ÉCOLE CENTRALE DES ARTS ET MANUFACTURES « ÉCOLE CENTRALE PARIS »

THÈSE

présentée par

Benjamin LEGROS

pour l'obtention du

GRADE DE DOCTEUR

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

SUJET :

OPTIMIZATION OF MULTI-CHANNEL AND MULTI-SKILL CALL CENTERS

soutenue le 13 décembre 2013
devant un jury composé de :

Stephen E. CHICK — **Président**
Professeur, INSEAD, Fontainebleau

Zeynep AKSIN KARAESMEN — **Rapporteur**
Professeur, Koç University, Istanbul

Raik STOLLETZ — **Rapporteur**
Professeur, University of Mannheim, Mannheim

Rob VAN DER MEI — **Examinateur**
Professeur, Vrije Universiteit Amsterdam, Amsterdam

Sébastien THOREL — **Examinateur**
Interact-iv, Paris

Yves DALLERY — **Directeur de thèse**
Professeur, Ecole Centrale Paris, Paris

Ger KOOLE — **Co-encadrant**
Professeur, Vrije Universiteit Amsterdam, Amsterdam

Oualid JOUINI — **Co-encadrant**
Maître de conférences, Ecole Centrale Paris, Paris

2013ECAP0072

Remerciements

Je tiens à exprimer toute ma gratitude à Zeynep Aksin et Raik Stolletz pour avoir accepté l'évaluation de cette thèse en tant que rapporteurs.

Je remercie, également, Stephen Chick pour l'honneur qu'il me fait en présidant le jury de cette thèse, ainsi que Rob Van Der Mei et Sébastien Thorel pour avoir accepté d'en être examinateurs.

Je remercie Yves Dallery pour l'encadrement et le soutien qu'il m'a accordé.

Je tiens à remercier la société Interact-iv et à travers elle Sébastien Thorel, Mohamed Jribi et Olivier Aimé. Nos échanges ont été à l'origine de nombreuses problématiques étudiées dans cette thèse. Ils ont permis un réel dialogue entre nos univers de travail.

Je remercie sincèrement Oualid Jouini pour ces trois années de travail et d'échanges, pour son encadrement et ses qualités humaines. Qualités humaines salutaires à certains moments.

Je remercie aussi chaleureusement Ger Koole pour son investissement sur mon travail, son expérience partagée de la recherche sur les centres d'appels et son accueil à Amsterdam.

Je garderai un excellent souvenir des discussions diverses que j'ai eues avec mes collègues du laboratoire.

Enfin, je rends hommage à tous les membres du Laboratoire Génie Industriel. En particulier Sylvie, Delphine et Corinne pour leur aide et leur soutien à de nombreux moments pendant ces trois années..

Benjamin.

ÉCOLE CENTRALE DES ARTS ET MANUFACTURES « ÉCOLE CENTRALE PARIS »

THÈSE

présentée par

Benjamin LEGROS

pour l'obtention du

GRADE DE DOCTEUR

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

SUJET :

OPTIMIZATION OF MULTI-CHANNEL AND MULTI-SKILL CALL CENTERS

soutenue le 13 décembre 2013
devant un jury composé de :

Zeynep AKSIN KARAESMEN Professeur, Koç University, Istanbul	**Rapporteur**
Raik STOLLETZ Professeur, University of Mannheim, Mannheim	**Rapporteur**
Stephen E. CHICK Professeur, INSEAD, Fontainebleau	**Examinateur**
Rob VAN DER MEI Professeur, Vrije Universiteit Amsterdam, Amsterdam	**Examinateur**
Sébastien THOREL Interact-iv, Paris	**Examinateur**
Yves DALLERY Professeur, Ecole Centrale Paris, Paris	**Directeur de thèse**
Ger KOOLE Professeur, Vrije Universiteit Amsterdam, Amsterdam	**Co-encadrant**
Oualid JOUINI Maître de conférences, Ecole Centrale Paris, Paris	**Co-encadrant**

Remerciements

Je tiens à exprimer toute ma gratitude à Zeynep Aksin et Raik Stolletz pour avoir accepté l'évaluation de cette thèse en tant que rapporteurs.

Je remercie, également, Stephen Chick pour l'honneur qu'il me fait en présidant le jury de cette thèse, ainsi que Rob Van Der Mei et Sébastien Thorel pour avoir accepté d'en être examinateurs.

Je remercie Yves Dallery pour l'encadrement et le soutien qu'il m'a accordé.

Je tiens à remercier la société Interact-iv et à travers elle Sébastien Thorel, Mohamed Jribi et Olivier Aimé. Nos échanges ont été à l'origine de nombreuses problématiques étudiées dans cette thèse. Ils ont permis un réel dialogue entre nos univers de travail.

Je remercie sincèrement Oualid Jouini pour ces trois années de travail et d'échanges, pour son encadrement et ses qualités humaines. Qualités humaines salutaires à certains moments.

Je remercie aussi chaleureusement Ger Koole pour son investissement sur mon travail, son expérience partagée de la recherche sur les centres d'appels et son accueil à Amsterdam.

Je garderai un excellent souvenir des discussions diverses que j'ai eues avec mes collègues du laboratoire.

Enfin, je rends hommage à tous les membres du Laboratoire Génie Industriel. En particulier Sylvie, Delphine et Corinne pour leur aide et leur soutien à de nombreux moments pendant ces trois années..

Benjamin.

Contents

Chapter 1

Introduction

In this chapter, we give a general introduction of the manuscript. First, we describe the context and the motivation of the thesis. Second, we highlight the approach, the structure and the main contributions of this work.

1.1 Context and Motivation

A call center is a service system. It is a facility designed to support the delivery of some interactive service via telephone communications, email, chat, etc. The definition of a call center is continuously changing with technological development, 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. This thesis focuses on operations management issues for multi-skill and multi-channel call centers. In what follows, we first present the context of the call center industry and the related operations management issues. Second, we focus on the motivation of this work, and present our collaboration with the French consulting company Interact-iv, which was at the origin of most of the addressed problems in this thesis.

Context. Call centers make up a large and growing part of the global economy. They are very labor-intensive operations, employing millions of persons across the globe. Call centers serve as the public face in various areas and industries: insurance companies, emergency centers, banks, information centers, help-desks, tele-marketing, etc. The success of call centers is due to the technological advances in information and communications systems, see Pinedo et al. (1999). The most important call centers equipments are the Interactive Voice Response (IVR), the Automated Call Distributor (ACD), and the Computer Telephone Integration (CTI). These technologies have grown cheaper, more reliable, and more sophisticated. Moreover, these advances enabled various call center tasks which requires multiple skills and channels.

In multi-skill call centers, the call assignment strategy of Skills-Based Routing (SBR), is used to assign incoming calls to the most suitable agent. The report of Holman et al. (2007) made on 2500 call centers in 17 countries with 475,000 employees points out that 56% of call centers use SBR strategies. These strategies are an enhancement to ACD systems.

Next, the development of alternative channels goes together with an adaptation to impatient customers with higher expectations. The recent report of ICMI (2013), based on the analysis of 361 large contact centers, presents the increasing use of new channels and the related research issues. In particular, they point out that outbound tasks require intensive integration with inbound ones in most call centers. Although the inbound calls remain present in most call centers (98%), emails are also widely used (89%). Moreover, outbound calls (76%), Web (70%) and chats (40%) are important and developing channels. We refer the reader to the description of the general context of multi-channel call centers management in Chapter 7 of Koole (2013).

Due to the operational difficulties to find better solutions than intuitive ones, managers have a continuous interest in the related research disciplines. The literature on operations management in call centers has focused on the following issues: demand forecasting, quality of service, capacity planning, queueing, call routing, staffing and agents scheduling. The related main academical dis-

ciplines are Mathematics and Statistics, Operations Research, Industrial Engineering, Information Technology, Human Resource Management, as well as Psychology and Sociology. The overall purpose of this literature is helping the manager in improving the management of their call centers. We refer the reader to the complete surveys of the academic literature on call center operations management by Gans et al. (2003) and Akşin et al. (2007).

The goal of the present thesis is to contribute to the operations management research in multi-skill and multi-channel call centers. The purpose is to enhance our understanding of such complex systems, so as we obtain useful guidelines for the practitioners.

Motivation. In what follows, we want to motivate the problems under consideration. These are related to flexibility in multi-skill call centers and also to the routing issues in multi-channel call centers.

The concept of flexibility is related to the ability of a company to efficiently match its capacity to an uncertain demand with multiple types. A wide literature has focused on the distribution of skills par agent. Increasing the number of skills per agent goes together with a better use of the resources but also with more costly resources. A well known studied and efficient configuration is *chaining*, first pointed out by Jordan and Graves (1995) in the context of manufacturing systems. In this configuration each agent has only two skills and the distribution of the skills corresponds to a chain. A precise definition of the chaining architecture will be given in Chapter 2. Developing intelligent configurations such as chaining is very interesting for practitioners. The value of these configurations is that they capture the benefits of pooling by only having a limited flexibility. However, the robustness of chaining fails in the case of asymmetric parameters (Sheikhzadeh et al. (1998)). The situations with asymmetric parameters arise in practice. The typical example is that of an European multilingual call center where customers call from several countries. For instance in Bluelink (a service provider of Air France KLM), each agent speaks two languages: her own native language and English. The workload is unbalanced ranging from only some few calls from a given

country to several thousand of calls from another country. For such cases in practice, it is important to develop new architectures that allow on the one hand to account for demand asymmetry, and on the other hand to capture the benefits of pooling with only a limited flexibility.

Call centers require a very accurate match of demand and supply. Since the volatility of call arrival patterns is high, there is often a mismatch between demand and the scheduled number of inbound agents. Moreover, even if the demand is accurately forecasted, a considerable overcapacity should be scheduled to be able to deal with the random Poisson fluctuations of the demand. To prevent idle overcapacity and to limit the necessity to have extremely accurate forecast, inbound calls are sometimes mixed with other types of channels which have a less strict allowable delay, such as emails or outbound calls. This is referred to as *(call) blending*. It arises in the context of multi-channel call centers. Next, we describe some motivational examples for blended operations issues.

In practice, we may find situations where a conversation between an agent and a customer contains a *natural break*. For example, an agent of an internet hotline asks the customer to reboot her modem or her computer which may take some time where no interactions can take place. It is also often the case that a call center agent of an electricity supplier company asks the customer for the serial number of her electricity meter box. Another example is that of commercial call centers with a financial transaction during the call conversation. Inside an underway conversation, the agent is then free to do another task if needed. For an efficient use of the agent time, there might be an opportunity to route the less urgent jobs (emails) to agents, not only when the system is empty of calls, but also during the call conversations. An interesting research question here is how should be the routing rules as a function of the system parameters.

Even in the classical case of a single stage call conversation, the ACD programming is still a complex task for a blending situation. Bhulai and Koole (2003) and Gans and Zhou (2003b) show that efficient assignment policies are those with agent reservation for inbound calls. The

main complexity comes from the fluctuation of the system parameters, in particular those of the jobs arrival processes. For instance from the statistical analysis of a call center data provided by Interact-iv, we observe that during a given period the arrival volume can triple from one day to another. The reasons of the fluctuations is hard to determine and an observation can hardy be duplicated on a future period. Given this fact, one could focus on developing routing policies with a continuous adaptation of the agent reservation threshold, while using at a minimal level the forecasted system parameters.

In the context of highly congested call centers, the use of a callback option can be proposed to customers so as to balance workload and avoid excessive abandonments. Since a callback option transforms an inbound call into an outbound one, the issue in the management of this option is somewhat similar to that of a blended situation. Some practical specific problem can be pointed out: What should be the routing rules of the jobs in the ACD, in order to optimize the system performance in terms of the waiting times of inbound and outbound jobs?

This work is done, in its major part, under a collaboration with the French consulting company Interact-iv. Interact-iv sells software, advice and methods to call centers. The customers of Interact-iv are for a large part small multi-channel call centers. Through the collaboration, the purpose of Interact-iv is to provide to its customers (call center managers) solutions that are thoroughly supported quantitatively. We had the opportunity to work on various issues of multi-channel call centers and had access to real call center data. This collaboration offered a wealth of learning opportunities.

1.2 Structure and Main Contributions

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

The current thesis can be divided into two parts. The topic of the first part is the design of

multi-skill call center architectures. It corresponds to Chapter 2. The topic of the second part is the optimal routing in multi-channel call centers. It corresponds to Chapters 3, 4 and 5.

In Chapter 2, we focus on architectures with limited flexibility for multi-skill call centers. The context is that of call centers with asymmetric parameters: unbalanced workload, different service requirements, a predominant customer type, unbalanced abandonments and high costs of cross-training. The well known architectures with limited flexibility such as chaining fail against such asymmetry. We propose a new architecture referred to as single pooling with only two skills per agent. We provide a comparison framework between chaining and single pooling and demonstrate the efficiency of single pooling under various situations of asymmetry. We also develop analytical results for particular single pooling models, in order to get some sense on the effect of arrival asymmetry on performance. This Chapter is based on Legros et al. (2012) (under second round revision in *International Journal of Production Economics*).

In the second part, we focus on routing problems in multi-channel call centers. In Chapter 3, we consider a blended call center with calls arriving over time and an infinitely backlogged queue of emails. The call service is characterized by three successive stages where the second one is a break. We define parameters of control for the routing of emails between or inside calls treatment. Next, we develop a method based on the analysis of Markov chains in order to derive the performance measures of interest for calls and for emails. We focus on optimizing the email routing parameters. In addition, we develop an approximation method for the system performance evaluation under the light-traffic regime. We also propose an approximation method to extend the results to the multi-server case. We derive various structural results and conclude that all the time at least one of the two email routing parameters has an extreme value. This chapter is based on Legros et al. (2013c) (submitted to *Stochatic Systems*).

In Chapter 4, we examine a threshold policy that reserves agents for inbound calls. We study a general non-stationary model where the call arrival follows a non-homogeneous Poisson process. The

optimization problem consists of maximizing the throughput of outbound tasks under a constraint on the waiting time of inbound calls. We propose an efficient adaptive threshold policy easy to implement in the Automatic Call Distributor (ACD). This scheduling policy is evaluated through a comparison with the optimal performance measures found in the case of a constant stationary arrival rate, and also a comparison with other intuitive adaptive threshold policies in the general non-stationary case. This chapter is based on Legros et al. (2013a) (submitted to *IIE Transactions*).

In Chapter 5, we consider a call center model with a callback option, which allows to transform an inbound call into an outbound one. The optimization problem consists of minimizing the expected waiting time of the outbound calls while respecting a service level constraint on the inbound ones. We propose a routing policy with two thresholds, one on the reservation of the agents for inbound calls, and another on the number of waiting outbound calls. A curve relating the two thresholds is determined. This chapter is based on the ongoing paper Legros et al. (2013b).

In Chapter 6, we close the thesis by giving general concluding remarks and highlighting directions for future research.

Chapter 2

A Flexible Architecture for Call Centers with Skill-Based Routing

We focus on architectures with limited flexibility for multi-skill call centers. The context is that of call centers with asymmetric parameters: unbalanced workload, different service requirements, a predominant customer type, unbalanced abandonments and high costs of cross-training. The most knowing architectures with limited flexibility such as chaining fail against such asymmetry. In this paper, we propose a new architecture referred to as single pooling with only two skills per agent and we demonstrate its efficiency. We conduct a comprehensive comparison between this novel architecture and chaining. As a function of the various system parameters, we delimit the regions where either chaining or single pooling is the best. Single pooling leads to a better performance than chaining while being less costly under various situations of asymmetry: asymmetry in the number of arrivals, in the service durations, in the variability of service times, or in the service level requirements. It is also shown that these observations are more apparent for situations with a large number of skills, or for those with a large call center size.

2.1 Introduction

Context and Motivation. The concept of flexibility is related to the ability of a company to efficiently match its capacity to an uncertain demand with multiple types. The need for flexibility arises in a wide range of manufacturing systems. It also extends to service systems, such as call centers, where different types of customers ask for a quasi-instantaneous processing. Resource flexibility in call centers reduces to cross-training agents, which allows to improve both the utilization and the performance. Since cross-training agents is achieved with higher operating costs, resource flexibility could result in a trade-off between performance and cost. The performance is measured through operational indicators such as the expected waiting time, the probability of waiting, and the waiting time distribution, or also through human resource aspects that result in a higher efficiency of the agents. Cross-training may improve the agent motivation and provides a career path. In this chapter, we only focus on the operational indicators.

The process flexibility problem have been studied in different directions, such as machine sharing, multi-stage supply chains, queueing systems and flexible workforce scheduling. Here we consider flexibility questions in the context of queueing models for call centers. A wide literature has focused on contrasting two extreme situations. The *full flexible* architecture (FF) versus the the *full dedicated* (FD) one. In the FF model, each agent is fully cross-trained for all call types. In most situations in which call types have similar service duration requirements, FF would require less agents than any other architecture, in order to reach a given predefined service level. The reason is that it benefits from the economies of scale, which absorb stochastic variability (Borst et al. (2004)). However the agents in FF are too costly and even sometimes impossible to find. As commented by Marengo (2004), the multilingual Compaq call center certainly could not find or train agents to speak eleven languages! In the other extreme situation of the FD model, an agent is only trained to handle a single call type. Agents are then less costly, but FD would require a larger staffing level to reach the same service level as in FF or any other architecture.

Full flexibility and full dedication, however, are only two extreme situations. A well known and studied intermediate configuration is *chaining*, first pointed out by Jordan and Graves (1995). In the chaining model, each call type can be assigned to one of two adjacent agent teams, and each agent can handle calls from two adjacent types. Sheikhzadeh et al. (1998), Gurumurthi and Benjaafar (2004), and Jordan et al. (2004) prove that chaining, with an appropriate linkage between demand and resource types, behaves just as well as full flexibility. In the context of Constant Work in Process (CONWIP) serial production lines, Hopp et al. (2004) showed that the impact of forming a complete chain of skill sets can be substantial in increasing throughput. Wallace and Whitt (2005) consider the problem of routing and staffing in multi-skill call centers. They again confirm the principal that a little flexibility has the potential to achieve the performance of total flexibility. Using simulation they demonstrate that the performance, with an appropriate and limited cross-training of agents (two skills per agent) such as in chaining, is almost as good as when each agent has all skills.

Developing intelligent configurations such as chaining is very interesting for practitioners. The value of these configurations is that they capture the benefits of pooling by only having a limited flexibility. However, the robustness of chaining fails in the case of asymmetric demand (Sheikhzadeh et al. (1998)). By asymmetric demand, we mean different workload intensities and service time requirements, and also different variabilities in inter-arrival and service times. For such cases in practice, it is important to develop new architectures that allows from the one hand to account for demand asymmetry, and from the other hand to capture the benefits of pooling with only a limited flexibility.

In this chapter, we consider skill-based routing (SBR) call centers with two particular features: demand asymmetry and costly/difficult agent training. The typical example is that of an European multilingual call center where customers call from several countries. It is difficult for managers to find agents speaking more than two languages. For instance in Bluelink (the service provider of Air

France KLM), each agent speaks two languages: her own native language and English. Note that Bluelink is more interested in agents speaking two languages than those speaking three or more languages. The reason is that the latter often feel themselves over-qualified. They are therefore likely to leave the company faster than the others, which increases the turnover. The workload is also unbalanced ranging from only some few calls from a given country to several thousand of calls from another country. Another example is that of post-sales service call centers of retailers such as Darty and Fnac which are French distributors of white goods, telecommunications products, information technology, but also internet services, photo services or travel services. We also give the example of retail banking call centers where questions are with regard to savings or stock exchange for examples. The main characteristics in the previous examples are (i) the demand is unbalanced, (ii) the nature of the required agent skills can be very different which make difficult or too costly the agent training, and (iii) one may find a predominant and "easy" type of questions that could be handled by most of the agents without any particular training, for example the English task in a multilingual call center, account information and simple bank tasks in banking, order tracking and payment for retailers, etc.

Main findings. Motivated by this prevalence in practice, we propose in this chapter a new call center architecture that can be used instead of chaining. For such cases, applying chaining is too costly and difficult to implement (many combinations of two tasks per agent are even hard to obtain). Moreover, existing literature have shown that chaining is not appropriate for such demand situations: unbalanced workload of the "difficult" tasks and a predominant "easy task". As proven in Bassamboo et al. (2010), the tailored pairing architecture is efficient for small systems including those with asymmetries. However, this architecture requires an important number of cross-trained teams which might be again difficult to implement in practice. We propose a new organizational model, referred to as *single pooling*, where we dedicate a team of agents to each difficult type of calls, and the easy type of calls have access to all agents from all teams. Balancing the workload

among the agents in this way captures the benefits of pooling without requiring every agent to process every call type.

A concise definition of our model will be given later. We do not claim that our model is better than chaining in all cases, but only in the particular range of parameters as shown in the call center examples above. The value of our architecture is that it has a low degree of flexibility (each agent handles one difficult type and the easy task) while behaving in terms of performance as a fully flexible call center. This is important in practice since additional flexibility often comes at the cost of high operating overhead. Hence, the results of our analysis have significant managerial implications.

Using simulation, we conduct a comprehensive comparison between this novel architecture and chaining. As a function of the various system parameters, we delimit the regions where either chaining or single pooling is the best. Our key findings are highlighted next. Single pooling leads to better performance while being less costly than chaining under various situations of asymmetry between the customer types: asymmetry in the number of arrivals, in the service duration, in the variability of service times, or in the service level requirements. Moreover, we conclude that these observations are more apparent for situations with a large number of skills, or for those with a large call center size. In practice, the issue of limiting the flexibility appears more in large call centers, rather than in small ones with a few number of agents. In small call centers, the number of customer types is often very limited or they are very similar in terms of the required agent skills, so that the agents are usually full-flexible. Hence, there is often no need for managers to deal with cross-training questions. These insights show that there might be opportunities for managers of call centers to improve performance using the single pooling architecture.

