005. Working paper, submitted to Journal of Applied Probability.

[50] F. Guillemin and D. Pinchon. Excursions of Birth and Death Processes, Orthogonal Polynomials, and Continued Fractions. *Journal of Applied Probability*, 36:752–770, 1999.

[51] P. Guo and P. Zipkin. Analysis and Comparaison of Queues with Different Levels of Delay Information. 2004. Working paper, Duke University.

[52] S. Halfin and W. Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29:567–588, 1981.

[53] A. Harel. Convexity of the Erlang Loss Formula. *Operations Research*, 38:499–505, 1990.

[54] A. Harel. Convexity Results for Single-Server Queue and for Multiserver Queues with Constant Service Times. *Journal of Applied Probability*, 27:465–468, 1990.

[55] A. Harel and P. Zipkin. The Convexity of a General Performance Measure for Multiserver Queues. *Journal of Applied Probability*, 24:725–736, 1987.

[56] A. Harel and P. H. Zipkin. Strong Convexity Results for Queueing Systems. *Operations Research*, 35:405–418, 1987.

[57] W.J. Hopp and M.P. van Oyen. Agile Workforce Evaluation: A Framework for Cross-Training and Coordination. *IIE Transactions*, 36:919–940, 2004.

[58] T.Y. Huang. Analysis and Modeling of a Threshold Based Priority Queueing System. *Computer Communications*, 24:284–291, 2001.

[59] M. Hui and D. Tse. What to Tell Customer in Waits of Different Lengths: an Integrative Model of Service Evaluation. *Journal of Marketing*, 60:81–90, 1996.

[60] M. Hui and L. Zhou. How Does Waiting Duration Information Influence Customers' Reactions to Waiting for Services? *Journal of Applied Social Psychology*, 26:1702–1717, 1996.

[61] D. L. Jagerman. Some Properties of the Erlang Loss Function. *The Bell System Technical Journal*, 53:525–551, 1974.

[62] A. A. Jagers and E. A. van Doorn. Convexity of Functions Which Are Generalizations of the Erlang Loss Function and the Erlang Delay Function. *SIAM Review*, 33:281–282, 1991.

[63] G. Jongbloed and G.M. Koole. Managing Uncertainty in Call Centers using Poisson Mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

[64] O. Jouini and Y. Dallery. Estimating and Announcing Waiting Times in Multiple Customer Class Call Centers. *Proceedings of INCOM 2006*, 2:371–376, 2006.

[65] O. Jouini and Y. Dallery. Moments of First Passage Times in General Birth-Death Processes. 2006. Submitted for publication.

[66] O. Jouini and Y. Dallery. Monotonicity Properties for Multiserver Queues with Reneging and Finite Waiting Lines. 2006. To appear in Probability in the Engineering and Informational Sciences.

[67] O. Jouini and Y. Dallery. Predicting Queueing Delays for Multiclass Call Centers. *Proceedings of VALUETOOLS 2006*, 2006.

[68] O. Jouini, Y. Dallery, and O. Z. Akşin. Modeling Call Centers with Delays Information. 2006. Working paper, Ecole Centrale Paris and Koç University.

[69] O. Jouini, Y. Dallery, and R. Nait-Abdallah. Analysis of the Impact of Team-Based Organizations in Call Centers Management. 2006. To appear in Management Science.

[70] O. Jouini, A. Pot, Y. Dallery, and G. Koole. Real-time Dynamic Scheduling Policies for Multiclass Call Centers with Impatient Customers. 2006. Working paper. Ecole Centrale Paris and Vrije Universiteit Amsterdam.

[71] S. Karlin and J. McGregor. The Classification of Birth and Death Processes. *Trans. Amer. Math. Soc.*, 86:366–401, 1957.

[72] S. Karlin and J. McGregor. The Differential Equation of Birth and Death Processes, and the Setieltjes Moment Problem. *Trans. Amer. Math. Soc.*, 85:489–546, 1957.

[73] K. Katz, B. Larson, and R. Larson. Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage. *Sloan Management Review*, pages 44–53, 1991.

[74] J. Keilson. A Review of Transient Behavior in Regular Diffusion and Birth-Death Processes. Part i. *Journal of Applied Probability*, 1:247–266, 1964.

[75] J. Keilson. A Review of Transient Behavior in Regular Diffusion and Birth-Death Processes. Part ii. *Journal of Applied Probability*, 1:247–266, 1964.

[76] J. Keilson. *Markov Chain Models - Rarity and Exponentiality*. Springer-Verlag, New York, 1979.

[77] J. Keilson. On the Unimodality of Passage Time Densities in Birth-Death Processes. *Statist. Neerlandica*, 35:49–55, 1981.

[78] O. Kella and U. Yechiali. Waiting Times in the Non-Preemptive Priority M/M/c Queue. *Stochastic Models*, 1:257–262, 1985.