The rest of the chapter is organized as follows. In Section 2.2 we review some of the literature related to this chapter. In Section 2.3 we describe chaining and single pooling models, and provide the comparison framework. In Section 2.4, we develop analytical results for particular single pooling

models, in order to get some sense on the effect of arrival asymmetry on performance. In Section 2.5, we use simulation to compare between the two call center models under various situations of asymmetry on the system parameters. Section 2.6 concludes the chapter and highlights some future research.

2.2 Literature Review

There is an extensive and growing literature on call centers. We refer the reader to Gans et al. (2003) and Akşin et al. (2007) for an overview. We review in what follows some of the literature related to this work.

Impact of Pooling. The value of pooling comes from the creation of flexibility. The general known intuition is that pooled systems are more effective than independent ones. The impact of pooling has been first studied in Smith and Whitt (1981). They show that pooling always leads to a better performance in terms of the expected delay in queue. Akşin and Karaesmen (2007) investigate the impact of the call center size on the opportunity to add flexibility. They demonstrate that a small call center will benefit more from adding flexibility than a large one.

Benjaafar (1995) studies the impact of pooling for a variety of manufacturing, telecommunication and computer systems. He considers a multi-processing system consisting of several facilities and shows that in some situations of heterogeneity in the workloads, increasing flexibility can deteriorate performance. Mandelbaum and Reiman (1998) consider stochastic service systems modeled as queueing networks. The service of a customer amounts to a collection of tasks. They show that adding flexibility does not automatically improve performance. They point out that adding a partial flexibility could be devastating for a queueing network. Recently, van Dijk and van Der Sluis (2008) show in the context of SBR call centers that without any clever routing rules and under a high variability in the call types and the resources, pooling could deteriorate the performance in terms of the average waiting time. In their work, Tekin et al. (2009) investigate the efficiency

benefits achievable via cross-training in SBR call centers. They conclude that under first come, first served (FCFS), the pooling of the dedicated teams is appropriate for heavy workloaded teams. However, it is not necessarily the case for teams with light workloads. They then use the difference between the arrival rates of the customers as a choice parameter based on which the decision of pooling would be taken or not. Inspired by the results of Smith and Whitt (1981), they also conclude that pooling teams could be counterproductive if services time means are very different from one customer type to another (for example when one is six times higher than the other ones).

Flexible Architectures. The most fundamental work on flexibility is that by Jordan and Graves (1995) for the automobile assembly plants, but it can be also applied to broader manufacturing system settings. They conduct an extensive simulation study and conclude that "a little flexibility can achieve almost all the benefits of total flexibility" under a configuration referred to as chaining, with two product types per plant. They demonstrate that the expected shortfall and capacity utilization of chaining resources are close to those under a full flexible configuration. Garavelli (2001) considers the setting of job shop cellular manufacturing systems used to perform batch production. He finds similar results to those by Jordan and Graves (1995) through a comparison between the performance of a full dedicated system, a full flexible one and chaining. Similar results are found by Garavelli (2003) in a complex supply chain environment, requiring the coordination of many plants producing good to customers located in different places. Again in the context of cellular manufacturing systems, Albino and Garavelli (1999) analyze the benefits of a limited flexibility. Starting from two industrial case studies concerning in-house metalworking shops, Nomden and van der Zee (2008) find by simulation that a chained distribution of routes behave very well.

For queueing systems, Gurumurthi and Benjaafar (2004) compare different scenarios of adding flexibility under different routing policies. They prove that the value of chaining decreases for an asymmetric demand. Hopp and van Oyen (2004) consider the question of how to cross-train a worker to two skills in the context of serial production lines. They conclude that a novel strategy called skill

chaining strategy is more robust against variability than a cherry-picking strategy (a team is full flexible) when demand is symmetric. The cherry-picking strategy in a serial production line can be seen as similar to single pooling, where the customers are the machines, and the bottleneck machine represents the easy type of calls. Tomlin and Wang (2005) consider the context of unreliable supply chains that produce multiple products. They study four canonical supply chain design strategies, where one of them, referred to as dual-source flexible, has been already proposed by Chevalier et al. (2004) in the context of call centers. They refine the prevailing intuition that a flexible network is preferable to a dedicated network by proving that this intuition is valid if either the resource investments are perfectly reliable or the firm is risk neutral. In a similar setting to ours, Robbins and Harrison (2010) introduce an SBR call center queueing model with two customer types, referred to as partial pooling. They consider two dedicated agent teams for each customer type, and one cross-trained team for both types. They show that cross-training a small number of agents can deliver a substantial benefit. They also find the level of cross-training that minimizes staffing costs, while satisfying a service level constraint. Bassamboo et al. (2010) study the flexibility problem with a newsvendor network model of resource portfolio investment. They conduct a comparison between chaining and tailored pairing. They show that a system that combines dedicated and cross-trained agents is asymptotically optimal. They also show using simulation experiments that the "tailored pairing" design is superior for small systems, including systems with asymmetries. The tailored pairing architecture might be one of the best propositions in the literature to deal with the asymmetric parameters. However as already mentioned above, we can not retain this architecture as a reference in our project. The reason is that for call centers with many skills, working under tailored pairing may lead to non-realizable situations. In such a case, the number of two-skills combinations could be very high.

Garnett and Mandelbaum (2001) argue on the importance of adapting the system architecture to the asymmetry in the customer arrival rates. In summary, chaining is robust according to its

ability to support variability. It however fails when demand is asymmetric. It can be also too expensive to train agents on various combinations of two skills. For these situations, we propose and analyze in this project a new efficient configuration of a queueing call center model.

Agent Skills, Staffing and Routing. In an SBR call center, agents can often only be trained for a subset of skills. One key management issue is to determine the subset of skills that will be considered, and the number of agents for each subset of skills. Pinker and Shumsky (2000) build a learning model where the quality of service is related to the employee experience. According to their model, the benefits of flexibility are not guaranteed. It is true that a flexible agent can treat more customers, but the quality of service would not be as good as it would be with a dedicated agent. They also compare between different system sizes and show that specialization is preferred in large systems and complete pooling is preferred in small systems. For medium size systems, a mix of flexibility and specialization would be appropriate. In a call center context, Wallace and Whitt (2005) conclude using an extensive simulation study that, when you add skills to an agent, most of the benefits is taken going from one skill per agent to two skills per agent. These results tend to support the idea of limiting the number of skills per agent.

As for the problems of staffing and routing, we refer the reader to the survey by Gans et al. (2003), where the authors present the square-root staffing rule. Borst et al. (2004) revisited the square-root rule by including principles of routing based on agent costs. To optimize the staffing level in an SBR call center, Henderson and Mason (1998) combine simulation and integer programming with cutting plane methods. Atlason et al. (2008) provide interesting properties of a cutting plane method for staffing and prove that it outperforms traditional staffing heuristics which are based on analytical queueing methods. Some other works have investigated the impact of the type of the agent contract on the required staffing level in order to reach a given service level: Ren and Zhou (2008) consider piece-meal and pay-per-call-resolved contracts and propose other contracts that coordinate both staffing and effort.

Borst and Seri (2000) present a routing heuristic that assigns customers to the available agent with the most specialized set of skills which is a generalization of the "specialist-first" principle. Chevalier et al. (2004) show that in terms of performance a 20/80 model (20% of generalists and 80% of specialists) performs almost as good as a full flexible model, moreover, it has lower operating costs. In a manufacturing setting, Sheikhzadeh et al. (1998) evaluate the difference in performance between strict priority, longest queue first and random priority in a comparison between chaining and full flexibility. They remark that the longest queue first policy is the best one. In order to perform a coherent comparison between chaining and our model, we first choose the appropriate routing rules and also staffing levels.

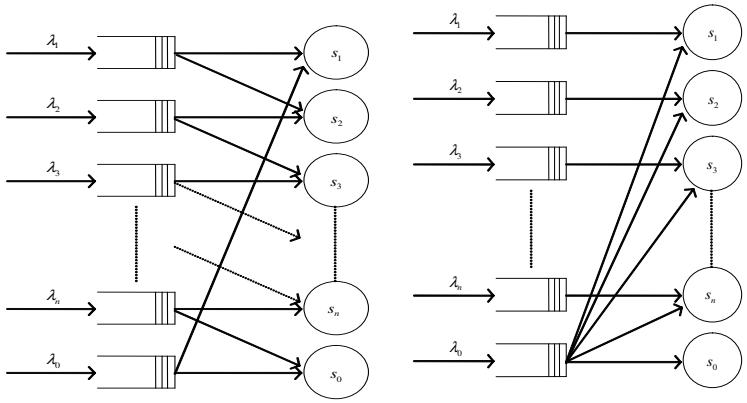
2.3 Problem Setting

We consider call center models with $n+1$ call types (types 0, 1, ..., n). Customer types 1, 2, ..., n, referred to as also regular types are those requiring specific agent skills 1, 2, ..., n, respectively, while customers 0 can be handled by any agent without a particular "sophisticated" training as required for the regular types. In other words, skill 0 is an easy skill. The mean arrival, service and abandonment rates of customers type i are λ_i, μ_i and γ_i, respectively ($i = 0, 1, ..., n$). The agents are organized in homogeneous teams, i.e., all agents from a given team have the same set of skills. In this chapter we only consider agent teams with at most two skills per agent. We define an economic framework as follows. We assume that skill 0 costs 1, and that skill i costs $1+t_i$ (for $i = 1, \cdots, n$). For two skills i and j, the cost is $1+t_{i,j}$ (for $i, j \in \{0, \cdots, n\}$). Since skill 0 is the easy skill, we assume that $t_{i,0} \leq t_{i,j}$ (for $i, j \in \{0, \cdots, n\}$).

We are interested in the performance in terms of the expected waiting time in the queue of each customer type i taken in service, denoted by W_i, for $i = 0$, 1, ..., n. We denote the objective service level for a type i by W_i^*, for $i = 0$, 1, ..., n. In what follows, we describe the two models that we compare in this chapter: chaining and single pooling. We do not consider the architectures

developed by Borst and Seri (2000), and Bassamboo et al. (2010), because in our context they are too costly or even very hard to implement by a call center manager due to the specificity of the agent skills.

Chaining and single pooling models are shown in Figures 2.1(a) and 2.1(b), respectively. In the chaining model, the skills of the teams are such that they form a chain. The value of the well known chaining model comes from its capability to smooth the workload over the agent teams. In the single pooling model, customers type 0 benefit from a complete pooling, whereas the other types have only access to one dedicated team. The value of this architecture is that it allows to appropriately handle situations of asymmetry in demand as we will show later. Single pooling can be seen as a dual of the architecture proposed by Chevalier et al. (2004), with dedicated single skill agent teams and one team of agents with all skills.



(a) Chaining (b) Single pooling

Figure 2.1: Call center configurations

The functioning we consider for chaining and single pooling is intuitive and easy to implement in practice. Under chaining, a customer upon her arrival has access to agents from two teams. If at least an agent is available in one of them, then the customer is routed to the team with the higher proportion of idle agents (number of idle agents in a team over the total number of agents in that team). If this proportion is the same for the two teams, then she is equiprobably routed to one of the two teams. Otherwise if all agents from the two teams are busy upon her arrival, the customer

waits in her queue (each customer type has its own queue). An agent can handle customers from two queues. Within each queue, the discipline of service is FCFS. When an agent becomes idle, she selects to service one of the customers that are waiting in the two queues, if any. The priority is given to the customer with the longest waiting time.

For single pooling, the routing rules are as shown in Figure 2.1(b). The discipline of service in each one of the $n+1$ queues is FCFS. A customer type i (for $i = 1, ...n$) can be served by only an agent from its associated team. A customer type 0 however can be served by any agent of any one of the $n+1$ teams. Upon arrival, a customer type 0 is in priority handled by an idle agent from team 0, if any. If not, she is handled by an idle agent from one of the teams of the regular customer types, if any. If more than one of those have at least one idle agent, then customer 0 is routed to the team with the higher proportion of idle agents. If many teams have the same highest proportion, then customer 0 is equiprobably routed to one of these teams. If all agents of all teams are busy, then customer 0 is placed in her queue. When an agent from one of the teams of the regular customers becomes free, it can serve either a regular customer or a customer 0. However a regular customer has a non-preemptive priority over a customer 0. This means that the idle agent deals with a call from her regular queue first (the first in line). If the queue of the associated regular type is empty, this agent provides service to a customer 0 (the first in line). We assume in our models that the queues are infinite.

In this project, we compare between the two models chaining and single pooling through simulations. In order to have a coherent comparison we optimize their total staffing cost under the constraints $W_i \leq W_i^*$, for $i = 0, 1, ..., n$. We use greedy heuristics for the simulation based optimization step. We refer the reader to the details in Section A.1 of the appendix. For the staffing optimization of SP, we use an increasing greedy algorithm. Starting from an under-staffed situation (a full dedicated model with customers 0), we increase step by step the arrival rate of customers 0. In each iteration, we increment the number of agents in the various teams such that we strictly

reach the service level constraints. For chaining, we develop a decreasing greedy algorithm. The algorithm starts with an over-staffed situation using a full dedicated model, which is the worst for chaining since it ignores the links between the teams. We then use an efficient method suggested by Wallace and Whitt (2005) in order to correct the staffing levels to the chaining setting.

2.4 Particular Single Pooling Cases

The analytical analysis of the general case of single pooling is too complex. We consider in this section two particular Markovian cases of single pooling, for which, we develop exact and approximate results. The objective of this analysis is to obtain some sense on the effect of the parameters asymmetry on performance. A more comprehensive analysis of the effect of asymmetry is then conducted in Section 2.5 using simulation.

2.4.1 Three Customer Types

Consider a single pooling model with three customer types 0, 1 and 2. The arrival process of types 0, 1 and 2 is Poisson with rates λ_0, λ_1 and λ_2, respectively. There two agent teams 1 and 2 with sizes s_1 and s_2, respectively. The service rate, denoted by μ, is identical for all customer types. Using a Markov chain approach, we compute in what follows the expected waiting times for all customer types.

Let us define the stochastic process $\{(x(t), y(t), z(t))t \geq 0\}$, where $x(t)$ and $y(t)$ denote the number of busy agents in team i plus the number of waiting customers in queue i ($i = 1, 2$), and $z(t)$ denotes the number of waiting customers in queue 0, for an instant $t \geq 0$. Since inter-arrival and service times are Markovian, $\{(x(t), y(t), z(t))t \geq 0\}$ is a Markov chain. Let us denote the system steady-state probabilities by $\pi_{x,y,z}$, for x, y, $z \in \mathbb{N}$. Note that we can have $z \geq 1$ only when $x \geq s_1$ and $y \geq s_2$. From the Markov chain, one may write the following set of equations. We have $(\lambda_1 + \lambda_2 + \lambda_0)\pi_{0,0,0} = \mu(\pi_{1,0,0} + \pi_{0,1,0})$. For $x = 0$ and $y > 0$, $(\lambda_1 + \lambda_2 + \lambda_0)\pi_{0,y,0} =$

$\min(y+1,s_2)\mu\pi_{0,y+1,0}+\mu\pi_{1,y,0}+\lambda_2\pi_{0,y-1,0}$. For $x>0$ and $y=0$, $(\lambda_1+\lambda_2+\lambda_0)\pi_{x,0,0}=\min(x+1,s_1)\mu\pi_{x+1,0,0}+\mu\pi_{x,1,0}+\lambda_1\pi_{x-1,y,0}$. For $0<x<s_1$ or $0<y<s_2$, there are 5 cases: If $y-1>\frac{s_2}{s_1}x$, $(\lambda_1+\lambda_2+\lambda_0+\min(x,s_1)\mu+\min(y,s_2)\mu)\pi_{x,y,0}=\min(x+1,s_1)\mu\pi_{x+1,y,0}+\min(y+1,s_2)\mu\pi_{x,y+1,0}+(\lambda_1+\lambda_0)\pi_{x-1,y,0}+\lambda_2\pi_{x,y-1,0}$. If $y-1=\frac{s_2}{s_1}x$, $(\lambda_1+\lambda_2+\lambda_0+\min(x,s_1)\mu+\min(y,s_2)\mu)\pi_{x,y,0}=\min(x+1,s_1)\mu\pi_{x+1,y,0}+\min(y+1,s_2)\mu\pi_{x,y+1,0}+(\lambda_1+\lambda_0)\pi_{x-1,y,0}+(\lambda_2+\lambda_0/2)\pi_{x,y-1,0}$. If $y-1<\frac{s_2}{s_1}x$ and $y>\frac{s_2}{s_1}(x-1)$, $(\lambda_1+\lambda_2+\lambda_0+\min(x,s_1)\mu+\min(y,s_2)\mu)\pi_{x,y,0}=\min(x+1,s_1)\mu\pi_{x+1,y,0}+\min(y+1,s_2)\mu\pi_{x,y+1,0}+(\lambda_1+\lambda_0)\pi_{x-1,y,0}+(\lambda_2+\lambda_0)\pi_{x,y-1,0}$. If $y=\frac{s_2}{s_1}(x-1)$, $(\lambda_1+\lambda_2+\lambda_0+\min(x,s_1)\mu+\min(y,s_2)\mu)\pi_{x,y,0}=\min(x+1,s_1)\mu\pi_{x+1,y,0}+\min(y+1,s_2)\mu\pi_{x,y+1,0}+(\lambda_1+\lambda_0/2)\pi_{x-1,y,0}+(\lambda_2+\lambda_0)\pi_{x,y-1,0}$. If $y<\frac{s_2}{s_1}(x-1)$, $(\lambda_1+\lambda_2+\lambda_0+\min(x,s_1)\mu+\min(y,s_2)\mu)\pi_{x,y,0}=\min(x+1,s_1)\mu\pi_{x+1,y,0}+\min(y+1,s_2)\mu\pi_{x,y+1,0}+\lambda_1\pi_{x-1,y,0}+(\lambda_2+\lambda_0)\pi_{x,y-1,0}$. For $x=s_1$, $y=s_2$ and $z=0$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{s_1,s_2,0}=s_1\mu\pi_{s_1+1,s_2,0}+s_2\mu\pi_{s_1,s_2+1,0}+(\lambda_1+\lambda_0)\pi_{s_1-1,s_2,0}+(\lambda_2+\lambda_0)\pi_{s_1,s_2-1,0}+(s_1+s_2)\mu\pi_{s_1,s_2,1}$. For $x=s_1$ and $y>s_2$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{s_1,y,0}=s_1\mu\pi_{s_1+1,y,0}+s_2\mu\pi_{s_1,y+1,0}+(\lambda_1+\lambda_0)\pi_{s_1-1,y,0}+\lambda_2\pi_{s_1,y-1,0}+s_1\mu\pi_{s_1,y,1}$. For $x>s_1$ and $y=s_2$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{x,s_2,0}=s_1\mu\pi_{x+1,s_2,0}+s_2\mu\pi_{x,s_2+1,0}+\lambda_1\pi_{x-1,s_2,0}+(\lambda_2+\lambda_0)\pi_{x,s_2-1,0}+s_2\mu\pi_{x,s_2,1}$. For $x>s_1$, $y>s_2$ and $z=0$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{x,y,0}=s_1\mu\pi_{x+1,y,0}+s_2\mu\pi_{x,y+1,0}+\lambda_1\pi_{x-1,y,0}+\lambda_2\pi_{x,y-1,0}$. For $z>0$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{s_1,s_2,z}=s_1\mu\pi_{s_1+1,s_2,z}+s_2\mu\pi_{s_1,s_2+1,z}+\lambda_0\pi_{s_1,s_2,z-1}+(s_1+s_2)\mu\pi_{s_1,s_2,z+1}$. For $x=s_1$, $y>s_2$ and $z>0$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{s_1,y,z}=s_1\mu\pi_{s_1+1,y,z}+s_2\mu\pi_{s_1,y+1,z}+\lambda_0\pi_{s_1,y,z-1}+s_1\mu\pi_{s_1,y,z+1}$. For $x>s_1$, $y=s_2$ and $z>0$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{x,s_2,z}=s_1\mu\pi_{x+1,s_2,z}+s_2\mu\pi_{x,s_2+1,z}+\lambda_0\pi_{x,s_2,z-1}+s_2\mu\pi_{x,s_2,z+1}$. For $x>s_1$, $y>s_2$ and $z>0$, $(\lambda_1+\lambda_2+\lambda_0+(s_1+s_2)\mu)\pi_{x,y,z}=s_1\mu\pi_{x+1,y,z}+s_2\mu\pi_{x,y+1,z}+\lambda_0\pi_{x,y,z-1}$.

One may intuitively see from the Markov chain how the asymmetry in arrivals increases the performance of single pooling. The counterproductive states are those with waiting customers and idle agents at the same time, i.e., $x>s_1$ and $0\le y<s_2$, or $y>s_2$ and $0\le x<s_1$. When being in one of these two "bad" cases, the probabilities to take the direction of leaving them are $\frac{\lambda_0+\lambda_2+s_1\mu}{\lambda_1+\lambda_2+\lambda_0+(s_1+y)\mu}$ and $\frac{\lambda_0+\lambda_1+s_2\mu}{\lambda_1+\lambda_2+\lambda_0+(s_1+y)\mu}$, respectively. This shows for example that increasing the

proportion of customers 0 (one form of asymmetry) increases the performance in single pooling.

A further illustration is given next. The expected waiting times as a function of the steady-state probabilities are given by

$$W_1 = \frac{1}{\lambda_1}\left(\sum_{x=s_1}^{+\infty}\sum_{y=0}^{+\infty}(x-s_1)\pi_{x,y,0} + \sum_{x=s_1}^{+\infty}\sum_{y=s_2}^{+\infty}\sum_{z=1}^{+\infty}(x-s_1)\pi_{x,y,z}\right),$$
$$W_2 = \frac{1}{\lambda_2}\left(\sum_{y=s_2}^{+\infty}\sum_{x=0}^{+\infty}(y-s_2)\pi_{x,y,0} + \sum_{y=s_2}^{+\infty}\sum_{x=s_1}^{+\infty}\sum_{z=1}^{+\infty}(y-s_2)\pi_{x,y,z}\right),$$
$$W_0 = \frac{1}{\lambda_0}\sum_{z=0}^{+\infty}\sum_{x=s_1}^{+\infty}\sum_{y=s_2}^{+\infty} z\pi_{x,y,z}.$$

The performance measures above are computed numerically. We solve the steady-state equations relating the state probabilities using a state space truncation, with a sufficiently high precision (six digits beyond the decimal point). Let us now denote by p the proportion of customers 0 among all arriving customers, $p = \dfrac{\lambda_0}{\sum_{i=0}^{n}\lambda_i}$. Figure 2.2 shows how W_1 and W_0 considerably improve in p.