[79] F.P. Kelly. Loss Networks. *Annals of Applied Probability*, 1:319–378, 1991.

[80] M. Kijima. *Markov Processes for Stochastic Modeling*. Chapman & Hall, 1997. First edition.

[81] L. Kleinrock. *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication, 1975.

[82] L. Kleinrock. *Queueing Systems, Computer Applications*, volume II. A Wiley-Interscience Publication, 1976.

[83] G. Koole. Convexity in Tandem Queues. *Probability in the Engineering and Informational Sciences*, 18:13–31, 2004.

[84] G. Koole. A Formula for Tail Probabilities of Cox Distributions. *Journal of Applied Probability*, 41:935–938, 2004.

[85] G. Koole and A. Mandelbaum. Queueing Models of Call Centers An Introduction. *Annals of Operations Research*, 113:41–59, 2002. abridged version.

[86] G. Koole, M. Nuyens, and R. Righter. The Effect of Service Time Variability on Maximum Queue Lengths in MX/G/1 Queues. *Journal of Applied Probability*, 42, 2005. To appear.

[87] G. Koole and A. Pot. A Note on Profit Maximization and Monotonicity for Inbound Call Centers. 2005. Submitted for publication.

[88] G. Koole and A. Pot. An Overview of Routing and Staffing Algorithms in Multi-Skill Customer Contact Center. 2006. Submitted for publication.

[89] A. Krinik and Y. Sourouri. Taylor Series Solutions of Classical Queueing Systems. *Abstracts American Mathematical Society*, 11, 1990.

[90] C.R. Larson. Perspectives on Queues: Social Justice and the Psychology of Queueing. *Operations Research*, 35:895–905, 1987.

[91] H. L. Lee and M. A. Cohen. A Note on the Convexity of Performance Measures of M/M/C Queueing Systems. *Journal of Applied Probability*, 20:920–923, 1983.

[92] P. Leguesdron, J. Pellaumail, G. Rubino, and B. Sericola. Transient Analysis of the M/M/1 Queue. *Advances in Applied Probability*, 25:702–713, 1993.

[93] Y. Lu and M.S. Squillante. Scheduling to Minimize General Functions of the Mean and Variance of Sojourn Times in Queueing Systems. 2004. Working Paper, IBM Research Division.

[94] D. Maister. Psychology of Waiting Lines. *Harvard Business School Cases*, pages 71–78, 1984.

[95] A. Mandelbaum. Call Centers (Centres): Research Bibliography with Abstracts. 2002. Version 3, 137 pages. Downloadable from ie.technion.ac.il/serveng/References/ccbib.pdf.

[96] A. Mandelbaum, Sakov A., and S. Zeltyn. Empirical Analysis of a Call Center. 2000. Technical Report, Technion.

[97] A. Mandelbaum and M.I. Reiman. On Pooling in Queueing Networks. *Management Science*, 44:971–981, 1998.

[98] A. Mandelbaum and S. Zeltyn. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. 2006. Working paper, Thechnion, Haifa, Israel.

[99] Y. H. Mao. Ergodic Degrees For Continuous-Time Markov Chains. *Science in China Ser. A*, 47:161–174, 2004.

[100] L. Meister and J. G. Shanthikumar. Concavity of the Throughput of Tandem Queueing System with Finite Buffer Storage Space. *Advances In Applied Probability*, 22:764–767, 1990.

[101] E. J. Messerli. Proof of a Convexity Property of the Erlang B Formula. *The Bell System Technical Journal*, 51:951–553, 1972.

[102] R. Nagarajan and D. Towsley. A Note on the Convexity of the Probability of a Full Buffer in the M/M/1/K queue. 1992. CMPSCI Technical Report TR 92-85, MIT.

[103] E. Nakibly. *Predicting Waiting Times in Telephone Service Systems.* 2002. Ph.D. Thesis, The Senate of the Technion.

[104] P. Naor. The Regulation of Queue Size by Levying Tolls. *Econometrica*, 37:15–24, 1969.

[105] E.L. Örmeci. Dynamic Admission Control in a Call Center with One Shared and Two Dedicated Service Facilities. *IEEE Transactions on Automatic Control*, 49:1157–1161, 2004.

[106] A. Pacheco. Second-Order Properties of the Loss Probability in M/M/s/s + c Systems. *Queueing Systems: Theory and Applications (QUESTA)*, 15:309–324, 1994.

[107] P. R. Parthasarathy. A Transient Solution to a M/M/1 Queue: A New Simple Approach. *Advances in Applied Probability*, 19:997–998, 1987.

[108] E. Pekoz. Optimal Policies for Multi-Server Non-Preemptive Priority queues. *Queueing Systems: Theory and Applications (QUESTA)*, 42:91–101, 2002.