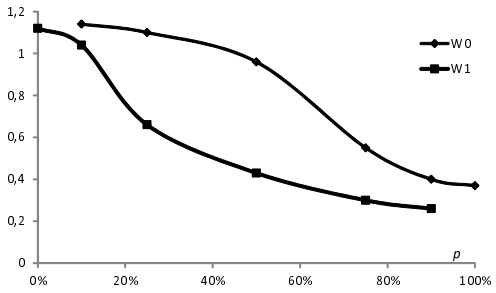


Figure 2.2: Impact of p on SP performance ($\mu_1 = \mu_2 = \mu_0 = 0.2$, $\sum_{i=0}^{2}\lambda_i = 4$, $\lambda_1 = \lambda_2$, $s_1 = s_2 = 12$)

2.4.2 A Fixed Point Approximation

We consider here a Markovian single pooling case with an arbitrary number of skills. There are $n+1$ customer types (type 0, and types $1, 2, ..., n$), n teams (no team 0), $n \geq 1$. The arrival rates are λ_0 and $\lambda_i = \lambda$ for $i = 1, ..., n$, and the service rates are $\mu_i = \mu$ for $i = 0, 1, ..., n$. Since the configuration is symmetric, we consider the same staffing level s in each team. In what follows, we

develop an approximation to compute the expected waiting time of regular customers type i, for $i = 1, ..., n$. The approximation is based on a Markov chain approach and a fixed point algorithm.

One can see that our model can be divided into n identical sub-systems. It suffices then to focus on the performance analysis of one of these sub-systems. A sub-system is a simple queueing system with s servers and an infinite queue. Two types of customers arrive to this sub-system: customers type i with a Poisson process with rate λ and customers type 0 with a general arrival process with mean arrival rate $\frac{\lambda_0}{n}$. (The arrival process of customers 0 to the whole system is Poisson. However, it becomes a general process at each sub-system because of the routing rules.)

Recall that customers 0 wait in their own queue before being routed to one of the sub-systems for an immediate processing. Because of the routing rule, customers 0 can be routed to a sub-system only if the number of customers in the sub-system is less or equal to $s - 1$. Also since we route customers 0 to the one of the less busiest sub-systems (with an equiprobable choice), the arrival rate of customers 0 is decreasing in the number of busy servers in a sub-system and it becomes 0 when all the s servers become busy.

Let us now define, for a sub-system, the stochastic process $\{E(t), t \geq 0\}$, where $E(t)$ denotes the number of customers in the system (queue + service). Note that the customers in the queue are only the regular customers, and those in service can be both regular or type 0 customers. We approximate customers 0 inter-arrival times by an exponential distribution with state-dependent rates. Since inter-arrival and service times are Markovian, $\{E(t), t \geq 0\}$ is a Markov chain as shown in Figure 2.3. The arrival rate δ_k denotes the state-dependent arrival rate of customers 0 when the number of customers in the sub-system is k, for $k = 0, ..., s - 1$ (no customers 0 arrive at the sub-system for $k \geq s$).

Assume that exactly s customers are in the sub-system and that a service completion occurs first before the next arrival epoch of a regular customer at this sub-system (Figure 2.3). Therefore, two possibilities may happen. The first possibility corresponds to the case of an empty queue 0.

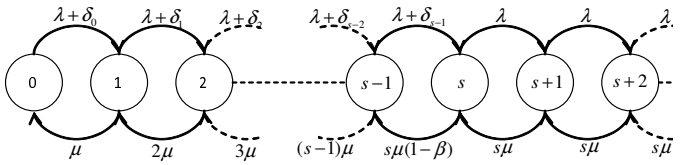


Figure 2.3: Markov chain associated to a sub-system of single pooling

We then move to state $s-1$. The second one corresponds to the case of a non-empty queue 0. We then stay in state s, because the server who just became idle immediately takes the customer 0 in the head of queue 0 into service. Let us denote by β the probability that queue 0 is not empty. Then the rate to move from state s to sate $s-1$ in the Markov chain is $s\mu(1-\beta)$.

Let us now assume that the stability condition of a sub-system holds, i.e., $\lambda + \frac{\lambda_0}{n} < s\mu$, and denote the stationary probabilities of the system states by π_k, for $k \geq 0$. We may then write

$$\pi_k = \frac{\prod_{i=0}^{k-1}(\lambda+\delta_i)}{k!\mu^k}\pi_0, \tag{2.1}$$

for $1 \leq k \leq s-1$, and

$$\pi_{s+k} = (\frac{\lambda}{s\mu})^k \frac{\prod_{i=0}^{s-1}(\lambda+\delta_i)}{s!\mu^s(1-\beta)}\pi_0, \tag{2.2}$$

for $k \geq 0$. Since all probabilities sum up to one, we obtain

$$\pi_0 = \frac{1}{1+\sum_{k=1}^{s-1}\frac{\prod_{i=0}^{k-1}(\lambda+\delta_i)}{k!\mu^k} + \frac{\prod_{i=0}^{s-1}(\lambda+\delta_i)}{s!\mu^s(1-\beta)}\frac{1}{1-\frac{\lambda}{s\mu}}}. \tag{2.3}$$

The difficulty to compute the stationary probabilities is that we do not have the values of δ_k $(k = 0, ..., s-1)$ and β. We use a fixed point algorithm to jointly compute them with the stationary probabilities. Let us now write δ_0, the arrival rate of customers 0 at a given sub-system when this sub-system is empty, as a function of the stationary probabilities of this sub-system. We use here a second approximation. We indeed assume that the states of the sub-systems are independent,

which is not true. Assume that our sub-system is the only one that is empty, i.e., each one of the other $n-1$ sub-systems have at least one customer in the system (queue + service). Using the approximation this occurs with probability $(1-\pi_0)^{n-1}$, then δ_0 is simply λ_0 in that case. Assume now that our sub-system and only another one are empty. Then δ_0 is $\frac{\lambda_0}{2}$ (equiprobable routing of customers 0 to one of the less busiest sub-systems). This occurs with probability $\pi_0(1-\pi_0)^{n-2}$ and there are $\binom{n-1}{1}$ combinations (where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ for $0 \leq k \leq n$). Continuing with the same reasoning and averaging over all possibilities, we obtain

$$\delta_0 = \lambda_0 \sum_{j=0}^{n-1} \frac{1}{j+1} \binom{n-1}{j} \pi_0^j (1-\pi_0)^{n-1-j}. \tag{2.4}$$

Since $\frac{1}{j+1}\binom{n-1}{j} = \frac{1}{n}\binom{n}{j+1}$, Equation (2.4) becomes

$$\delta_0 = \frac{\lambda_0}{n} \sum_{j=0}^{n-1} \binom{n}{j+1} \pi_0^j (1-\pi_0)^{n-1-j} = \frac{\lambda_0}{n\pi_0} \sum_{j=1}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j},$$

which leads to

$$\delta_0 = \lambda_0 \frac{1-(1-\pi_0)^n}{n\pi_0}. \tag{2.5}$$

In the same way, we obtain

$$\delta_k = \lambda_0 \frac{\left(1-\sum_{j=1}^{k-1}\pi_j\right)^n - \left(1-\sum_{j=1}^{k}\pi_j\right)^n}{n\pi_k}, \tag{2.6}$$

for $1 \leq k \leq s-1$. Let us now give the expression of β as a function of the stationary probabilities π_k, $k \geq 0$. Since the mean arrival rate of customers 0 at our sub-system is $\frac{\lambda_0}{n}$, we have

$$\sum_{k=0}^{s-1} \pi_k \delta_k + \beta s \mu = \frac{\lambda_0}{n}, \tag{2.7}$$

which implies

$$\beta = \frac{\frac{\lambda_0}{n} - \sum_{k=0}^{s-1} \pi_k \delta_k}{s\mu}. \tag{2.8}$$

In summary, from the one hand, Equations (2.1)-(2.3) give the stationary probabilities π_k ($k \geq 0$) as a function of δ_k ($0 \leq k \leq s-1$) and β. From the other hand, Equations (2.5), (2.6) and (2.8) give δ_k ($0 \leq k \leq s-1$) and β as a function of π_k ($k \geq 0$). As a consequence, we have a fixed point. We propose the following fixed point algorithm to compute it. In the first iteration, we choose $\delta_0 = \frac{\lambda_0}{n}$, $\delta_k = 0$ for $1 \leq k \leq s-1$, and $\beta = 0$. Then we compute π_k ($k \geq 0$) using Equations (2.1)-(2.3). From these π_k, we next compute the new values of δ_k ($0 \leq k \leq s-1$) and β using Equations (2.5), (2.6) and (2.8). In the second iteration, we use the latter values of δ_k and β to compute π_k. From these new π_k, we compute the new values of δ_k and β. We do the same in the third iteration, and so on. We stop the algorithm when the values of π_k ($k \geq 0$), δ_k ($0 \leq k \leq s-1$) and β converge to their limits with a given predefined precision (we have chosen a precision of 10^{-6} in the numerical experiments below). Proposition 1 proves the convergence of the fixed point algorithm.

Proposition 1 *The fixed point algorithm always converges.*

Proof. We use the Brouwer's theorem to prove the convergence. The Brouwer's theorem states that any continuous function from a convex compact subset K of an Euclidean space to itself has at least one fixed point. In what follows, we prove that the conditions of the Brouwer's theorem hold in our context.

After k iterations, the fixed point algorithm gives the vector $(\pi_0, \pi_1, \pi_2, \cdots, \pi_c)_k$ belonging to a convex compact, $[0;1]^{s+1}$, that is included in an Euclidean space, $\mathbb{R}^{s+1}$. From Equations (2.1)-(2.8), it is obvious to see that the function that allows to calculate $(\pi_0, \pi_1, \pi_2, \cdots, \pi_c)_{k+1}$ (iteration $k+1$) as a function of $(\pi_0, \pi_1, \pi_2, \cdots, \pi_c)_k$ is continuous (combination of continuous functions),

for $\pi_k \neq 0$ $(k = 0, ..., s-1)$. In what follows, we prove that this function is continuous in $\pi_k = 0$ $(k = 0, ..., s-1)$ by prolongation. From Equations (2.5) and (2.6), we have

$$\delta_k = \lambda_0 \frac{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n - \left(1 - \sum_{j=1}^{k} \pi_j\right)^n}{n\pi_k} = \lambda_0 \frac{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n}{n\pi_k} \left(1 - \frac{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n}{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n}\right),$$

for $k = 0, ..., s-1$, where by convention an empty sum is equal to 0. Calculating further, we obtain

$$\delta_k = \lambda_0 \frac{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n}{n\pi_k} \left(1 - \left(1 - \frac{\pi_k}{1 - \sum_{j=1}^{k-1} \pi_j}\right)^n\right),$$

for $k = 0, ..., s-1$. The Taylor expansion of δ_k as a function of π_k in the neighborhood of 0 is

$$\delta_k = \lambda_0 \frac{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n}{n\pi_k} \left(1 - (1 - n\, o(\pi_k))\right) = \lambda_0 \left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n + o(1),$$

where $o(1)$ is a function that converges to a finite limit as π_k goes to 0, for $k = 0, ..., s-1$. Since $\lambda_0 \left(1 - \sum_{j=1}^{k-1} \pi_j\right)^n$ is finite, δ_k is continuous by prolongation in $\pi_k = 0$, for $k = 0, ..., s-1$.

It remains now to focus on the issue for $\beta = 1$ in Equation (2.3). This case of $\beta = 1$ can not happen. The proof is as follows. Assume that $\beta = 1$. Equation (2.7) thus leads to $s\mu = \frac{\lambda_0}{n} - \sum_{k=0}^{s-1} \pi_k \delta_k$. Since δ_k and π_k $(0 \leq k \leq s-1)$ are positive, $s\mu \leq \frac{\lambda_0}{n}$. As a consequence the sub-system is unstable, which is absurd. This completes the proof of the convergence of the fixed point algorithm. □

Having in hand the stationary probabilities, we next compute for the regular customers the expected waiting time in the queue and the probability of delay. Recall that all sub-systems are identical because of the symmetry in the parameters. Using Little's law, the expected waiting time of a regular customer type i $(i = 1, ..., n)$ is given by

$$W_i = \frac{1}{\lambda} \sum_{k=1}^{\infty} k\pi_{s+k}, \tag{2.9}$$

Table 2.1: Fixed point approximation, $\mu = 0.2$

	λ	λ_0	s	$\frac{(\lambda+\lambda_0)/n}{s\mu}$	W_i		P_D	
					Simulation	Approximation	Simulation	Approximation
$n=1$	0.35	0.35	5	70%	0.581	0.581	37.78%	37.78%
	0.475	0.475	5	95%	1.672	1.672	87.78%	87.78%
	1.4	1.4	20	70%	0.035	0.035	9.36%	9.36%
	1.9	1.9	20	95%	0.359	0.359	75.54%	75.54%
	3.8	0	20	95%	3.777	3.777	75.54%	75.54%
$n=2$	0.35	0.7	5	70%	0.436	0.435	29.74%	28.28%
	0.475	0.95	5	95%	1.623	1.623	87.51%	85.23%
	1.4	2.8	20	70%	0.003	0.0027	0.74%	0.72%
	1.9	3.8	20	95%	0.320	0.290	61.03%	60.98%
	3.8	0	20	95%	3.777	3.777	75.54%	75.54%
$n=5$	0.35	1.75	5	70%	0.290	0.288	19.06%	18.76%
	0.475	2.375	5	95%	1.556	1.550	81.84%	81.38%
	1.4	7	20	70%	0.001	0.001	0.28%	0.27%
	1.9	9.5	20	95%	0.205	0.204	42.99%	42.97%
	3.8	0	20	95%	3.777	3.777	75.54%	75.54%
$n=10$	0.35	3.5	5	70%	0.259	0.252	17.21%	16.39%
	0.475	4.75	5	95%	1.516	1.504	79.07%	78.96%
	1.4	14	20	70%	0.0001	0.0001	0.23%	0.23%
	1.9	19	20	95%	0.167	0.167	35.23%	35.21%
	3.8	0	20	95%	3.777	3.777	75.54%	75.54%

for $i = 1, ..., n$, and its probability of delay denoted by $P_{D,i}$ is

$$P_{D,i} = \sum_{k=1}^{\infty} \pi_{s+k}, \tag{2.10}$$

for $i = 1, ..., n$. The approximation for both W_i and $P_{D,i}$ works very well for the regular customer types, however it does not for customers 0 because of their complex routing. The comparison between the approximate results using the fixed point algorithm and the exact ones using simulation are given in Table 2.1. Note that in the extreme situations of $n = 1$ or $\lambda_0 = 0$, our method gives the exact results.

Table 2.1 reveals that our approximation yields very accurate estimates, while slightly over-estimating the performance. It gives lower values for W_i and $P_{D,i}$ than those from simulation. An explanation would be as follows. In our approximation, we assume that all sub-systems are independent one of another. In reality, the routing rule leads to a fair sharing of customers between the sub-systems. Therefore, when a given sub-system is almost busy, the other ones are likely to be almost busy. Thus, the arrival rates δ_k (for high values of k close to $s - 1$) should be in reality

higher than those in the approximation, which implies that the latter would give better performance (lower waiting and lower probability of delay) than simulation does.

Using the above approximate analysis, we illustrate in Figure 2.4 how the asymmetry in arrivals (by increasing p) improves performance (expected waiting time W_i for $i = 1, ..., n$).

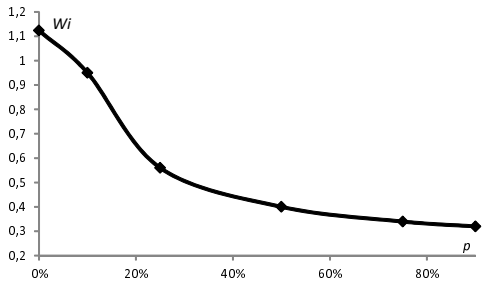


Figure 2.4: Impact of p on SP performance ($n = 4$, $\mu_i = \mu_0 = 0.2$, $\sum_{i=0}^{2} \lambda_i = 8$, $\lambda_i = \lambda_j$, $s_i = 12$ for $i, j = 1, ..., 4$)

2.5 Effect of the Parameters Asymmetry

In this section, we present the results of the comparison between chaining and single pooling. We use simulation experiments to optimize the call center staffing. In using simulation for call center operations management, we are following longstanding practice, see for example Wallace and Whitt (2005).

We simplify the cost model such that the SP cost is upper bounded and that of chaining is lower bounded. All the numerical comparisons are based on the lower and upper bounds values. This makes the results pessimistic for SP and optimistic for chaining, i.e., the performance of SP is in reality better than what we present. The cost of single pooling is $\sum_{i=0}^{n}(1+t_{i,0})s_i$. This is upper bounded by $(\sum_{i=0}^{n} s_i)\max_i(1+t_{i,0})$. The cost of chaining is $(1+t_{0,1})s_0+(1+t_{1,2})s_1+\cdots(1+t_{n,0})s_n$ and is lower bounded by $(1+t_{0,1})s_0 + (1+\min_{i,j}(1+t_{i,j}))(\sum_{i=1}^{n-1} s_i) + (1+t_{n,0})s_n$. Let us now simplify the problem as follows. An agent with skills 0 and i ($i = 1, ..., n$) costs 1. An agent with skills i and j ($i, j = 1, ..., n$ and $i \neq j$) costs $1 + t$, $t \geq 0$. In this simplification, we have

$\max_i(1+t_{i,0}) = 1$ and $\min_{i,j}(1+t_{i,j}) = 1+t$ $(i,j = 1,...,n$ and $i \neq j)$. The parameter t is then the incremental cost of an agent with two regular skills compared to that with a regular skill and skill 0.

Design of Experiments. As we are interested in the effect of asymmetry of the parameters on performance, we propose various forms of asymmetry. For customers 0, we define the parameters p and p' to measure the relative importance in arrivals and service durations, respectively. They are given by $p = \frac{\lambda_0}{\sum_{i=0}^{n}\lambda_i}$ and $p' = \frac{\frac{1}{\mu_0}}{\sum_{i=0}^{n}\frac{1}{\mu_i}}$. We measure the asymmetry between the arrival rates of regular customers by $V = \frac{\lambda_1}{\lambda_2} = \frac{\lambda_2}{\lambda_3} = \cdots = \frac{\lambda_{n-1}}{\lambda_n}$, and that between service durations by $U = \frac{1/\mu_1}{1/\mu_2} = \frac{1/\mu_2}{1/\mu_3} = \cdots = \frac{1/\mu_{n-1}}{1/\mu_n}$. We also consider for customers 0 the asymmetry in the variability of service times, measured by the coefficient of variation of its distribution and denoted by cv_s. We consider other forms of asymmetry in terms of the required service level and also the time to abandon for customers 0 relatively to those for the regular customers. These effects are studied in the settings of small and large call centers, and also in the settings of small and large number of skills. Although the considered forms of asymmetries do not cover all possibilities, they allow to obtain the main useful conclusions.

The approach to conduct the simulation experiments is as follows. Due to the high number of parameters, we first run experiments by separately treating one parameter at a time. In a systematic way, we vary one parameter while holding all the others constant. Second to see the possible interaction effects, we simultaneously vary the values of more than one of them at a time. For the values of the parameters, we choose wide ranges that allow to cover most of call center situations in practice. For the rest of the chapter, inter-arrival are assumed to be Markovian. Service times are also assumed to be Markovian, except in Section 2.5.2. The abandonment rates are assumed to be equal to zero, except for Section 2.5.4.

2.5.1 Asymmetry in Arrival Rates

We want to understand the effect of the asymmetry in the demand. We separate the study into two steps. First, we construct the asymmetry only on the arrival rate of customers 0. Second, we construct it by differentiating between all the arrival rates of all customer types.

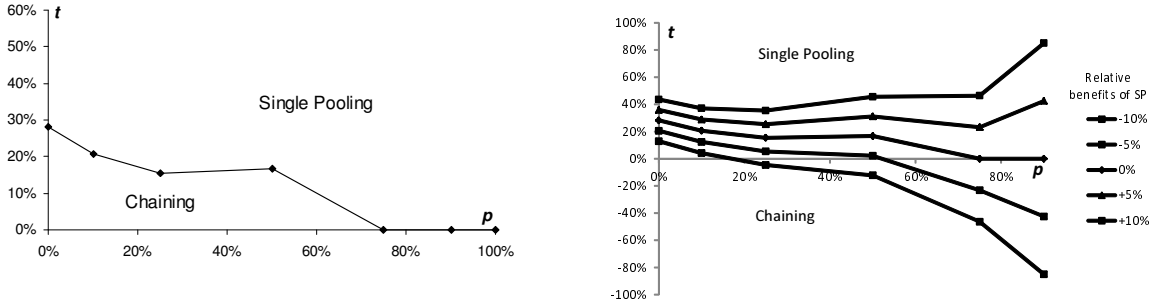
Asymmetry on Customers 0

To isolate the impact of $p = \frac{\lambda_0}{\sum_{i=0}^{n} \lambda_i}$, we assume that all customer types have the same expected service time, and all the arrival rates of the regular customers are the same, $\lambda_i = \lambda$ for $i = 1, ..., n$ ($V = 1$). In particular, we are interested to know, for the different ranges of p, which one of the models would be preferred to the other. We choose call center examples with $n = 4$, i.e., 5 agent teams and 5 skills including skill 0. The results are shown in Table 2.2 and Figures 2.5(a) and 2.5(b).