[109] M. P. Pierson and W. Whitt. A Statistically-Fit Markovian Approximation of a Basic Call-Center Model. 2006. Working paper, Columbia University.

[110] M. Pinedo, Seshadri S., and Shantikumar J.G. Call Centers in Financial Services: Strategies, Technologies, and Operations. *In E.L. Melnick, P. Nayyar, M.L. Pinedo, and S. Seshadri, editors, Creating Value in Financial Services: Strategies, Technologies, and Operations*, 1990. Kluwer.

[111] V. Pla, V. Casares-Giner, and Martínez. On a Multiserver Finite Buffer Queue with Impatient Customers. *Proceedings of the ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems*, 2004.

[112] A. Pot. *Routing and Planning Algorithms for Multi-Skill Contact Centers.* 2006. Ph.D. Thesis, Vrije Universiteit Amsterdam, The Netherlands.

[113] W. H. Randolph. *Queueing Methods for Services and Manufacturing.* Prentice Hall, 1991.

[114] S. I. Rosenlund. Upwards Passage Times in the Non-Negative Birth-Death Process. *Scand. J. Statist.*, 4:90–92, 1977.

[115] M.H. Rothkopf and P. Rech. Perspectives on Queues: Combining Queues is not Always Beneficial. *Operations Research*, 35:906–909, 1987.

[116] R. Schonberger. *World Class Manufacturing: The Lessons of Simplicity Applied.* Free Press, New York, 1986. 10-11.

[117] L.E. Schrage and L.W. Miller. The Queue M/G/1 with the Shortest Remaining Processing Time Discipline. *Operations Research*, 14:670–684, 1966.

[118] M. Shaked and J. G. Shanthikumar. Stochastic Convexity and Its Applications. *Advances in Applied Probability*, 20:427–446, 1988.

[119] J. G. Shanthikumar. Stochastic Majorization of Random Variables with Proportional Equilibrium Rate. *Advances in Applied Probability*, 19:854–872, 1987.

[120] D.R. Smith and W. Whitt. Resource Sharing for Efficiency in Traffic Systems. *The Bell System Technical Journal*, 60:39–55, 1981.

[121] U. Sumita. On Conditional Passage Time Structure of Birth-Death Processes. *Journal of Applied Probability*, 21:10–21, 1984.

[122] L. Takâcs. *Introduction to the Theory of Queues.* Oxford University Press, 1960.

[123] A. M. K. Tarabia. Transient Analysis of M/M/1/N Queue - An Alternative Approach. *Tamkang Journal of Science and Engineering*, 3:263–266, 2000.

[124] S. Taylor. Waiting for Service: The Relationship Between Delays and Evaluations of Service. *Journal of Marketing*, 58:56–69, 1994.

[125] E. Tekin, W.J. Hopp, and M.P. varOyen. Pooling Strategies for Call Center Agent Cross-Training. 2004. Submitted for publication.

[126] H. Y. Tu and H. Kumin. A Convexity Result for a Class of GI/G/1 Queueing Systems. *Operations Research*, 31:948–950, 1983.

[127] N.M. van Dijk and E. van der Sluis. Check-in Computation and Optimization by Simulation and IP in Combination. *European Journal of Operational Research*, 171:1152–1168, 2006.

[128] R.B. Wallace and W. Whitt. A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*, 2005.

[129] A. R. Ward and W. Whitt. Predicting Response Times in Processor-Sharing Queues. *Proceedings of the Fields Institute Conference on Communication Networks*, 2000.

[130] A.R. Ward and P.W. Glynn. A Diffusion Approximation for a Markovian Queue with Reneging. *Queueing Systems: Theory and Applications (QUESTA)*, 43:103–128, 2003.

[131] R. R. Weber. On the Marginal Benefit of Adding Servers to G/GI/m Queues. *Management Science*, 26:946–951, 1980.

[132] R. R. Weber. A Note on Waiting Times in Single Server Queues. *Operations Research*, 31:950–951, 1983.

[133] W. Whitt. Blocking when Service is Required from Several Facilities Simultaneously. *AT&T Technical Journal*, 64:1807–1856, 1985.

[134] W. Whitt. Counterexamples for Comparisons of Queues with Finite Waiting Rooms. *Queueing Systems: Theory and Applications (QUESTA)*, 10:271–278, 1992.

[135] W. Whitt. Improving Service by Informing Customers about Anticipated Delays. *Management Science*, 45:192–207, 1999.

[136] W. Whitt. Partitioning Customers into Service Groups. *Management Science*, 45:1579–1592, 1999.

[137] W. Whitt. Predicting Queueing Delays. *Management Science*, 45:870–888, 1999.

[138] W. Whitt. Stochastic Models for the Design and Management of Customer Contact Centers: Some Research Directions. 2002. Working paper, Columbia University.