Table 2.2: Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = V = 1$, $p' = 20\%$, $n = 4$)

	Chaining						SP	Crossing value
p	t=0%	t=5%	t=10%	t=25%	t=50%	t=100%		(Chaining = SP)
0%	49	50.95	52.9	58.75	68.5	88	60	t=28.21%
10%	49	50.7	52.4	57.5	66	83	56	t=20.58%
25%	48	49.3	50.6	54.5	61	74	52	t=15.38%
50%	49	49.9	50.8	53.5	58	67	52	t=16.67%
75%	51	51.55	52.1	53.75	56.5	62	51	t=0%
90%	51	51.3	51.6	52.5	54	57	51	t=0%
100%	47	47	47	47	47	47	47	t=0%

Since any agent in SP has skills 0 and i (i.e., costs 1), the staffing cost of SP does not depend on t. In Table 2.2, the column *Crossing value* gives the value of t for which the two models chaining and SP are equivalent. Below this threshold chaining is better than SP and viceversa (see Figure 2.5(a)). Consider small values of t. Table 2.2 reveals that chaining performs well for small values of p. The best situation for chaining is reached in the symmetric case (identical arrival rates). The performance of SP improves as p increases. For small values of p, SP approaches FD which has the worst performance. For high values of p, customers 0 are first preponderant and second benefit



(a) Preference zone

(b) Relative benefits

Figure 2.5: Comparing single pooling and chaining ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = V = 1$, $p' = 20\%$, $n = 4$)

from pooling, which highly improves the performance of SP. With $t = 0$, SP and chaining become equivalent for values of $p \geq 75\%$.

For higher values of t, SP goes ahead of chaining. The reason is related to the increase of the costs of the agents with two skills i and j $(i, j = 1, ..., 4)$. It suffices to have $t = 15.38\%$ to outperform the best performance of chaining (the symmetric case). For any t beyond 30%, SP is systematically better than chaining whatever is p.

We also measure the relative benefits between SP and chaining. Figure 2.5(b) provides, for various values of the relative benefits, the associated curve of t as a function of p. We observe that the sensitivity of the relative benefit as a function of t decreases in p. The reason is that the number of customers 0 increases in p, which decreases the number of agents with two regular skills in chaining (i.e., decreases the cost sensitivity in t).

The main conclusion here is that SP can be better than chaining when the demand for skill 0 is important and/or when skill 0 is less costly than the other ones. The main idea is that as type 0 dominates, they profit in SP from a total pooling from all teams, while chaining, they do profit from a partial pooling from only two teams.

Asymmetry on the other Arrival Rates

The parameter p, that is defined on customers 0, is one way of measuring asymmetry in arrivals. Here, we focus on the asymmetry between regular customer types, measured by $V = \frac{\lambda_1}{\lambda_2} = \frac{\lambda_2}{\lambda_3} = \frac{\lambda_3}{\lambda_4}$. The simulation results for the cases $V = 2$ and 5 are shown in Table 2.3. The experiments for the case $V = 1$ reduces to those given in Table 2.2.

Table 2.3: Impact of V ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = 1$, $p' = 20\%$, $n = 4$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$V = 2$	0%	50	51.8	53.6	59	68	57	t=19.44%
	10%	50	51.55	53.1	57.75	65.5	56	t=19.35%
	25%	49	50.3	51.6	55.5	62	53	t=15.38%
	50%	48	48.8	49.6	52	56	51	t=18.75%
	75%	50	50.55	51.1	52.75	55.5	52	t=18.18%
	90%	52	52.3	52.6	53.5	55	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$V = 3$	0%	50	51.8	53.6	59	68	56	t=16.67%
	10%	50	51.45	52.9	57.25	64.5	55	t=17.24%
	25%	49	50.25	51.5	55.25	61.5	52	t=12.00%
	50%	49	50	51	54	59	52	t=15.00%
	75%	50	50.75	51.5	53.75	57.5	52	t=13.33%
	90%	52	52.2	52.4	53	54	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$V = 5$	0%	49	51.05	53.1	59.25	69.5	54	t=12.20%
	10%	50	51.5	53	57.5	65	54	t=13.33%
	25%	50	51.25	52.5	56.25	62.5	52	t=8.00%
	50%	50	50.7	51.4	53.5	57	52	t=14.29%
	75%	51	51.4	51.8	53	55	52	t=12.50%
	90%	52	52.25	52.5	53.25	54.5	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%

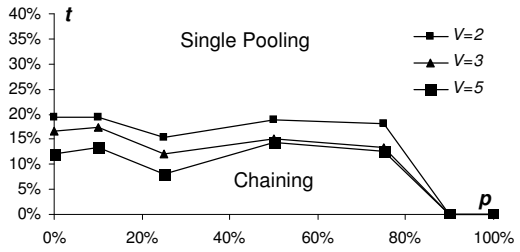


Figure 2.6: Preference zone ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $U = 1$, $p' = 20\%$, $n = 4$)

Table 2.3 and Figure 2.6 reveal that the performance of SP increases in V. An intuitive explanation is as follows. Remark that the team size $s_i = s(\lambda_i)$ is increasing and concave in λ_i, for $i = 1, ..., n$. Applying then the Jensen inequality leads to $\sum_{i=1}^{n} s(\lambda_i) \leq n \cdot s\left(\frac{\sum_{i=1}^{n} \lambda_i}{n}\right)$. In this inequality, the left hand side corresponds to the overall staffing level for an arbitrary situation, i.e., with arbitrary values of λ_is. As for the right hand side, it gives the overall staffing level for a symmetric situation, i.e., all the λ_is are identical. We also observe from Table 2.3 that the performance of chaining is however relatively insensitive to V. Note that we change each time the configuration of chaining such that the large teams are close to each others in order to create more pooling effect. This is better than having small teams each of which connected to a large team.

2.5.2 Asymmetry in Service Rates

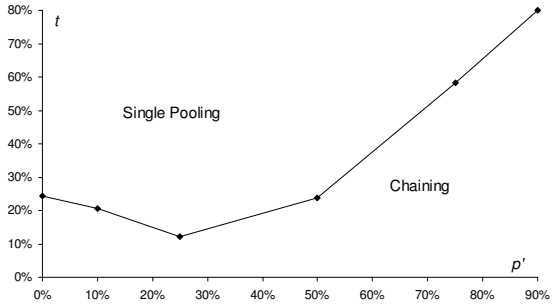
In this section, we focus on the comparison between chaining and SP with regard to the asymmetry in the customer service times. We first define the asymmetry only on customers 0, and second on all customer types.

Asymmetry on Customers 0

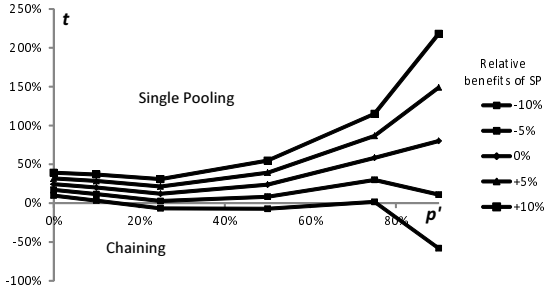
We measure the asymmetry on customers 0 by $p' = \frac{\frac{1}{\mu_0}}{\sum_{i=0}^{n} \frac{1}{\mu_i}}$. The asymmetry here is defined by the difference between the value of the mean service time of customers 0 and that of the regular types. The results are shown in Table 2.4 and Figure 2.7(a).

Table 2.4: Impact of p' ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)

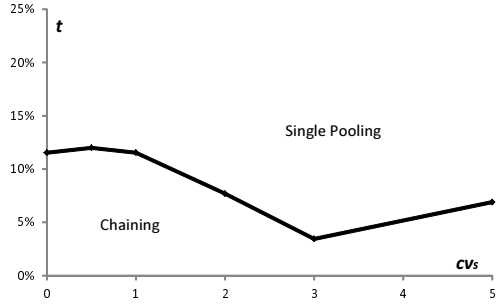
p'	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
0%	60	62.45	64.9	72.25	84.5	72	t=24.49%
10%	59	60.95	62.9	68.75	78.5	67	t=20.51%
25%	58	59.65	61.3	66.25	74.5	62	t= 12.12%
50%	60	61.05	62.1	65.25	70.5	65	t=23.81%
75%	61	61.6	62.2	64	67	68	t=58.33%
90%	65	65.25	65.5	66.25	67.5	69	t=80.00%



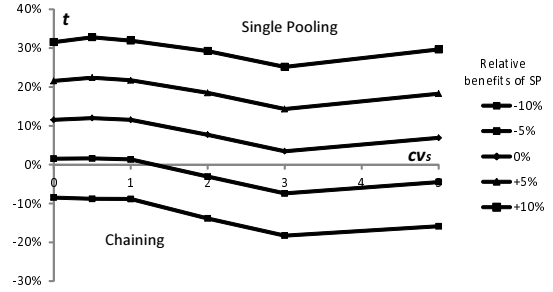
(a) Preference zone ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)



(b) Relative benefits ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$)



(c) Effect of variability in service times($\lambda_0 = 2$, $\lambda_i = 1.5$, $\mu_0 = \mu_i = 0.2$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)



(d) Relative benefits ($\lambda_0 = 2$, $\lambda_i = 1.5$, $\mu_0 = \mu_i = 0.2$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)

Figure 2.7: Preference zone

From Table 2.4, we observe that the performance of both models chaining and SP improves in p' (from 0 until the symmetric case for $p' = 25\%$). The reason is that for chaining we are approaching the symmetric case where it behaves well, and for SP we are profiting better from the pooling effect when all service times are statistically identical. However the performance of the two models deteriorates in p' (for p' above 25%), and no model performs well for a high asymmetry in service times. The explanation is related to a phenomenon referred to as the *blocking effect*. The blocking effect is the situation where the agents are excessively blocked by customers 0 (who are in need of large service times) which deteriorates the waiting time of the regular customers. This phenomenon is more apparent for single pooling since in the latter customers 0 have access to all teams, whereas in chaining they do only have access to two teams. We refer the reader to Tekin et al. (2009) for more details on how pooling could be counterproductive when service times are very different. Recall that this situation with a slow service rate for customers 0 is out of our context. In our

context, customers 0 are in need of an easy skill, and are therefore likely to be served within a short duration. We also measure the relative benefits between SP and chaining as a function of p' (see Figure 2.7(a)). Similarly to the effect of p, we again observe that the sensitivity of the relative benefit as a function of t decreases in p'. The reason is that the service time duration of customers 0 increases in p', which decreases the number of agents with two regular skills in chaining (i.e., decreases the cost sensitivity in t).

In what follows, we go further by defining the asymmetry on the variability of customers 0 service times. We choose to measure this variability by the coefficient of variation (ratio of standard deviation over expected value), denoted by cv_s. We consider a log-normal distribution for the service times of customers 0 (inter-arrival times of all types, and service times of all regular types are Markovian). The choice of the log-normal distribution is based on the call center statistical analysis in Brown et al. (2005). The results are shown in Table 2.5 and Figure 2.7(c). We draw the same conclusions as those for service rates. Due to the blocking effect, both models do not behave well as the variability is increasing. Figure 2.7(d) reveals that the relative benefit as a function of t is not sensitive to the variation of cv_s. To the contrary to the case for p and p', the arrival and service rates of regular types do not vary here.

Table 2.5: Impact of variability in service times ($\mu_i = \mu_0 = 0.2$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = V = 1$, $\sum_{i=0}^{4} \lambda_i = 8$, $n = 4$)

cv_s	0%	5%	10%	25%	50%		value of t
0	49	50.3	51.6	55.5	62	52	11.54%
0.5	49	50.25	51.5	55.25	61.5	52	12.00%
1	50	51.3	52.6	56.5	63	53	11.54%
2	54	55.3	56.6	60.5	67	56	7.69%
3	62	63.45	64.9	69.25	76.5	63	3.45%
5	64	65.45	66.9	71.25	78.5	66	6.90%

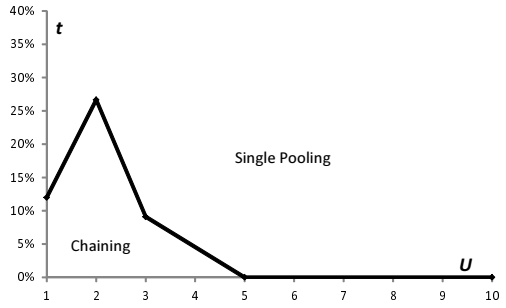
Asymmetry on the Other Service Rates

We examine the impact of asymmetry by defining it on all service times. The service times can be now different from one regular customer to another. Recall that the ratio U is defined by

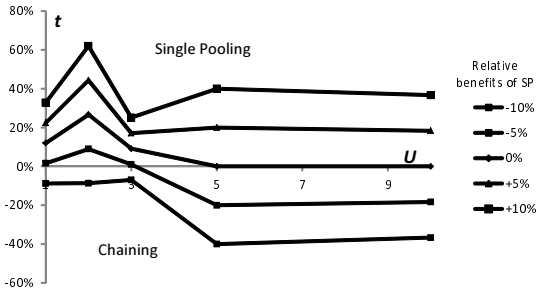
$U = \frac{1/\mu_1}{1/\mu_2} = \frac{1/\mu_2}{1/\mu_3} = \frac{1/\mu_3}{1/\mu_4}$. We also consider cases with a high proportion of customers 0, $p = 50\%$. This can be seen as a worst case for SP, since the blocking effect is more apparent in such a case. The simulation results are shown in Table 2.6, and Figures 2.8(a) and 2.8(b).

Table 2.6: Impact of U ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $V = 1$, $n = 4$)

U	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
1	49	50.25	51.5	55.25	61.5	52	t=12.00%
2	49	49.75	50.5	52.75	56.5	53	t=26.67%
3	50	51.65	52.3	54.25	57.5	53	t=9.09%
5	52	52.65	53.3	55.25	58.5	52	t=0.00%
10	55	55.75	56.5	58.75	62.5	55	t=0.00%



(a) Preference zone



(b) Relative benefits

Figure 2.8: Preference zone ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$ for $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $V = 1$, $n = 4$)

From the numerical results we observe that SP is preferred to chaining for a wide range of parameters. The performance of SP is quite insensitive to the asymmetry defined by U. The reason is that whatever is U, the agent teams in SP are divided to two types. One first type with two teams where customers 0 are served faster than regular customers (positive effect), and a second type with two teams where customer 0 are served slower than regular customers (negative effect of blocking). The performance of chaining is however decreasing in asymmetry. In chaining, each team receives two customer types with different service times, which creates a negative blocking effect in all teams and deteriorates as a consequence the performance. In general for both single pooling and chaining with $U \neq 1$, regular customers require different mean service times. We then

have regular customers that are served faster than others. The slowly served ones block the teams in which they are routed to. This is more apparent in chaining because regular customers are routed to two teams (and to only one in SP). We also measure the relative benefits between SP and chaining. Figure 2.8(b) reveals that this benefit as a function of t is not sensitive to the variation of U. The reason is that although the service rates of regular types do vary, the total staffing level for the regular types do not.

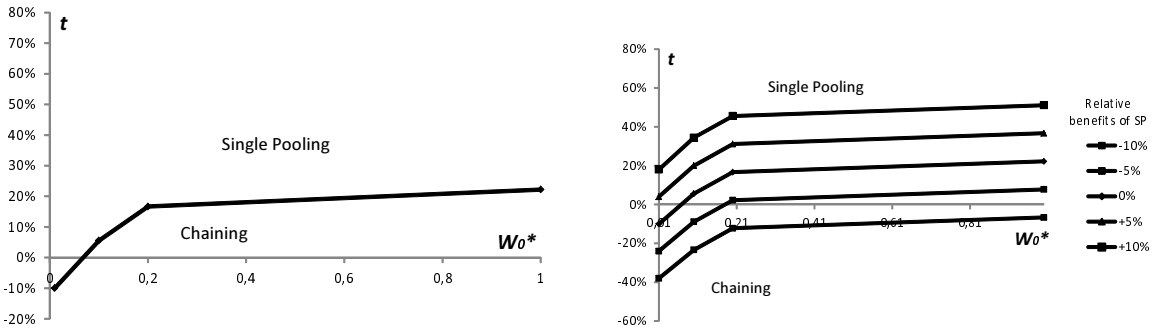
2.5.3 Asymmetry in the Service Level Constraints

We define the asymmetry on the service level of customers 0, W_0^*. The results are shown in Table 2.7 and Figure 2.9(a).

Table 2.7: Impact of W_0^* ($\lambda_0 = 4$, $\lambda_i = 1$, $\mu_i = \mu_0 = 0.2$ and $W_i^* = 0.2$ for $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)

W_0^*	Chaining t=0%	t=5%	t=10%	t=25%	SP	Crossing value (Chaining = SP)
0.01	58	59	60	63	56	t=-10.00%
0.1	51	51.9	52.8	55.5	52	t=5.56%
0.2	49	49.9	50.8	53.5	52	t=16.67%
1	48	48.9	49.8	52.5	52	t=22.22%

We observe as expected that SP behaves better than chaining in the case of a high asymmetry in the service levels. Chaining is requiring higher staffing levels than needed for some customer types. The agent teams are less correlated in SP than in chaining. This gives more flexibility under SP to adjust the size of the teams as required. However, the strong link between the chains in chaining forces the size of the teams to be adjusted with regard to the high requirement of some customer types while it is not needed for other types. As for the relative benefits between SP and chaining, we observe from Figure 2.9(b) that it is not sensitive to the variation of W_0^*. Since the parameters related to the regular types do not vary, the associated staffing levels do not change also.



(a) Preference zone (b) Relative benefits

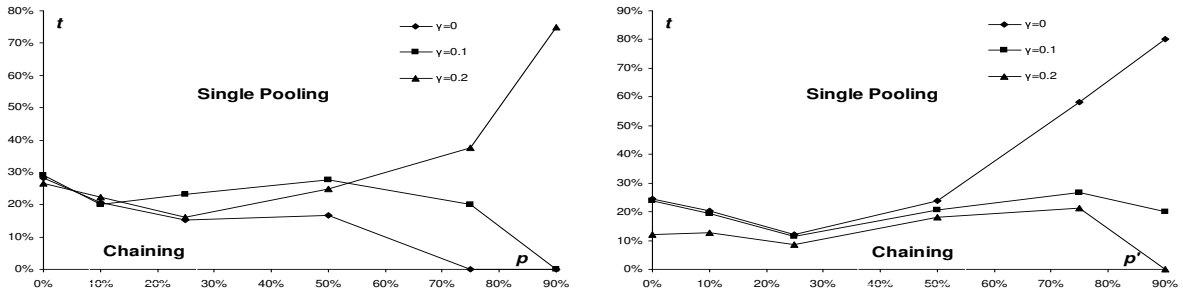
Figure 2.9: Preference zone ($\lambda_0 = 4$, $\lambda_i = 1$, $\mu_i = \mu_0 = 0.2$ and $W_i^* = 0.2$ for $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $U = V = 1$, $n = 4$)

2.5.4 Asymmetry in Abandonments

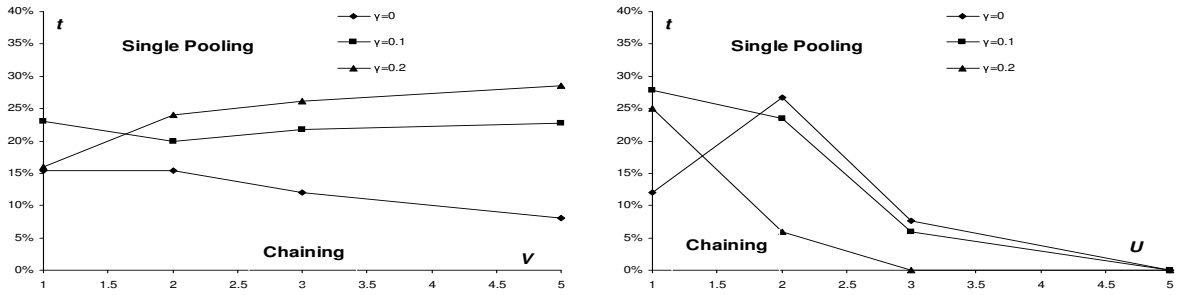
We allow in this section customers to abandon. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time she will abandon and be lost. Abandonment is an important feature in call centers. We first investigate the impact of abandonment on the performance of single pooling and chaining. We then investigate the effect of the asymmetry in the abandonment rate of customers 0. Recall that that γ_i denotes the abandon rate of customers i, for $i = 0, \cdots n$. In the experiments below, times before abandonment are assumed to be exponentially distributed. Note that with customer abandonment, new performance measures do appear for waiting times. Since the sojourn time in queue may end up with a start of service or an abandonment, we distinguish the conditional waiting time given service, that given abandonment, and the unconditional one. We focus here on the conditional waiting time given service.

Impact of Abandonment. We investigate the impact of abandonment on the performance of SP and chaining in various situations of asymmetries. We consider homogeneous abandonments for all customer types, $\gamma_i = \gamma$ for $i = 0, \cdots n$. The results are shown in Figures 2.10(a)-2.10(d). Further results are also given in Tables A.6-A.9 in Section A.3 of the appendix. An important observation here is that the effect of the parameters asymmetry changes in the presence of abandonment. For

example, to the contrary to the results with no abandonment, the performance of SP deteriorates in p, but improves in p'. The reason is that the abandonment of customers reduces the arrivals to service, which in turn reduces the asymmetry. This can be seen from Table 2.8, where the the probability to abandon of customers 0 increases in p.



(a) Impact of p ($\mu_i = \mu_0 = 0.2$, $\sum_{i=0}^{2} \lambda_i = 8$, $\lambda_i = \lambda_j$, (b) Impact of p' ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_0^* = W_i^* = 0.2$, for $i, j = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$, $n = 4$) $n = 4$)



(c) Impact of V ($\lambda_0 = 2$, $\mu_0 = \mu_i = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $W_0 = W_i^* = W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$, (d) Impact of U ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$, 0.2, $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $V = 1$, $n = 4$) $n = 4$)

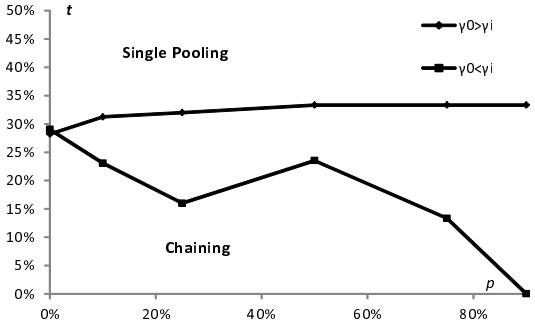
Figure 2.10: Impact of abandonment

Asymmetry in Abandonment. Consider the asymmetry in the abandonment rates measured by the relative difference between the abandonment rate of customers 0 compared to those of the regular customers. The results are shown in Figures 2.11(a)-2.11(d). Further results are also given in Tables A.10-A.13 in Section A.3 of the appendix. We again observe an important impact of the abandonment on the performance of SP and chaining. This impact mainly depend on how the abandonment affects the asymmetry. For example, we observe from Figure 2.11(a) that when regular customers have higher abandonment rates than customers 0, the asymmetry in terms of

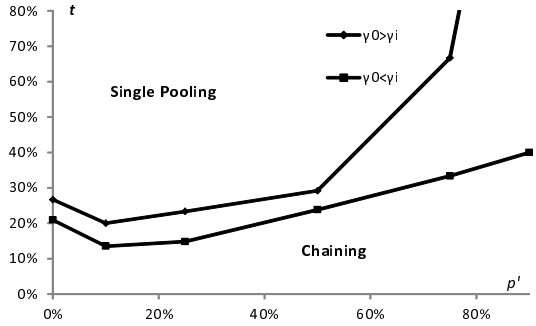
Table 2.8: Probability of abandonment ($\mu_i = \mu_0 = 0.2$, $\sum_{i=0}^{2} \lambda_i = 8$, $\lambda_i = \lambda_j$, $W_0^* = W_i^* = 0.2$, $\gamma_i = \gamma_0 = \gamma$ for $i, j = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

	$\gamma = 0.1$				$\gamma = 0.2$			
	Single Pooling		Chaining		Single Pooling		Chaining	
p	Type i	Type 0	Type i	Type 0	Type i	Type 0	Type i	Type 0
0%	3.04%		1.68%		7.00%		3.20%	
10%	2.05%	0.00%	2.24%	1.95%	5.00%	0.03%	3.58%	3.13%
25%	1.84%	0.01%	1.52%	1.79%	4.30%	0.07%	4.44%	4.24%
50%	1.73%	0.08%	1.69%	2.11%	4.18%	0.40%	3.82%	4.13%
75%	1.70%	0.53%	1.09%	1.27%	4.12%	1.45%	3.80%	3.47%
90%	1.68%	1.14%	0.29%	1.13%	4.10%	2.62%	4.63%	3.33%
100%		1.67%		1.67%		4.25%		4.25%

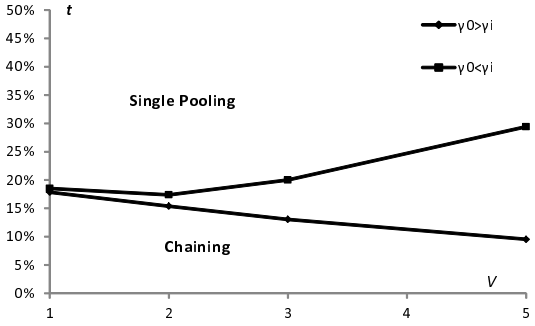
p is accentuated (which further improves SP performance). In the opposite case however, the asymmetry in p reduces because of the abandonment of customers 0.



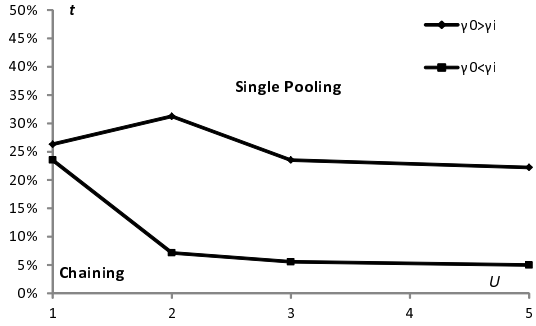
(a) Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)



(b) Impact of p' ($W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $i = 1, ..., 4$, $p = 50\%$, $U = V = 1$, $n = 4$)



(c) Impact of V ($\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\mu_i = \mu_0 = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$, $n = 4$)



(d) Impact of U ($\lambda_0 = 4$, $W_0^* = W_i^* = 0.2$, $\mu_0 = 0.2$, $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $V = 1$, $n = 4$)

Figure 2.11: Impact of the asymmetry in abandonment

Table 2.9: Impact of the Call Center Size ($\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$ for $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

	Small Call Center ($\sum_{i=0}^{4} \lambda_i = 1$)					Large Call Center ($\sum_{i=0}^{4} \lambda_i = 100$)				
	Chaining			SP	Crossing value	Chaining			SP	Crossing value
p	$t = 0\%$	$t = 5\%$	$t = 10\%$		(Chaining = SP)	$t = 0\%$	$t = 5\%$	$t = 10\%$		(Chaining = SP)
0%	12	12.4	12.8	16	$t = 50\%$	513	534.6	556.2	536	$t = 5.32\%$
10%	12	12.4	12.8	16	$t = 50\%$	513	531.15	549.3	518	$t = 1.38\%$
25%	11	11.3	11.6	16	$t = 83.33\%$	513	527.2	541.4	513	$t = 0\%$
50%	12	12.25	12.5	15	$t = 60\%$	513	522.1	531.2	513	$t = 0\%$
75%	13	13.25	13.5	13	$t = 0\%$	515	519.45	523.9	513	$t = -2.25\%$
90%	12	12.15	12.3	11	$t = -33.33\%$	517	519	521	513	$t = -10.00\%$
100%	9	9	9	9	$t = 0\%$	513	513	513	513	$t = 0\%$

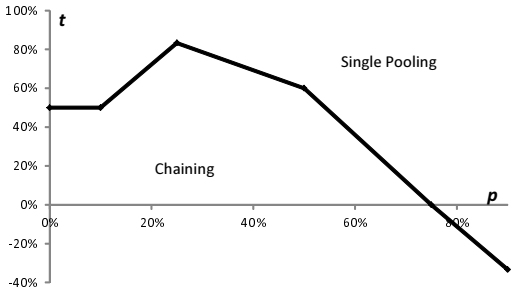
2.5.5 Impact of the Call Center Size

We focus in this section on the impact of the size of the call center on the comparison between the two models. Akşin and Karaesmen (2007) showed that a small call center benefits more from a flexible architecture than a larger one. From the simulation experiments conducted here, we confirm this conclusion. The results are shown in Table 2.9 and Figure 2.12. In Table 2.10 provides the achieved expected waiting times for the optimal staffing levels.

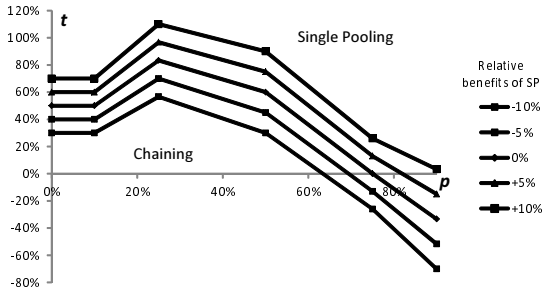
Table 2.10: Expected waiting times ($\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$ for $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

	Small Call Center ($\sum_{i=0}^{4} \lambda_i = 1$)				large Call Center ($\sum_{i=0}^{4} \lambda_i = 100$)			
	Single Pooling		Chaining		Single Pooling		Chaining	
p	W_i	W_0	W_i	W_0	W_i	W_0	W_i	W_0
0%	0.08		0.06		0.18		0.20	
10%	0.07	0.00	0.06	0.06	0.15	0.20	0.19	0.19
25%	0.05	0.00	0.09	0.08	0.08	0.19	0.19	0.20
50%	0.04	0.00	0.07	0.05	0.05	0.17	0.18	0.20
75%	0.19	0.01	0.07	0.02	0.04	0.20	0.17	0.19
90%	0.19	0.03	0.06	0.05	0.02	0.19	0.15	0.18
100%		0.10		0.10		0.17		0.17

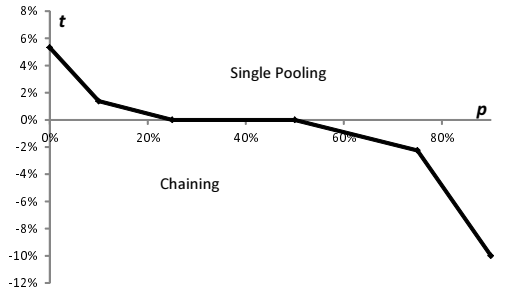
Because of the small teams, the lack of the pooling effect in small call centers makes the threshold values of t higher than those in large call centers. However in large call centers, the team sizes are quite large in the sense that we have a less need to the chains. This makes SP better than chaining even under the symmetric case of arrival rates. From Table 2.10 we observe that to the contrary to small call centers, the service level constraints are saturated for large call centers. Because of



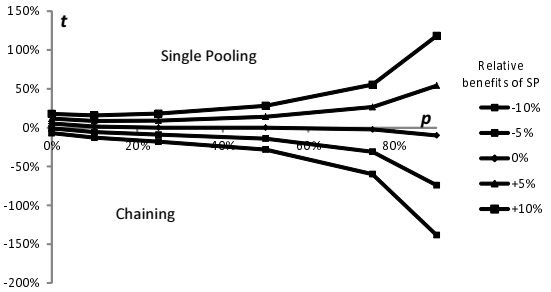
(a) Small Call Center ($\sum_{i=0}^{4} \lambda_i = 1$)



(b) Relative benefits ($\sum_{i=0}^{4} \lambda_i = 1$)



(c) Large Call Center ($\sum_{i=0}^{4} \lambda_i = 100$)



(d) Relative benefits ($\sum_{i=0}^{4} \lambda_i = 100$)

Figure 2.12: Preference zone ($\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$ for $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$, $n = 4$)

the discrete nature of staffing levels, the impact of adding or removing an agent on performance is higher in small call centers. For the same reason, the staffing levels of the regular teams do not vary much in small call centers. This makes the relative benefits between SP and chaining not sensitive to the variation of p in small call centers, while the opposite is true for large call centers (see Figures 2.12(b) and 2.12(d)).

2.5.6 Impact of the Number of Skills

In this section, we investigate the effect of the number of skills (denoted by $N = n + 1$). For two cases with different number of skills, it is not possible to keep at the same time a constant workload on each team and a constant overall workload. We choose to separately treat each situation.

Constant Workload per Team. We consider identical service rates for all customer types. In the experiments below, the ratio $\frac{\sum_{i=0}^{n} \lambda_i}{N}$ is then hold constant. The results are presented in Table 2.11 and Figure 2.13(a). We observe that SP behaves much better than chaining as the number of

skills increases. Figure 2.13(a) shows that for $N = 10$, the crossing value of t should be negative for high values of p (this means that SP is better in all cases). Single pooling behaves much better than chaining for the following two reasons. First as N increases, the flexibility in chaining decreases. A customer type in the chaining configuration has access to a fewer proportion of agent as N increases (the gap with the full flexible model increases). The second reason is related to the impact of the constant ratio $\frac{\sum_{i=0}^{n}\lambda_i}{N}$, which increases the overall size of the call center as N increases. Having large call centers makes SP more efficient (see Section 2.5.5).

Table 2.11: Impact of the number of skills ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{n}\lambda_i/N = 2$, $i = 1, ..., n$, $U = V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$N = 3$	0%	36	37.8	39.6	45	54	40	t=11.11%
	10%	37	38.05	39.1	42.25	47.5	40	t=14.29%
	25%	37	37.75	38.5	40.75	44.5	39	t=13.33%
	50%	37	37.4	37.8	39	41	37	t=0.00%
	75%	36	36.15	36.3	36.75	37.5	36	t=0.00%
	90%	36	36.05	36.1	36.25	36.5	36	t=0.00%
	100%	36	36	36	36	36	36	t=0.00%
$N = 4$	0%	48	49.95	51.9	57.75	67.5	54	t=15.38%
	10%	48	49.45	50.9	55.25	62.5	52	t=13.79%
	25%	47	48.15	49.3	52.75	58.5	50	t=13.04%
	50%	48	48.8	49.6	52	56	48	t=0.00%
	75%	48	48.5	49	50.5	53	48	t=0.00%
	90%	47	47.25	47.5	48.25	49.5	47	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$N = 5$	0%	60	62.55	65.1	72.75	85.5	68	t=15.69%
	10%	59	61	63	69	79	67	t=20.00%
	25%	58	59.6	61.2	66	74	64	t=18.75%
	50%	59	60.1	61.2	64.5	70	61	t=9.09%
	75%	60	60.75	61.5	63.75	67.5	61	t=6.67%
	90%	61	61.3	61.6	62.5	64	61	t=0.00%
	100%	57	57	57	57	57	57	t=0.00%
$N = 10$	0%	116	121	126	141	166	144	t=28.00%
	10%	115	119.6	124.2	138	161	135	t=21.74%
	25%	115	118.8	122.6	134	153	126	t=14.47%
	50%	117	119.75	122.5	130.75	144.5	117	t=0.00%
	75%	120	121.65	123.3	128.25	136.5	114	t=-18.18%
	90%	122	122.95	123.9	126.75	131.5	110	t=-63.16%
	100%	109	109	109	109	109	109	t=0.00%

Constant Overall Workload. We again consider identical service rates for all customer types. The summation $\sum_{i=0}^{n}\lambda_i$ is then hold constant. The results are presented in Table 2.12 and Figure 2.13(b). We distinguish two effects depending on p. For small values of p, the preference zone for SP reduces. The opposite is true for large values of p. The reason is related to the decreasing of

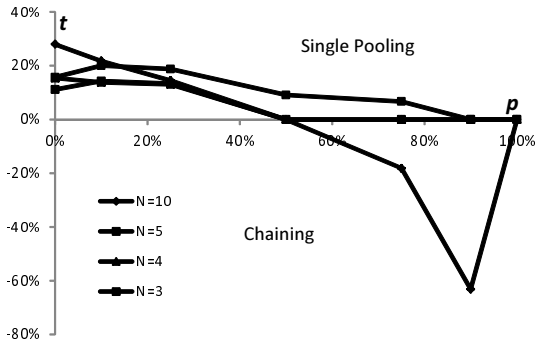
the size of each team as N increases. Since we keep constant the overall workload, increasing the number of skills implies a lower demand per skill, which requires less agents per team. This makes the effect of pooling predominant. For the case of large p, the large number of customers 0 benefits from pooling under SP. For the case of small p, the system contains more regular customers, each of which benefits in chaining from the pooling of two adjacent teams.

Table 2.12: Impact of p, t and N on the staffing cost ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{n} \lambda_i = 8$, $i = 1, ..., n$, $U = V = 1$)

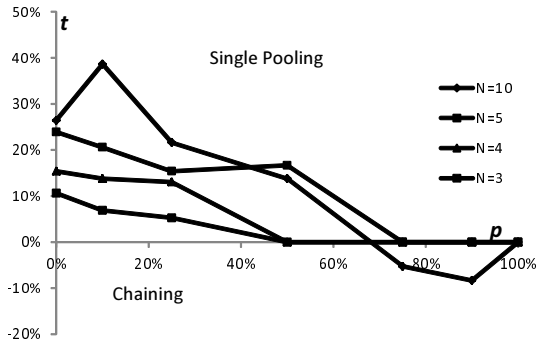
		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$N = 3$	0%	47	49.35	51.7	58.75	70.5	52	t=10.64%
	10%	47	48.45	49.9	54.25	61.5	49	t=6.90%
	25%	47	47.95	48.9	51.75	56.5	48	t=5.26%
	50%	47	47.5	48	49.5	52	47	t=0.00%
	75%	47	47.2	47.4	48	49	47	t=0.00%
	90%	47	47.05	47.1	47.25	47.5	47	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$N = 4$	0%	48	49.95	51.9	57.75	67.5	54	t=15.38%
	10%	48	49.45	50.9	55.25	62.5	52	t=13.79%
	25%	47	48.15	49.3	52.75	58.5	50	t=13.04%
	50%	48	48.8	49.6	52	56	48	t=0.00%
	75%	48	48.5	49	50.5	53	48	t=0.00%
	90%	47	47.25	47.5	48.25	49.5	47	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$N = 5$	0%	49	51.3	53.6	60.5	72	60	t=23.91%
	10%	49	50.7	52.4	57.5	66	56	t=20.59%
	25%	48	49.3	50.6	54.5	61	52	t=15.38%
	50%	49	49.9	50.8	53.5	58	52	t=16.67%
	75%	51	51.55	52.1	53.75	56.5	51	t=0.00%
	90%	51	51.3	51.6	52.5	54	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$N = 10$	0%	58	60.65	63.3	71.25	84.5	72	t=26.42%
	10%	55	57.2	59.4	66	77	72	t=38.64%
	25%	55	56.85	58.7	64.25	73.5	63	t=21.62%
	50%	56	57.45	58.9	63.25	70.5	60	t=13.79%
	75%	57	57.95	58.9	61.75	66.5	56	t=-5.26%
	90%	56	56.6	57.2	59	62	55	t=-8.33%
	100%	47	47	47	47	47	47	t=0.00%

2.5.7 Mix of Asymmetry

In this section we mix the effects of more than a parameter at a time. We propose to interact the effects of p and p', U and p, U and p', U and V, and also all of them. The results are presented in Tables 2.13-2.17 and Figures 2.14(a)-2.14(d). From the numerical results, we observe that the individual effects are still present, but they may accumulate or make up for one another.



(a) Preference zone ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{n} \lambda_i/N = 2$, $i = 1, ..., 4$, $U = V = 1$)



(b) Preference zone ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{n} \lambda_i = 8$, $i = 1, ..., 4$, $U = V = 1$)

Figure 2.13: Preference zone

One important observation is that two asymmetries may lead to a bad performance for SP. For example SP behaves well in each one of the asymmetric situations ($U = 2$ and $V = 1$) and ($U = 1$ and $V = 1/3$) in isolation. However, it does not behave well for the mixed situation ($U = 2$ and $V = 1/3$). In such a situation, the customers types with large arrival rates are the faster to be served, and viceversa. Therefore, the different customer types workloads are likely to be symmetric. For the same reason, SP behaves well in the situation ($U = 2$ and $V = 3$) because the mix of asymmetries further accentuates the asymmetry in workloads.

Tables 2.14 and 2.15 reveal also that the most predominant effects are those of p (because of pooling) and p' (because of blocking). Various scenarios of mixed asymmetries are considered in Table 2.17. We find again that SP behaves well in large call centers (the first four scenarios). Scenarios 3 and 7 are similar in terms of the values of p and p' (high values for the two parameters). This means that the effect of pooling and blocking are highly present in both scenarios. An important observation here is that scenario 3 is the best among scenarios 1-4, while scenario 7 is the worst among scenarios 5-8. This gives an indication on the direct competition between the effects of p and p'. In large call centers, the pooling effect created by customers 0 is predominant over the blocking effect, and the opposite is true in small call centers.

Main Conclusions. In summary, the numerical analysis of this section confirms that single pooling

Table 2.13: Impact of p and p' ($\mu_i = \mu_j$ and $\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$,$\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $U = V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$p' = 10\%$	0%	52	53.95	55.9	61.75	71.5	64	t=30.77%
	10%	51	52.7	54.4	59.5	68	60	t=26.47%
	25%	46	47.45	48.9	53.25	60.5	52	t=20.69%
	50%	39	39.9	40.8	43.5	48	43	t=22.22%
	75%	35	35.6	36.2	38	41	35	t=0.00%
	90%	31	31.3	31.6	32.5	34	31	t=0.00%
	100%	24	24	24	24	24	24	t=0.00%
$p' = 20\%$	0%	49	50.95	52.9	58.75	68.5	60	t=28.21%
	10%	49	50.7	52.4	57.5	66	56	t=20.58%
	25%	48	49.3	50.6	54.5	61	52	t=15.38%
	50%	49	49.9	50.8	53.5	58	52	t=16.67%
	75%	51	51.55	52.1	53.75	56.5	51	t=0.00%
	90%	51	51.3	51.6	52.5	54	51	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$p' = 50\%$	0%	24	24.95	25.9	28.75	33.5	40	t=84.21%
	10%	40	41.05	42.1	45.25	50.5	53	t=61.90%
	25%	53	53.85	54.7	57.25	61.5	63	t=58.82%
	50%	75	75.6	76.2	78	81	82	t=58.33%
	75%	97	97.4	97.8	99	101	100	t=37.50%
	90%	111	111.2	111.4	112	113	112	t=25.00%
	100%	112	112	112	112	112	112	t=0.00%

performs better than chaining for various cases of asymmetry in the system parameters. In the case of a predominance of customers 0 and/or an important asymmetry in the arrival rates of the regular types (captured by V), SP is more robust than chaining even for small differences between the costs of a regular skill and that of skill 0. Because of the blocking effect, the performance of both chaining and SP deteriorates in the asymmetry defined by the service time duration of customers 0 relatively to that of regular customers. This is more apparent in single pooling because customers 0 have access to all teams, while in chaining they do only have access to two teams. We have also observed that SP is more robust than chaining against an increasing asymmetry between the service times of regular types. Since the teams under SP are less inter-dependent than under chaining, SP is again preferred in the case of an asymmetry between the objective service levels. We therefore avoid over-staffing situations that may happen in chaining. Another important feature is that of abandonment, because it may affect the asymmetry of the parameters. Finally, all above conclusions are more apparent for the situations with a large number of skills, or for those with a

Table 2.14: Impact of p and U ($\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$,$\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p' = 20\%$, $V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$U = 2$	0%	49	50.8	52.6	58	67	57	t=22.22%
	10%	49	50.6	52.2	57	65	55	t=18.75%
	25%	49	50.15	51.3	54.75	60.5	53	t=17.39%
	50%	49	49.75	50.5	52.75	56.5	53	t=26.67%
	75%	51	51.5	52	53.5	56	53	t=20.00%
	90%	53	53.35	53.7	54.75	56.5	53	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$U = 3$	0%	51	52.4	53.8	58	65	55	t=14.29%
	10%	50	51.4	52.8	57	64	53	t=10.71%
	25%	50	51.5	53	57.5	65	53	t=10.00%
	50%	50	51.3	52.6	56.5	63	52	t=7.69%
	75%	51	52.3	53.6	57.5	64	52	t=3.85%
	90%	52	53.35	54.7	58.75	65.5	52	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%
$U = 5$	0%	52	53.2	54.4	58	64	55	t=12.50%
	10%	51	51.95	52.9	55.75	60.5	53	t=10.53%
	25%	52	53.2	54.4	58	64	53	t=4.17%
	50%	52	52.05	52.1	52.25	52.5	52	t=0.00%
	75%	52	52.05	52.1	52.25	52.5	52	t=0.00%
	90%	52	52.15	52.3	52.75	53.5	52	t=0.00%
	100%	47	47	47	47	47	47	t=0.00%

large call center size.