[139] W. Whitt. Engineering Solution of a Basic Call-Center Model. *Management Science*, 51:221–235, 2005.

[140] W. Whitt. Sensitivity of Performance in the Erlang-A Queueing Model to Changes in the Model Parameters. *Operations Research*, 54:247–260, 2006.

[141] R.W. Wolff. Poisson Arrivals See Time Averages. *Operations Research*, 30:223–231, 1982.

[142] S.H. Xu, R. Righter, and Shanthikumar J.G. Optimal Dynamic Assignment of Customers to Heterogeneous Servers in Parallel. *Operations Research*, 40:1126–1138, 1992.

[143] D. D. Yao and J. G. Shanthikumar. The Optimal Input Rate to a System of Manufacturing Cells. *Information Systems and Operational Research*, 25:57–65, 1987.

[144] D. Zakay. An Integrated Model of Time Estimation. *Times and Human Cognition: A Life Span Perspective*, 1989. Iris Levin and Dan Zakay, eds, Amsterdam: North Holland.

[145] S. Zeltyn and A. Mandelbaum. Call Centers with Impatient Customers: Many-Servers Asymptotics of the M/M/n+G Queue. *Queueing Systems: Theory and Applications (QUESTA)*, 51:361–402, 2005.

[146] E. Zohar, A. Mandelbaum, and Shimkin N. Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support. *Management Science*, 48:566–583, 2002.

Index

Résumé Depuis quelques années, les centres d'appels enregistrent une forte croissance dans le monde. Les entreprises s'orientent de plus en plus vers ce choix qui leur offre une relation privilégiée avec leurs clients. Ainsi, ils disposent d'un moyen convivial et peu coûteux pour fidéliser leurs clients tout en essayant d'en acquérir de nouveaux. Le sujet de cette thèse porte sur le développement et l'analyse de modèles stochastiques pour l'aide à la décision dans les centres d'appels.

Dans la première partie, nous considérons un centre d'appels où tous les agents sont groupés dans un même pool et les clients sont traités indifféremment par un des agents. Nous étudions les bénéfices de la migration depuis cette configuration vers un centre d'appels où les clients sont divisés en classes (appelées portefeuilles de clients). Chaque portefeuille de clients est servi par un pool de conseillers qui lui est exclusivement dédié. Ensuite, nous considérons un centre d'appels avec deux classes de clients impatients. Nous développons des politiques dynamiques pour l'affectation des clients (selon leurs types) aux différentes files d'attente. L'objectif étant lié aux qualités de service différentiées exprimées en terme du pourcentage des clients perdus, ainsi qu'en terme de la variance du temps d'attente. Enfin, nous étudions un centre d'appels qui annonce le délai d'attente à chaque nouveau client. Nous montrons les avantages de l'annonce sur les performances du centre d'appels.

Dans la deuxième partie, nous considérons un processus de naissance et de mort de forme générale. Nous calculons ensuite les moments de plusieurs variables aléatoires liées aux temps de premiers passages (ordinaires et conditionnels). Ensuite, nous montrons un résultat de concavité dans une file d'attente avec capacité limitée et avec une seule classe de clients impatients. Nous démontrons que la probabilité d'entrer en service est strictement croissante et concave en fonction de la taille de la file d'attente.

Mots-clefs centres d'appels, modèles stochastiques, files d'attente, chaînes de Markov, simulation, politique de routage, qualité de service, analyse transitoire, abandons, priorité non-préemptive

Abstract In the past few years, call centers have been introduced with great success by many service-oriented companies such as banks and insurance companies. They become the main point of contact with the customer, and an integral part of the majority of corporations. The large-scale emergence of call centers has created a fertile source of management issues. In this thesis, we focus on various operations management issues of call centers. The objective of our work is to derive, both qualitative and quantitative, results for practical management.

In the first part of the thesis, we investigate the impact of team-based organizations in call centers management. We develop queueing models that show that the benefits of the team based organization in providing more efficient answers to customers very often outweigh its drawback coming from the loss of pooling. Next, we consider a two-class call center and develop real-time scheduling policies that determine the rule of assignment of new arrivals to the waiting lines. We focus on service levels criteria related to the fraction of abandoning customers and the variance of queueing delays. Finally, we propose a call center model in which we provide information about delays to customers, and we quantify its effect upon performance.

In the second part of the thesis, we tackled the quantitative analysis of stochastic processes and queueing models. First, we derive several closed-form expressions of the moments of first passage times in general birth-death processes, and we point out their applications. Second, we investigate some monotonicity results for the probability of being served in markovian queueing systems with impatient customers.

Keywords call centers, stochastic models, queueing theory, Markov chains, simulation, scheduling policies, quality of service, transient analysis, reneging, non-preemptive priority.