2.6 Concluding Remarks

We focused on a fundamental problem in the design and management of SBR call centers, for which it is important to choose an intelligent architecture. We considered the context of call centers with unbalanced workload, different service requirements, a predominant customer type and high costs of cross-training. With these asymmetry in the parameters, the well known existing architectures such as chaining lose their robustness. For those particular cases, we proposed a new call center architecture (single pooling) and demonstrated its efficiency. SP allows to balance the workload among the agents in a way that captures the benefits of pooling, without requiring every agent to process every type of call. The results of the comparison between SP and chaining have significant managerial implications. We showed that SP behaves well in most cases of asymmetry in the parameters. There might be then opportunities for managers of call centers to improve

Table 2.15: Impact of p' and U ($\lambda_i = \lambda_j = 1.5$ for $i, j = 1, ..., 4$, $\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p = 25\%$, $V = 1$)

		Chaining					SP	Crossing value
	p'	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$U = 2$	0%	51	52.15	53.3	56.75	62.5	55	t=17.39%
	10%	50	51.2	52.4	56	62	54	t=16.67%
	25%	51	52.25	53.5	57.25	63.5	54	t=12.00%
	50%	51	52.25	53.5	57.25	63.5	56	t=20.00%
	75%	52	53.25	54.5	58.25	64.5	62	t=40.00%
	90%	52	53.25	54.5	58.25	64.5	68	t=64.00%
$U = 3$	0%	52	53.15	54.3	57.75	63.5	54	t=8.70%
	10%	51	52.2	53.4	57	63	53	t=8.33%
	25%	51	52.25	53.5	57.25	63.5	53	t=8.00%
	50%	51	52.25	53.5	57.25	63.5	56	t=20.00%
	75%	54	55.25	56.5	60.25	66.5	61	t=28.00%
	90%	56	57.3	58.6	62.5	69	67	t=42.31%
$U = 5$	0%	52	53.2	54.4	58	64	53	t=4.17%
	10%	51	52.25	53.5	57.25	63.5	52	t=4.00%
	25%	51	52.3	53.6	57.5	64	52	t=3.85%
	50%	52	53.3	54.6	58.5	65	54	t=7.69%
	75%	55	56.25	57.5	61.25	67.5	60	t=20.00%
	90%	58	59.3	60.6	64.5	71	66	t=30.77%

performance using the single pooling architecture.

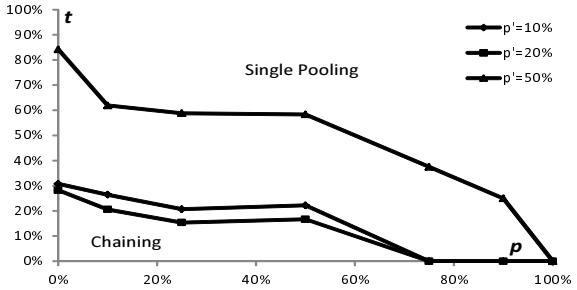
In a future research, it would be useful to extend the use of the fixed point algorithm to evaluate the performance measures of customers 0. Another interesting work is to generalize the functioning of single pooling in order to avoid the blocking effect in the case of long service times for customers 0.

Table 2.16: Impact of U and V ($\lambda_0 = 0.8$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ and $\sum_{i=0}^{4} \lambda_i = 8$, $p = 10\%$, $p' = 20\%$)

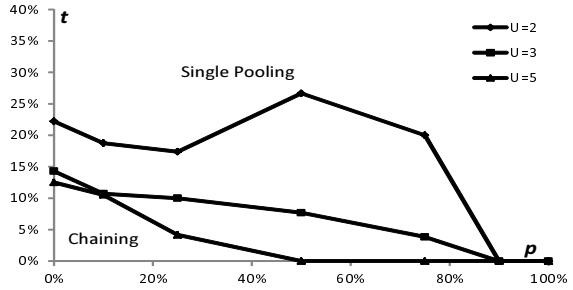
U	V	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
1	1/3	50	51.45	52.9	57.25	64.5	55	t=17.24%
	1/2	50	51.55	53.1	57.75	65.5	56	t=19.35%
	1	49	50.7	52.4	57.5	66	56	t=20.58%
	2	50	51.55	53.1	57.75	65.5	56	t=19.35%
	3	50	51.45	52.9	57.25	64.5	55	t=17.24%
2	1/3	26	26.8	27.6	30	34	34	t=50.00%
	1/2	31	32	33	36	41	37	t=30.00%
	1	49	50.6	52.2	57	65	55	t=18.75%
	2	71	72.6	74.2	79	87	76	t=15.63%
	3	81	82.3	83.6	87.5	94	84	t=11.54%
3	1/3	20	20.55	21.1	22.75	25.5	26	t=54.55%
	1/2	24	24.65	25.3	27.25	30.5	31	t=53.85%
	1	50	51.4	52.8	57	64	53	t=10.71%
	2	79	80.65	82.3	87.25	95.5	82	t=9.09%
	3	93	94.35	95.7	99.75	106.5	95	t=7.41%

Table 2.17: Impact of p, p', U and V ($W_i^* = 0.2$ for $i = 0, \cdots 4$)

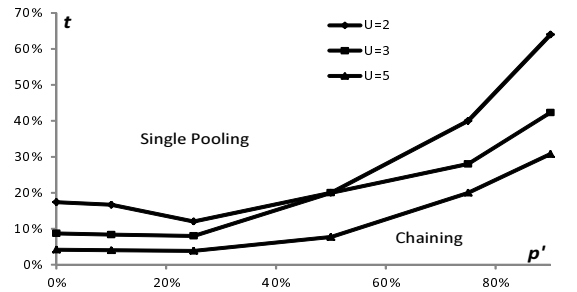
	Scenarios λ_1	λ_2	λ_3	λ_4	λ_0	μ_1	μ_2	μ_3	μ_4	μ_0	Chaining $t = 0\%$	$t = 10\%$	$t = 20\%$	SP	Crossing value (Chaining=SP)
Sc 1	1	2	3	4	5	0.05	0.1	0.2	0.5	1	78	83.5	89	87	16.36%
Sc 2	2	3	4	5	1	0.05	0.1	0.2	0.5	1	115	122.6	130.2	127	15.79%
Sc 3	1	2	3	4	5	1	0.5	0.2	0.1	0.05	179	184.9	190.8	184	8.47%
Sc 4	2	3	4	5	1	1	0.5	0.2	0.1	0.05	111	116.9	122.8	119	13.56%
Sc 5	0.1	0.2	0.3	0.4	0.5	0.05	0.1	0.2	0.5	1	14	15	16	18	40.00%
Sc 6	0.2	0.3	0.4	0.5	0.1	0.05	0.1	0.2	0.5	1	18	19.4	20.8	24	42.86%
Sc 7	0.1	0.2	0.3	0.4	0.5	1	0.5	0.2	0.1	0.05	26	26.7	27.4	30	57.14%
Sc 8	0.2	0.3	0.4	0.5	0.1	1	0.5	0.2	0.1	0.05	18	18.7	19.4	21	42.86%



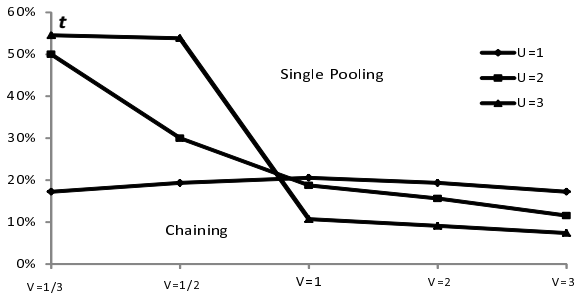
(a) Impact of p and p' ($\mu_i = \mu_j$ and $\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $U = V = 1$)



(b) Impact of U and p ($\lambda_i = \lambda_j$ for $i, j = 1, ..., 4$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p' = 20\%$, $V = 1$)



(c) Impact of U and p' ($\lambda_i = \lambda_j = 1.5$ for $i, j = 1, ..., 4$, $\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ $i = 1, ..., 4$, $p = 25\%$, $V = 1$)



(d) Impact of U and V ($\lambda_0 = 0.8$, $\mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$ and $\sum_{i=0}^{4} \lambda_i = 8$, $p = 10\%$, $p' = 20\%$)

Figure 2.14: Preference zone

Chapter 3

Optimal Email routing in a Multi-Channel Call Center

Motivated by the call center practice, we consider a blended call center with calls arriving over time and an infinitely backlogged queue of emails. Calls have a non-preemptive priority over emails. The call service is characterized by three successive stages where the second one is a break, i.e., there is no required interaction between the customer and the agent for a non-negligible duration. This leads to a new opportunity to efficiently split the agent time between calls and emails.

We focus on the optimization of the email routing to agents. Our objective is to maximize the throughput of emails subject to a constraint on the call waiting time. We develop a general framework with two parameters for the email routing to agents. One parameter controls the routing between calls, and the other does the control inside a call. We then derive various structural results with regard to the optimization problem and numerically illustrate them. Various guidelines to call center managers are provided. In particular, we prove for the optimal routing that all the time at least one of the two email routing parameters has an extreme value.

3.1 Introduction

Context and Motivation: Call centers are an important part of customers' service in many organizations. New technology-driven innovations are multiplying the opportunities to make more efficient use of an agent as she can handle different types of workflow, including inbound calls, outbound calls, email, chat, etc. However, several issues on the management of call center operations emerged also as a result of advanced technology. In this chapter, we consider a call center with two types of jobs, inbound calls and emails. We focus on how to efficiently share the agent time between the two types of jobs in order to improve the call center performance.

In practice, we encountered call center situations where inbound calls and emails are combined. This is called *blending*. The key distinction of problems with blending comes from the fact that emails are less urgent and can be inventoried to some extent, relative to incoming calls. Therefore managers are likely to give a strict priority to calls over emails. An important question here is what should be the best way of routing of emails (or the non-urgent job) to agents, i.e., as a function of the systems parameters and the service level constraints (on calls and emails) when should we ask the agent to treat emails between the call conversations (Bernett et al., 2002; Bhulai and Koole, 2003; Legros et al., 2013a). The email routing question is further important in the context of the call center applications we consider here. We encountered examples where a call conversation between an agent and a customer contains a *natural break*. We mean by this a time interval with no interaction between the agent and the customer. During the conversation, the agent asks the customer to do some necessary operations in her own (without the need of the agent availability). After finishing those operations, the conversation between the two parties can start again. Inside an underway conversation, the agent is then free to do another task if needed. For an efficient use of the agent time, one would think about the routing of the less urgent jobs (emails) not only when the system is empty of calls, but also during call conversations. In practice, such a situation often occurs. For example, an agent in an internet hotline call center asks the customer to reboot her

modem or her computer which may take some time where no interactions can take place. It is also often the case that a call center agent of an electricity supplier company asks the customer for the serial number of her electricity meter box. This box is usually located outside of the house and is locked, so, the customer needs some non-negligible time to get the required information. Another example is that of commercial call centers with a financial transaction during the call conversation. After some time from the start of the call conversation, the customer is asked to do an online payment on a website before coming back to the same agent in order to finish the conversation. The online paiement needs that the customer looks for her credit card, then she enters the credit card numbers, then she goes through the automated safety check with her bank (using SMS, etc.), which may take some minutes.

In the call center examples we encountered, the back office job could be a confirmation email of subscription or unsubscription, simple answers for various customer requests, etc. The answers to these emails usually consist on a set of preprepared text blocks that the agent should mix/adapt with the customer case. Some minutes are then enough for the agent to handle more than one email. For such situations, it is natural that call center managers think about using the opportunity to route emails to an agent during the break of an undergoing call conversation, and not only when no calls are waiting in the queue. The main advantages are *i)* an efficient use of the agent time and therefore better call center performance, *ii)* also, agents become less bored because of the diversity of activities, and therefore they are kept from falling into a rut.

Main Contributions: In this chapter, we consider a call center with an infinite amount of outbound jobs (emails), and inbound jobs (calls) arriving over time for which a break is required in the middle of the call conversation. Given this type of call centers, we are interested in optimizing its functioning by controlling how the resource is shared between the two types of jobs. Calls are more important than emails in the sense that calls request a quasi-instantaneous answer (waiting time in

the order of some minutes), however emails are more flexible and could be delayed for several tens of hours. An appropriate functioning is therefore that the agent works on calls as long as there is work to do for calls. The agent can then work on emails when she becomes free from calls, i.e., after a service completion when no calls are waiting in the queue, or during the call conversation break. We assume that calls have a non-preemptive priority over emails, which means that if a call is busy with an email (that has started after a service completion or during the break), the agent will finish first the email before turning to a new arrived call to the queue or a call that has accomplished the requested operations and wants to start again the conversation to finish her service. The non-preemption priority rule is coherent with the operations in practice and also to the call center literature (Bhulai and Koole, 2003; Deslauriers et al., 2007). It is not appropriate to stop the service of a low priority customer, and it is not efficient for the agent to stop the treatment of an email or to group emails for a simultaneous treatment. In Appendix in Section B.1 we prove that the simultaneous treatment is not an interesting opportunity in a call center.

We focus on the research question: when should the agent treat emails? Between calls, or inside a call conversation, or in both situations? Given the nature of the job types, a call center manager in practice would be interested in maximizing the number of treated emails while respecting some service level objective on the call waiting time (Bhulai and Koole, 2003). For calls, we are interested in the steady-state performance measures in terms of the expected waiting time, the probability that the waiting time is less than a given threshold, and the probability of delay. We do not consider the call waiting after the break because it is not perceived as badly as that before entering the first stage of service (less uncertainty for the customer because she has been already connected to an agent). For emails, we are interested in the steady-state performance in terms of the throughput of emails, i.e., the number of treated emails per unit of time.

The email routing problem considered in this chapter is a part of a collaboration between the authors and the French consulting company Interact-iv.com. In the small call center customers of

Interact-iv.com, an efficient control of the agent time is important. For those call centers, Interact-iv.com wants to implement, in the email dispatcher, an intelligent email routing algorithm adapted to the call system load.

Despite its prevalence, there are no papers in the call center literature addressing such a question. Most of the related papers only focus on the email routing between call conversations but not inside a call conversation. To answer this question, we develop a general framework with two parameters for the email routing to agents. One parameter controls the routing between calls, and the other does the control inside a call conversation. For the tractability of the analysis, we first focus on the single server case. We then discuss the extension of the results to the multi-server case. For the single server modeling, we first evaluate the performance measures using a Markov chain analysis. Second, we propose an optimization method of the routing parameters for the problem of maximizing the email throughput under a constraint on the service level of the call waiting time. As a function of the system parameters (the server utilization, the email service time, the severity of the call service level constraint, etc.), we derive various guidelines to managers. In particular, we prove for the optimal routing that all the time at least one of the two email routing parameters has an extreme value. As detailed later in this chapter, an extreme value means that the agent should do all the time emails inside a call (or between calls) or not at all. In other cases the parameters lead to randomized policies. We also solve our optimization problem by proposing 4 particular cases corresponding to the extreme values of the probabilistic parameters. We analytically derive the conditions under which one particular case would be preferred to another one. The interest from these particular cases is that they are easy to understand for agents and managers. Several numerical experiments are used to illustrate the analysis. To simplify the Markov chain analysis, we further propose an approximation method under the light-traffic regime.

The rest of the chapter is organized as follows. In Section 3.2 we review some of the related

literature. In Section 3.3, we describe the blended call center modeling and the optimization problem. In Section 3.4, we develop a method based on the analysis of Markov chains in order to derive the performance measures of interest for calls and for emails. In Section 3.5, we focus on optimizing the email routing parameters. We also provide various numerical illustrations and discuss the results. In Section 3.6, we develop an approximation method for the system performance evaluation under the light-traffic regime. We also propose an approximation method to extend the results to the multi-server case. Finally in Section 3.7, we provide some concluding remarks and directions for future research.

3.2 Literature Review

There are three related streams of literature to this chapter. The first one deals with blended call centers. The second one is the Markov chain analysis for queueing systems with phase type service time distributions. The third one is related to the cognitive analysis, or in other words the ability for an agent to treat and switch between different job types.

The literature on blended call centers consists on developing performance evaluation and optimal blending policies. Deslauriers et al. (2007) develop a Markov chain for the modeling of a Bell Canada blended call center with inbound and outbound calls. The performance measures of interest are the rate of outbound calls and the waiting time of inbound calls. Through simulation experiments they prove the efficiency of their Markov chain model to reflect reality. Brandt and Brandt (1999) develop an approximation method to evaluate the performance of a call center model with impatient inbound calls and infinitely patient outbound calls of lower priority than the inbound traffic. Bhulai and Koole (2003) consider a similar model to the one analyzed in this chapter, expect that the call service is done in a single stage without a possible break. The model consists on inbound and outbound jobs where the inbound jobs have a non-preemptive priority over the outbound ones. For the special case of identically distributed service times for the two jobs, they optimize the outbound

jobs routing subject to a constraint on the expected waiting time of inbound jobs. Gans and Zhou (2003a) study a call center with two job types where one of the jobs is an infinitely backlogged queue. They develop a routing policy consisting on the reservation of servers in order to maximize the throughput on the jobs of the infinitely backlogged queue. Armony and Maglaras (2004a) analyze a similar model with a callback option for incoming customers. The customer behavior is captured through a probabilistic choice model. Other references include (Bernett et al., 2002; Keblis and Chen, 2006; Pichitlamken et al., 2003).

The analysis in this chapter is also related to the analysis of queueing systems with phase type service time distributions. We model the call service time through three successive exponentially distributed stages, where the second stage may also overlap with the service of one or several emails with an exponential time duration for each. The performance evaluation of such systems involves the stationary analysis of Markov chains and is usually addressed using numerical methods. We refer the reader to Kleinrock (1975) for simple models with Erlang service time distributions. For more complex systems, see Neuts (1982); Sze (1984); Bolotin (1994); Brown et al. (2005); Guo and Zipkin (2008). Our approach to derive the performance measures is based on first deriving the stationary system state probabilities for two-dimension and semi-infinite continuous time Markov chains. One may find in the literature three methods for solving such models. The first one is to truncate the state space, see for example Seelen (1986) and Keilson et al. (1987). The second method is called spectral expansion (Daigle and Lucantoni, 1991; Mitrani and Chakka, 1995; Choudhury et al., 1995). It is based on expressing the invariant vector of the process in terms of the eigenvalues and the eigenvectors of a matrix polynomial. The third one is the matrix-geometric method, see Neuts (1981). The approach relies on determining the minimal positive solution of a non-linear matrix equation. The invariant vector is then expressed in terms of powers of itself. In our analysis, we reduce the problem to solving cubic and quartic equations, for which we use the method of Cardan and Ferrari (Gourdon, 1994).

Finally, we briefly mention some studies on human multi-tasking, as it is the case for the agents in our setting. Gladstones et al. (1989) show that a simultaneous treatment of tasks is not efficient even with two easy tasks because of the possible interferences. In our models, we are not considering successive tasks in the sense that an agent can not talk to a customer and at the same time treats an email. More interestingly, Charron and Koechlin (2010) studied the capacity of the frontal lobe to deal with different tasks by alternation (as here for calls and emails). They develop the notion of *branching*: capacity of the brain to remember information while doing something else. They show that the number of tasks done alternatively has to be limited to two to avoid loss of information. Dux et al. (2009) showed that training and experience can improve multi-tasking performance. The risk from alternating between two tasks is the loss of efficiency because of switching times. An important aspect to avoid inefficiency as pointed out by Dux et al. (2009) and Charron and Koechlin (2010) is that the alternation should be at most between two tasks quite different in nature (like inbound and outbound jobs).

3.3 Problem Description and Modeling

For tractability and a better understanding of the results, we first focus on a single server queueing model. We then extend the analysis to the multi-server case in Section 3.6.2 using simulation. We consider a single server queue with two types of jobs: calls and emails. The arrival process of customers is assumed to be Poisson with mean arrival rate λ. We assume to have an infinite amount of emails that are waiting to be treated in a dedicated first come, first served (FCFS) queue with an infinite capacity. One can think of a call center that stores a sufficiently large number of emails of a given day and handles them the next one.

Upon arrival, a call is immediately handled by the agent, if available. If not, the call waits for service in an infinite FCFS dedicated queue. Calls have a non-preemptive priority over emails. This means that the idle agent deals with a call first (the first in line). If the queue of calls is empty,

this agent can handle an email (the first in line) from the queue of emails. Non-preemption priority is a natural assumption for our application. An agents in practice prefers to finish answering an underway email rather than starting it over later on. This is also preferred from an efficiency perspective. We assume in our models that there is no call abandonment or retrial.

As mentioned in Section 3.1, we consider call center applications where the communication between the agent and the customer includes a break (the customer does not need the agent availability). We model the service time of a call by 3 successive stages. The first stage is a conversation between the two parties. The second stage is the break, i.e., no interactions between the two parties. The third and final step is a again a conversation between the two parties. The service completion occurs as soon as the third stage finishes. We model each stage duration as an exponentially distributed random variable. The service rates of the first, second and third stages are denoted by μ_1, μ_2 and μ_3, respectively. This Markovian assumption, which is common in modeling in service operations, is reasonable for systems with high service time variability where service times are typically small but there are occasionally long service times. An agent handle an email within one single step without interruption. The time duration of an email treatment is random and assumed to be exponentially distributed with rate μ_0.

We are interested in an efficient use of the agent time between calls and emails. More concretely, we want to answer the question when should we treat emails for the following optimization problem

$$\begin{cases} \text{Maximize the throughput of emails} \\ \text{subject to a service level constraint on the call waiting time in the queue.} \end{cases} \tag{3.1}$$

To solve Problem (3.1), we propose a general model for the routing of emails to the agent. It is refereed to as *probabilistic model* or *Model PM* and is described below. Recall that the call center has an infinite number of emails. Then defining the routing of emails consists of determining whether or not to start an email each time the agent becomes idle inside a call conversation, or after a call service completion with no waiting calls in the queue. Note that when the agent becomes idle

during the second service stage of a currently served call, she cannot start to serve a new waiting call from the queue, if any. Such an overlap between two or more different call treatments would necessarily disturb the agent, which leads to errors and work inefficiency (Gladstones et al., 1989).

Probabilistic Model (Model PM): We distinguish the two situations when the agent is available to handle emails between two call conversations, or inside a call conversation.

Between two calls: just after a call service completion (as soon as the third stage finishes) and no waiting calls are in the queue, the agent treats one or more emails with probability p (independently of any other event), or does not work on emails at all with probability $1-p$. In the latter case, the agent simply remains idle and waits for a new call arrival to handle it. In the former case (with probability p), she selects a first email to work on. After finishing the treatment of this email, there are two cases: either a new call has already arrived and it is now waiting in the queue, or the queue of calls is still empty. If a call has arrived, the agent handles that call. If not, she selects another email, and so on. At some point in time, a new call would arrive while the agent is working on an email. The agent will then handle the call as soon as she finishes the email treatment.

Inside a call: Just after the end of the first stage of an underway call service (regardless wether there are other waiting calls in the queue or not), the agent treats one or more emails with probability q (independently of any other event), or does not work on emails at all with probability $1-q$. In the latter case, the agent simply remain idle and waits for the currently served customer to finish her operations on her own (corresponding to the second call service stage, i.e., the agent break). As soon as the customer finishes on herself her second service stage, the agent starts the third and last service stage. In the former case (with probability q), she selects a first email to work on. After finishing the treatment of this email, there are two cases: either the currently served customer has already finished her second service stage, or not yet. If she does, the agent starts the third stage of the customer call service. If not, she selects another email, and so on. At some point in time, the

Table 3.1: Particular cases of Model PM

Model	Description
Model 1	$p = q = 0$, no treatment of emails
Model 2	$p = 1$ and $q = 0$, systematic treatment of emails only between two calls
Model 3	$p = 0$ and $q = 1$, systematic treatment of emails only during the break
Model 4	$p = q = 1$, systematic treatment of emails between two calls and during the break

currently served call would finish her second service while the agent is working on an email. The agent will then handle the call as soon as she finishes the the email treatment.

We further consider next 4 particular cases of Model PM as shown in Table 3.1. Although these models might appear to be too restrictive to solve Problem (3.1), we show later their merit in Section 3.5.2 when we focus on the optimization of p and q in Model PM. Moreover, they have the advantage of being easy to implement in practice, easy to understand by managers, and easy to follows by agents. Note that in Model 1, the throughput of emails is zero. The interest from Model 1 is in the extreme case of a very high workload of calls or a very restrictive constraint on the call waiting time.

3.4 Performance Analysis

In this section we provide an exact method to characterize the call waiting time in the queue and the email throughput for Model PM (Section 3.4.1) and its extreme cases (Section 3.4.2). Our approach consists on using a Markov chain model to describe the system states and compute their steady-state probabilities. The computation of some of the steady-state probabilities involves the resolution of cubic (third degree) or quartic (fourth degree) equations for which we use the Cardan-Ferrari method.

3.4.1 Model PM

Let us define the random process $\{(x(t), y(t)), t \geq 0\}$ where $x(t)$ and $y(t)$ denote the state of the agent and the number of waiting calls in the queue at a given time $t \geq 0$, respectively. We have

$y(t) \in \{0, 1, 2, ...\}$, for $t \geq 0$. The possible values of $x(t)$ (corresponding to the possible states of the agent), for $t \geq 0$, are

- "Agent working on the first stage of a call service" denoted by $x(t) = A$,

- "Idle agent that is waiting for the call to finish her second stage of service" denoted by $x(t) = B$,

- "Agent working on an email while an underway call has already finished her second stage of service and is waiting for the agent to start her third stage of service" denoted by $x(t) = B'$,

- "Agent working on the third stage of a call service" denoted by $x(t) = C$,

- "Agent working on an email between two call conversations" denoted by $x(t) = M$,

- "Agent idle between two call conversations" denoted by $x(t) = 0$.

Since call inter-arrival times, call service times in each stage, and email service times are exponentially distributed, $\{(x(t), y(t)), t \geq 0\}$ is a Markov chain (Figure 3.1).

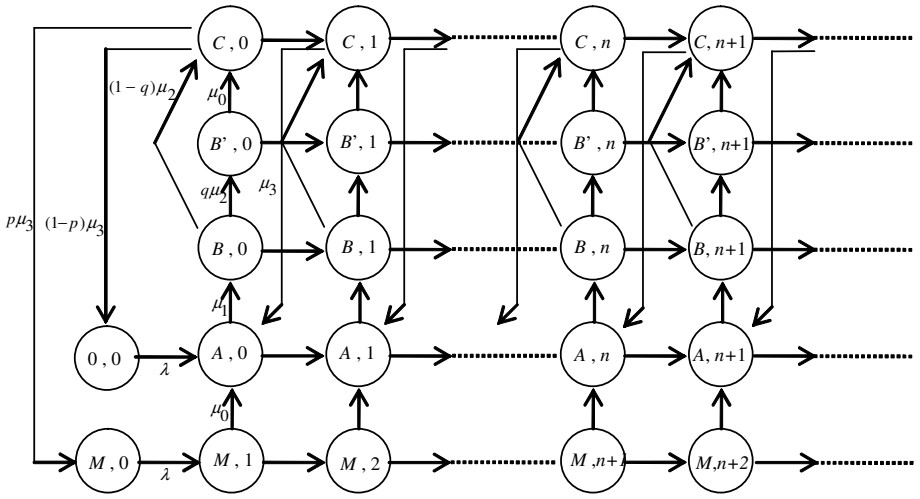


Figure 3.1: Markov chain for Model PM

For ease of exposition, we denote by P_0 the probability to be in state $(0, 0)$, and for $n \geq 0$ we denote by a_n, b_n, b'_n, c_n and m_n the probabilities to be in state (A, n), (B, n), (B', n), (C, n) and (M, n), respectively. We also define $\rho_i = \frac{\lambda}{\mu_i}$, for $i \in \{0, 1, 2, 3\}$. In Proposition 2, we give the probability of delay of a call (probability of waiting) denoted by P_D and the throughput of emails denoted by T. Note that the stability condition of Model PM is $\lambda < \frac{q}{\mu_0} + \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}$.

Proposition 2 *For Model PM, we have*

$$P_D = 1 - \frac{1-p}{1+p\rho_0}(1-\rho_1-\rho_2-q\rho_0-\rho_3),$$
$$T = \mu_0\left(\frac{1+\rho_0}{1+p\rho_0}p\,(1-\rho_1-\rho_2-q\rho_0-\rho_3)+q(\rho_2+\rho_0)\right).$$

Proof. From the Markov chain of Model PM, we have

$$c_0 = \rho_3(P_0+m_0),$$
$$c_n = \rho_3(a_{n-1}+b_{n-1}+b'_{n-1}+c_{n-1}+m_n),$$

for $n \geq 1$. Then

$$\sum_{n=0}^{\infty} c_n = \rho_3\left\{P_0+m_0+\sum_{n=0}^{\infty}\left(a_n+b_n+b'_n+c_n\right)+\sum_{n=1}^{\infty}m_n\right\}. \tag{3.2}$$

Since all system state probabilities sum up to 1, i.e., $P_0+\sum_{n=0}^{\infty}(a_n+b_n+b'_n+c_n+m_n)=1$, Equation (3.2) becomes

$$\sum_{n=0}^{\infty} c_n = \rho_3. \tag{3.3}$$

For the state $(M,0)$, we have $p\mu_3 c_0 = \lambda m_0$, or equivalently $c_0 = \rho_3\frac{m_0}{p}$. Therefore $c_0 = \rho_3\frac{P_0}{1-p}$. We then may write

$$P_0 = \frac{1-p}{p}m_0. \tag{3.4}$$

From the Markov chain, we also have $\mu_2\sum_{n=0}^{\infty}b_n = \mu_3\sum_{n=0}^{\infty}c_n = \mu_0\sum_{n=0}^{\infty}b'_n+(1-q)\mu_2\sum_{n=0}^{\infty}b_n =$

$\mu_1 \sum_{n=0}^{\infty} a_n$. Using Equation (3.3), we then obtain

$$\sum_{n=0}^{\infty} b_n = \rho_2, \ \sum_{n=0}^{\infty} b'_n = q\rho_0, \sum_{n=0}^{\infty} a_n = \rho_1. \tag{3.5}$$

For state (M, n), $n \geq 1$, we have $m_n = (\frac{\rho_0}{1+\rho_0})^n m_0$. Therefore $\sum\limits_{i=0}^{\infty} m_i = m_0(1+\rho_0)$. Using now Equation (3.5) together with the normalization condition implies $m_0 = \frac{p}{1+p\rho_0}(1-\rho_1-\rho_2-\rho_3-q\rho_0)$, and Equation (3.4) then becomes

$$P_0 = \frac{1-p}{1+p\rho_0}(1-\rho_1-\rho_2-q\rho_0-\rho_3)).$$

A new call enters service immediately upon arrival, if and only if the system is in state $(0, 0)$. Since the call arrival process is Poison, we use the PASTA property to state that the steady-state probabilities seen by a new call arrival coincide with those seen at an arbitrary instant. Thus $P_D = 1 - P_0$, or

$$P_D = 1 - \frac{1-p}{1+p\rho_0}(1-\rho_1-\rho_2-\rho_3-q\rho_0).$$

As for the email throughput, it is given by $T = \mu_0 \left(q \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty} b'_i + \sum_{i=0}^{\infty} m_i\right)$, which may be also written as $T = \mu_0 \left(\frac{1+\rho_0}{1+p\rho_0} p \left(1-\rho_1-\rho_2-\rho_3-q\rho_0\right) + q(\rho_2+\rho_0)\right)$. This finishes the proof of the proposition. □

Let us now define W, a random variable, as the call waiting time in the queue, and $\mathrm{P}(W < t)$ as its cumulative distribution function (cdf) for $t \geq 0$. Conditioning on a state seen by a new call

arrival and averaging over all possibilities, we state using PASTA that

$$\mathrm{P}(W < t) = P_0 \cdot 1 + \sum_{n=0}^{+\infty} (\mathrm{P}(W < t, (A, n)) \cdot a_n + \mathrm{P}(W < t, (B, n)) \cdot b_n + \mathrm{P}(W < t, (B', n)) \cdot b'_n)$$

$$+ \mathrm{P}(W < t, (C, n)) \cdot c_n + \mathrm{P}(W < t, (M, n)) \cdot m_n). \tag{3.6}$$

For $n \geq 0$, the quantities $\mathrm{P}(W < t, (A, n))$, $\mathrm{P}(W < t, (B, n))$, $\mathrm{P}(W < t, (B', n))$, $\mathrm{P}(W < t, (C, n))$ and $\mathrm{P}(W < t, (M, n))$ are the cdf of the conditional call waiting times in the queue, given that a new arriving call finds the system in states (A, n), (B, n), (B', n), (C, n) and (M, n), respectively. In the Markov chain of Model PM, these conditional random variables correspond to first passage times to state $(0, 0)$ starting from the system state upon a new call arrival. They are convolutions of independent exponential random variables with arbitrarily rates, not necessarily all equal (Erlang random variable) or all distinct. Using the results in Amari and Misra (1997), we can explicitly derive the expressions of $\mathrm{P}(W < t, (A, n))$, $\mathrm{P}(W < t, (B, n))$, $\mathrm{P}(W < t, (B', n))$, $\mathrm{P}(W < t, (C, n))$ and $\mathrm{P}(W < t, (M, n))$, for $n \geq 0$, as shown in Section B.2 of the Appendix.

It remains now to compute the probabilities a_n, b_n, b'_n, c_n and m_n in n, for $n \geq 0$. From the Markov chain of Model PM, we can write the following iterative equations

$$\lambda X_{n-1} = A X_n, \tag{3.7}$$

for $n \geq 1$, where

$$X_n = \begin{pmatrix} a_n \\ b_n \\ b'_n \\ c_n \\ m_n \end{pmatrix},$$

for $n \geq 0$ is the vector of probabilities to be computed and

$$A = \begin{pmatrix} \mu_1 & -\lambda & -\lambda & -\lambda & -\lambda \\ -\mu_1 & \lambda+\mu_2 & 0 & 0 & 0 \\ 0 & -q\mu_2 & \lambda+\mu_0 & 0 & 0 \\ 0 & -(1-q)\mu_2 & -\mu_0 & \lambda+\mu_3 & 0 \\ 0 & 0 & 0 & 0 & \lambda+\mu_0 \end{pmatrix}.$$

The first step to solve Equation (3.7) is to find the eigenvalues of the matrix $\frac{1}{\lambda}A$. These are solutions of the equation $\det(\frac{1}{\lambda}A - yI) = 0$ with y as variable. One obvious eigenvalue is $1 + \frac{1}{\rho_0}$ (see the last line of A), and the remaining ones are those of a 4×4 matrix (derived from $\frac{1}{\lambda}A$ by removing the last line and the last column) and they are solutions of the following quadric equation

$$\sigma_4 y^4 - (3\sigma_4 + \sigma_3)y^3 + (3\sigma_4 + 2\sigma_3 + \sigma_2)y^2 - (\sigma_4 + \sigma_3 + \sigma_2 + \sigma_1)y + 1 + \rho_0(1-q) = 0, \qquad (3.8)$$

with y as variable, $\sigma_1 = \rho_0 + \rho_1 + \rho_2 + \rho_3$, $\sigma_2 = \rho_0\rho_1 + \rho_0\rho_2 + \rho_0\rho_3 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3$, $\sigma_3 = \rho_0\rho_1\rho_2 + \rho_0\rho_1\rho_3 + \rho_0\rho_2\rho_3 + \rho_1\rho_2\rho_3$, and $\sigma_4 = \rho_0\rho_1\rho_2\rho_3$. Since the constant term $1 + \rho_0(1-q)$ in Equation (3.8) is strictly positive, zero cannot be a solution of that equation. Then, $\frac{1}{\lambda}A$ is invertible. Therefore the eigenvalues of λA^{-1} are solutions of

$$(1 + \rho_0(1-q))x^4 - (\sigma_4 + \sigma_3 + \sigma_2 + \sigma_1)x^3 + (3\sigma_4 + 2\sigma_3 + \sigma_2)x^2 - (3\sigma_4 + \sigma_3)x + \sigma_4 = 0, \qquad (3.9)$$

where $x = \frac{1}{y}$. We solve the quadric Equation (3.9) using the Cardan-Ferrari method. In Section B.4 of the Appendix, we describe the details of this method.

The explicit expressions of the probability components of the vector X_n, for $n \geq 0$, can be derived, however they are too cumbersome for Model PM. We go further in providing their expressions for the extreme cases of Model PM in Section 3.4.2 and also using a light-traffic approximation in Section 3.6.1. In all cases, an exact numerical method is straightforward and easy to implement.

Numerical illustrations are shown later in Section 3.5.

Let us now compute the expected call waiting time in Model PM, denoted by $E(W)$. Consider first a model similar to Model PM except that emails can only be treated inside a call conversation. We denote this model by Model PM'. With a little thought, one can see that the expected call waiting time in Model PM is that of Model PM' plus $\frac{p}{\mu_0}$. This can be easily proven using the memoryless property of the email service duration and sample path arguments. The main idea of the proof is as follows. Consider each first call of the busy periods of Model PM'. The same calls arrive also at Model PM but not necessarily enter immediately service as in Model PM'. Each one of these calls in Model PM, will arrive either at a system that is empty of calls, or not. In the first case and with probability p, she will be delayed compared to Model PM' by the residual duration of an email treatment (exponential with rate μ_0). All the calls arriving after her (and seeing a system non-empty of calls) will be delayed by the same amount of time. In the second case, if the call arrives at a system non-empty of calls, then this means that the previous busy period of calls in Model PM has been delayed by an amount of time corresponding to the residual time of an email treatment. Then this call and all of those who arrive after her and see a system non-empty of calls will be delayed with the same amount of time, and so on and so forth.

Let us now compute the expected waiting time in Model PM', denoted by $E(W')$. We use the Pollaczeck-Kinchin result for an M/G/1 queue. From Pollaczeck (1930), we have $E(W') = \frac{\rho^2(1+cv^2)}{2\lambda(1-\rho)}$, where cv is the coefficient of variation of the service distribution (its standard deviation over its expected value) and ρ is the equivalent workload. Because of the possibility to do emails between calls, the random variable representing the service time duration, say S, can be written as $S = S_1 + S_2 + US_0 + S_3$, where S_i, a random variable, follows an exponential distribution with rate μ_i, for $i = 0, ..., 3$, and U follows a binomial distribution with parameter q. We denote by $E(Z)$ and $V(Z)$ the expected (first moment) and variance of a given random variable Z. The first moment

of S is given by

$$E(S) = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{q}{\mu_0} + \frac{1}{\mu_3},$$

and its variance can be written as (using the independence between S_i and S_j for $i \neq j \in \{0, ..., 3\}$)

$$V(S) = V(S_1) + V(S_2) + V(US_0) + V(S_3) = \frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{2q - q^2}{\mu_0^2} + \frac{1}{\mu_3^2}.$$

Then

$$cv^2(S) = \frac{\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{2q-q^2}{\mu_0^2} + \frac{1}{\mu_3^2}}{\left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{q}{\mu_0} + \frac{1}{\mu_3}\right)^2}.$$

After some algebra, we obtain

$$E(W') = \frac{(\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2 + 2q\rho_0(\rho_0 + \rho_1 + \rho_2 + \rho_3)}{2\lambda(1 - (\rho_1 + \rho_2 + q\rho_0 + \rho_3))},$$

which leads to

$$E(W) = \frac{p}{\mu_0} + \frac{(\rho_1 + \rho_2 + \rho_3)^2 + \rho_1^2 + \rho_2^2 + \rho_3^2 + 2q\rho_0(\rho_0 + \rho_1 + \rho_2 + \rho_3)}{2\lambda(1 - (\rho_1 + \rho_2 + q\rho_0 + \rho_3))}. \tag{3.10}$$

This closes the performance measure analysis of Model PM.

3.4.2 Extreme Cases

We consider the 4 extreme cases of Model PM; Models 1,...,4. To derive the expressions of the email throughput, the call probability of delay, and the call expected waiting time, we simply apply the analysis of Section 3.4.1 and state the results as shown in Table 3.2.

One can derive the cdf of the call waiting time $\mathrm{P}(W < t)$ using Equation (3.6). For $n \geq 0$,

Table 3.2: Expressions of T, $E(W)$ and P_D for Models 1,...,4

	Model 1	Model 2
T	0	$\mu_0(1-\rho_1-\rho_2-\rho_3)$
$E(W)$	$\frac{(\rho_1+\rho_2+\rho_3)^2(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2}{(\rho_1+\rho_2+\rho_3)^2})}{2\lambda(1-\rho_1-\rho_2-\rho_3)}$	$\frac{1}{\mu_0}+E(W_1)$
P_D	$\rho_1+\rho_2+\rho_3$	1
	Model 3	**Model 4**
T	$\mu_0(\rho_0+\rho_2)$	$\mu_0(1-\rho_1-\rho_3)$
$E(W)$	$\frac{(\rho_1+\rho_2+\rho_3+\rho_0)^2(1+\frac{\rho_1^2+\rho_2^2+\rho_3^2+\rho_0^2}{(\rho_1+\rho_2+\rho_3+\rho_0)^2})}{2\lambda(1-\rho_1-\rho_2-\rho_3-\rho_0)}$	$\frac{1}{\mu_0}+E(W_3)$
P_D	$\rho_0+\rho_1+\rho_2+\rho_3$	1

the quantities $\mathrm{P}(W < t,(A,n))$, $\mathrm{P}(W < t,(B,n))$, $\mathrm{P}(W < t,(B',n))$, $\mathrm{P}(W < t,(C,n))$ and $\mathrm{P}(W < t,(M,n))$ can be derived using again the results of Amari and Misra (1997) as shown in Section B.2 of the appendix. Fortunately, the computation of the probabilities a_n, b_n, b'_n, c_n and m_n, for $n \geq 0$, simplifies further. In what follows, we avoid the matrix analysis (as developed in Section 3.4.1) by providing for each model the expression of each one of the latter probabilities as a function of u_n, where $u_n = a_n + b_n + b'_n + c_n$, for $n \geq 0$. We then show that u_n, for $n \geq 0$, satisfies a recurrent cubic or quartic linear equation that we solve using the Cardan-Ferrari method. Another method to compute $\mathrm{P}(W < t)$ is this case (service in three exponential stages) based on the Dunford decomposition is developed in Section B.3 of the Appendix.

Model 1: The Markov chain associated to Model 1 is given in Figure 3.2. For this model, we have $u_n = a_n + b_n + c_n$, for $n \geq 0$.

From the Markov chain of Model 1, we may write $c_n(\lambda+\mu_3) = \lambda c_{n-1} + \mu_2 b_n$ for $n \geq 1$, and $c_n = \rho_3 u_{n-1}$ for $n \geq 1$. This leads to $b_n = \rho_2\left((1+\rho_3)u_{n-1} - \rho_3 u_{n-2}\right)$ for $n \geq 2$. From the Markov chain, we also may write $b_n(\lambda+\mu_2) = \lambda b_{n-1} + \mu_1 a_n$ for $n \geq 1$. Combining the last two equations yields to $a_n = \rho_1(((1+\rho_3)\rho_2 + \rho_3 + 1)u_{n-1} - (\rho_2(1+2\rho_3)+\rho_3)u_{n-2} + \rho_3 u_{n-3})$ for $n \geq 3$. The

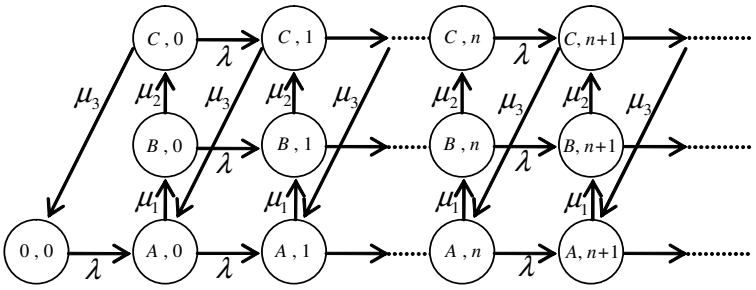


Figure 3.2: Markov chain for Model 1

equation $u_n = a_n + b_n + c_n$ for $n \geq 0$ is then equivalent to

$$u_{n+3} = ((1+\rho_1)(1+\rho_2)(1+\rho_3) - 1)\, u_{n+2} - (\rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3 + 2\rho_1\rho_2\rho_3)\, u_{n+1} + \rho_1\rho_2\rho_3 u_n,$$

which leads to

$$u_{n+3} = (\sigma_1 + \sigma_2 + \sigma_3)\, u_{n+2} - (\sigma_2 + 2\sigma_3)\, u_{n+1} + \sigma_3 u_n, \tag{3.11}$$

where $\sigma_1 = \rho_1 + \rho_2 + \rho_3$, $\sigma_2 = \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3$, and $\sigma_3 = \rho_1\rho_2\rho_3$. The cubic equation associated to the recurrent Equation (3.11) is $x^3 = (\sigma_1 + \sigma_2 + \sigma_3)\, x^2 - (\sigma_2 + 2\sigma_3)\, x + \sigma_3$, with x as variable. It remains now to apply the Cardan-Ferrari method in order compute u_n, for $n \geq 0$. The type of the solutions depends on the discriminant Δ of the cubic equation (see Section B.4 of the Appendix).

- If $\Delta > 0$, the cubic equation has one real solution denoted by x_1 and two complex solutions x_2 and its conjugate. We denote by $|x_2|$ the module of x_2 and θ its argument. Since u_n is real, it can be written as

  $$u_n = r x_1^n + s|x_2|^n cos(n\theta) + t|x_2|^n sin(n\theta),$$

  for $n \geq 0$, where the parameters $r, s, t \in \mathbb{R}$. These parameters are easily computed using the initial conditions given by u_0, u_1 and u_2.

- If $\Delta < 0$, the cubic equation has three real solutions denoted by x_1, x_2 and x_3. Thus

$$u_n = rx_1^n + sx_2^n + tx_3^n,$$

for $n \geq 0$, where again the parameters r, s and t are computed from the initial conditions.

- If $\Delta = 0$, the cubic equation has a simple solution denoted by x_1 and a double one denoted by x_2. We then have

$$u_n = rx_1^n + (sn + t)x_2^n,$$

for $n \geq 0$, and the real parameters r, s and t are again given by the initial conditions.

Model 2: The Markov chain associated to Model 2 is given in Figure 3.3. For this model, we have $u_n = a_n + b_n + c_n$, for $n \geq 0$. Following the same reasoning as that for Model 1, we obtain

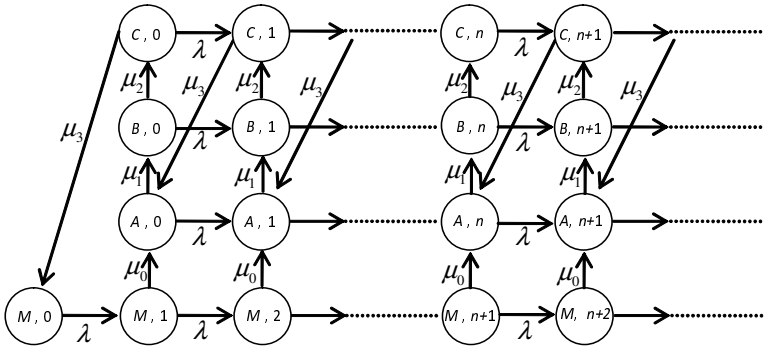


Figure 3.3: Markov chain for Model 2

$$u_{n+3} = [\sigma_1 + \sigma_2 + \sigma_3]u_{n+2} - [\sigma_2 + 2\sigma_3]u_{n+1} + \sigma_3 u_n + K\left(\frac{\rho_0}{1+\rho_0}\right)^n, \tag{3.12}$$

with the constant $K \in \mathbb{R}$. Since Equation (3.12) is similar to Equation (3.11) in the sense that the former has just an additional term proportional to $\left(\frac{\rho_0}{1+\rho_0}\right)^n$, we again use the solutions of Equation

(3.11) to easily obtain those of Equation (3.12).

Model 3: The Markov chain associated to Model 3 is given in Figure 3.4. For this model, we have $u_n = a_n + b_n + b'_n + c_n$, for $n \geq 0$.

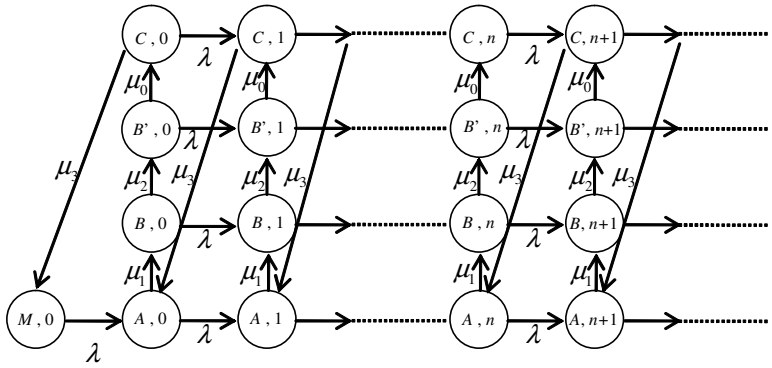


Figure 3.4: Markov chain for Model 3

From the Markov chain of Model 3, we may write $c_n(\lambda + \mu_3) = \lambda c_{n-1} + \mu_0 b'_n$ for $n \geq 1$, and $c_n = \rho_3 u_{n-1}$ for $n \geq 1$. So, $b'_n = \rho_0(1 + \rho_3)u_{n-1} - \rho_0\rho_3 u_{n-2}$ for $n \geq 2$. We also have from the Markov chain $b'_n(\lambda + \mu_0) = \lambda b'_{n-1} + \mu_2 b_n$, for $n \geq 1$. Therefore, $b_n = \rho_2(1 + \rho_0)(1 + \rho_3)u_{n-1} - \rho_2(1 + \rho_0)\rho_3 u_{n-2} - \rho_2\rho_0(1 + \rho_3)u_{n-2} + \rho_2\rho_0\rho_3 u_{n-3}$, for $n \geq 3$. From a state (B, n) for $n \geq 1$, we may write $b_n(\lambda + \mu_2) = \lambda b_{n-1} + \mu_1 a_n$, which leads to

$$
\begin{aligned}
a_n = {} & \rho_1(1 + \rho_2)(1 + \rho_0)(1 + \rho_3)u_{n-1} - \rho_1(1 + \rho_2)(1 + \rho_0)\rho_3 u_{n-2} - \rho_1(1 + \rho_2)\rho_0(1 + \rho_3)u_{n-2} \\
& + \rho_1(1 + \rho_2)\rho_0\rho_3 u_{n-3} - \rho_1\rho_2(1 + \rho_0)(1 + \rho_3)u_{n-2} + \rho_1\rho_2(1 + \rho_0)\rho_3 u_{n-3} + \rho_1\rho_2\rho_0(1 + \rho_3)u_{n-3} \\
& - \rho_1\rho_2\rho_0\rho_3 u_{n-4},
\end{aligned}
$$

for $n \geq 4$. From $u_n = a_n + b_n + b'_n + c_n$ we then state that

$$\begin{aligned} u_{n+4} = (&\rho_0 + \rho_1 + \rho_2 + \rho_3 + \rho_0\rho_1 + \rho_0\rho_2 + \rho_0\rho_3 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3 + \rho_0\rho_1\rho_2 \\ &+ \rho_0\rho_1\rho_3 + \rho_0\rho_2\rho_3 + \rho_1\rho_2\rho_3 + \rho_0\rho_1\rho_2\rho_3)u_{n+3} - (\rho_0\rho_1 + \rho_0\rho_2 + \rho_0\rho_3 + \rho_1\rho_2 \\ &+ \rho_1\rho_3 + \rho_2\rho_3 + 3(\rho_0\rho_1\rho_2 + \rho_0\rho_1\rho_3 + \rho_0\rho_2\rho_3 + \rho_1\rho_2\rho_3) + 3\rho_0\rho_1\rho_2\rho_3)u_{n+2} \\ &+ (\rho_0\rho_1\rho_2 + \rho_0\rho_1\rho_3 + \rho_0\rho_2\rho_3 + \rho_1\rho_2\rho_3 + 2\rho_0\rho_1\rho_2\rho_3)u_{n+1} - \rho_0\rho_1\rho_2\rho_3 u_n, \end{aligned}$$

for $n \geq 0$, or equivalently

$$u_{n+4} = [\sigma_1 + \sigma_2 + \sigma_3 + \sigma_4]u_{n+3} - [\sigma_2 + 2\sigma_3 + 3\sigma_4]u_{n+2} + [\sigma_3 + 3\sigma_4]u_{n+1} - \sigma_4 u_n, \tag{3.13}$$

for $n \geq 0$, where $\sigma_1 = \rho_0 + \rho_1 + \rho_2 + \rho_3$, $\sigma_2 = \rho_0\rho_1 + \rho_0\rho_2 + \rho_0\rho_3 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3$, $\sigma_3 = \rho_0\rho_1\rho_2 + \rho_0\rho_1\rho_3 + \rho_0\rho_2\rho_3 + \rho_1\rho_2\rho_3$ and $\sigma_4 = \rho_0\rho_1\rho_2\rho_3$. In order to compute u_n we then solve the associated quadric equation using the Cardan-Ferrari method and the initial conditions u_0, u_1, u_2 and u_3.

Model 4: The Markov chain associated to Model 4 is given in Figure 3.5. For this model, we have $u_n = a_n + b_n + b'_n + c_n$, for $n \geq 0$.

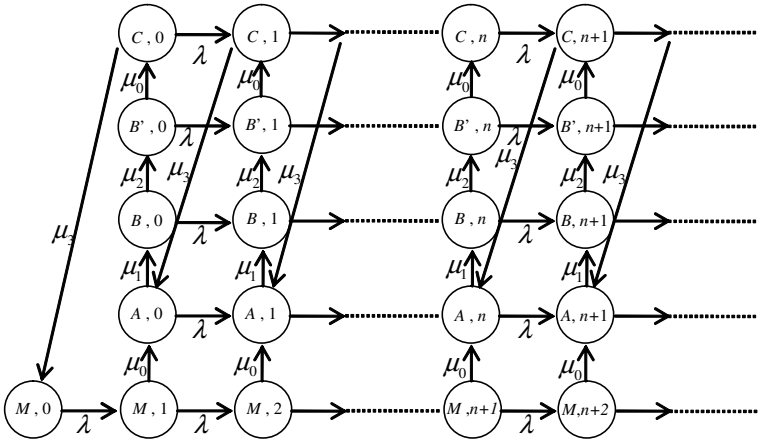


Figure 3.5: Markov chain for Model 4

Following the same reasoning as that for Model 3, we obtain

$$u_{n+4} = [\sigma_1 + \sigma_2 + \sigma_3 + \sigma_4]\, u_{n+3} - [\sigma_2 + 2\sigma_3 + 3\sigma_4]u_{n+2} + [\sigma_3 + 3\sigma_4]u_{n+1} - \sigma_4 u_n + K\left(\frac{\rho_0}{1+\rho_0}\right)^n, \tag{3.14}$$

with the constant $K \in \mathbb{R}$. Since Equation (3.14) is similar to Equation (3.13) in the sense that the former has just an additional term proportional to $\left(\frac{\rho_0}{1+\rho_0}\right)^n$, we again use the solution of Equation (3.13) to easily obtain those of Equation (3.14). This closes the characterization of the performance measures for the special cases Models 1,...,4.

3.5 Comparison Analysis and Insights

We start in Section 3.5.1 by a comparison analysis between the extreme cases Models 1,...,4. The comparison is based on the optimization problem (3.1). We derive various structural results and properties for this comparison. In particular, we investigate the impact of the mean arrival rate intensity of calls on the comparison between Models 1,...,4. One could think of a call center manager that adjusts the job routing schema as a function of the call arriving workload over the day. In Section 3.5.2 we focus on the general case Model PM. We prove that the optimization of the parameters of Model PM lead to extreme situations in the sense of a systematic email treatment of emails either between calls or inside a call conversation, which gives an interest in practice for Models 1,...,4.

3.5.1 Comparison between the Extreme Cases

We first compare between Models 1,...,4 based on their performance in terms of the email throughput, denoted by T_1,..., T_4, respectively. It is obvious that Model 4 is the best and Model 1 is the worst (no emails at all). Let us now compare between Models 2 and 3. From Table 3.2 we have

$T_2 = \mu_0(1 - \rho_1 - \rho_2 - \rho_3)$ and $T_3 = \mu_0(\rho_0 + \rho_2)$. Thus $T_3 > T_2$ is equivalent to

$$\lambda > \frac{1}{\frac{1}{\mu} + \frac{1}{\mu_2}},$$

where $\frac{1}{\mu} = \frac{1}{\mu_0} + \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}$. Since the stability condition for Model 3 is $\lambda < \mu$, Model 3 is better than Model 2 if

$$\frac{1}{\frac{1}{\mu} + \frac{1}{\mu_2}} < \lambda < \mu. \tag{3.15}$$

Denoting the left term in Inequality (3.15) by R, the condition under which $T_3 > T_2$ is then

$$R = \frac{1}{\frac{1}{\mu_0} + \frac{1}{\mu_1} + \frac{1}{\mu_3} + \frac{2}{\mu_2}} < \lambda < \mu. \tag{3.16}$$

From Inequality (3.16), we first see that treating emails only inside a call conversation (Model 3) becomes better that treating them only between calls (Model 2) is likely the case for high arrival workloads (in such a case, busy period durations are reduced). We also see that $\frac{\partial R}{\partial \mu_2} > 0$ for $\mu_2 > 0$, $\frac{\partial R}{\partial \mu_0} > 0$ for $\mu_0 > 0$, $\frac{\partial R}{\partial \mu_1} > 0$ for $\mu_1 > 0$, and $\frac{\partial R}{\partial \mu_3} > 0$ for $\mu_3 > 0$. This means that decreasing the expected duration of the call service second stage ($1/\mu_2$) relative to the expected durations of the other call service stages or the email service duration ($1/\mu_1$, $1/\mu_3$ and $1/\mu_0$) increases the range of arrival workloads where it is preferred to use Model 2 instead of Model 3. In other word, there is no sufficient time to treat emails inside the call conversation.

Comparison with a constraint on $E(W)$

As a function of the mean call arrival rate, we want to answer the question when should we treat emails (which model among Models 1 to 4 should a manger choose?) for the following problem

$$\begin{cases} \text{Maximize } T \\ \text{subject to } E(W) \leq w^*, \end{cases} \tag{3.17}$$

where w^* is the service level for the expected waiting time, $w^* > 0$. Let W_i, a random variable, denote the expected call waiting time in Model i, $i = 1, ..., 4$. It is clear that for some periods of a working day with a very high call arrival rate λ, the manager is likely to choose Model 1 (no emails), and for other periods with a very low λ, she is likely to choose Model 4 (emails between calls and inside a call). However for intermediate values of λ, the optimal choice is not clear. This is what we analytically analyze in what follows.

Under the condition of stability of Model i, $E(W_i)$ is continuous and strictly increasing in λ (see Table 3.1), for $i = 1, ..., 4$. The constraint $E(W_i) \leq w^*$ is then equivalent to $\lambda \leq \overline{\lambda}_i$, for $i = 1, ..., 4$, where

$$\begin{aligned}
\overline{\lambda}_1 &= \frac{2w^*}{2w^*\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{3}\frac{1}{\mu_i^2}}, \\
\overline{\lambda}_2 &= \frac{2\left(w^* - \frac{1}{\mu_0}\right)}{2\left(w^* - \frac{1}{\mu_0}\right)\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{3}\frac{1}{\mu_i^2}}, \\
\overline{\lambda}_3 &= \frac{2w^*}{2w^*\left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=0}^{3}\frac{1}{\mu_i^2}}, \\
\overline{\lambda}_4 &= \frac{2\left(w^* - \frac{1}{\mu_0}\right)}{2\left(w^* - \frac{1}{\mu_0}\right)\left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right) + \left(\sum_{i=0}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=0}^{3}\frac{1}{\mu_i^2}}.
\end{aligned} \tag{3.18}$$

For a given λ and under the condition of stability of Model i ($i = 1, ..., 4$), the choice of Model i happens if $\lambda \leq \overline{\lambda}_i$ and $T_i = \max_{j \in \{1,...,4\},\ \lambda \leq \overline{\lambda}_j}(T_j)$. When $\lambda \leq \overline{\lambda}_4$, the choice is obviously for Model 4. When $\lambda \leq \overline{\lambda}_1$ and $\lambda > \overline{\lambda}_i$ for $i = 2, 3, 4$ the only possibility is Model 1. Proposition 3 provides the conditions under which an optimal choice of Model 2 or Model 3 may happen.

Proposition 3 *The following holds:*

1. *For $\lambda < \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}$, it exists some values of λ for which Model 2 is optimal if and only if $\overline{\lambda}_2 > 0$.*

2. *For* $\lambda < \frac{1}{\mu_0} + \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}$, *it exists some values of* λ *for which Model 3 is optimal if and only if*

$$\begin{cases} R \leq \overline{\lambda}_3 \\ \\ or \\ \\ \overline{\lambda}_2 < \overline{\lambda}_3. \end{cases}$$

3. *We have* $\overline{\lambda}_2 < \overline{\lambda}_3$ *if and only if* $\frac{1}{\mu_0} < w^* < \overline{w}^*$, *where*

$$\overline{w}^* = \frac{1}{2}\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3}\frac{1}{\mu_i^2} + 6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}} - \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right).$$

Proof. See Section C.2 of the appendix. □

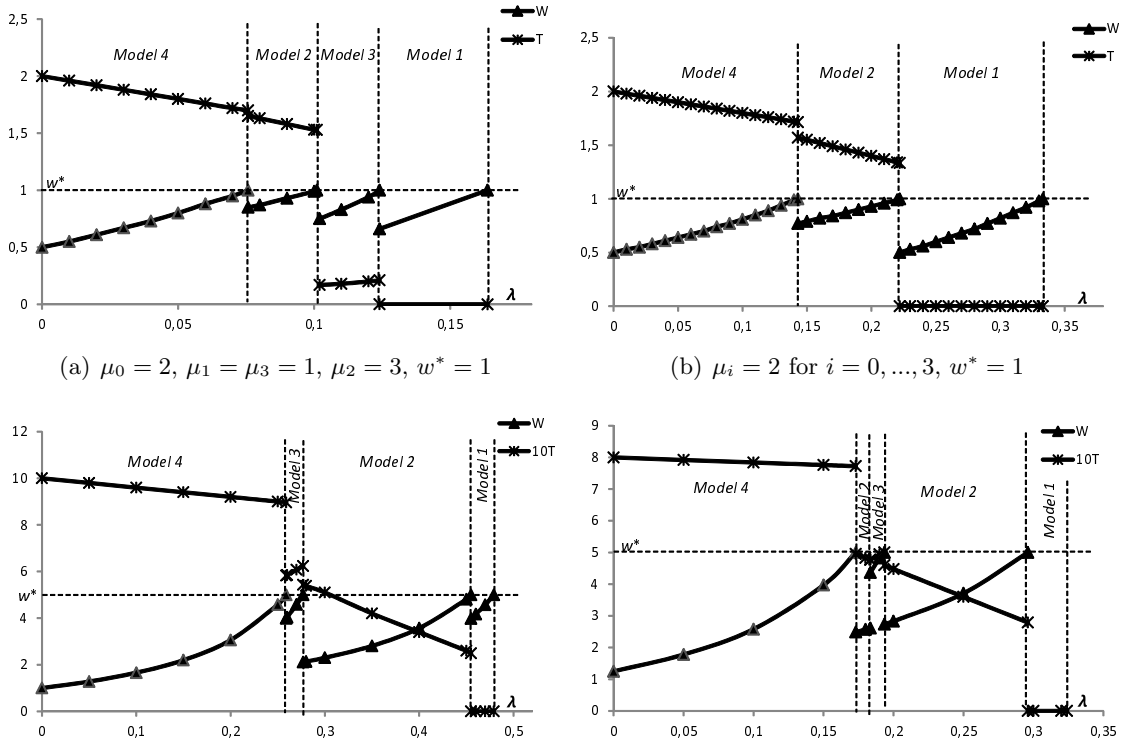
Using Equation (3.18), the condition in the first statement of Proposition 3 may be rewritten as

$$\begin{cases} w^* > \frac{1}{\mu_0} \\ \\ \text{or} \\ \\ w^* < \frac{1}{\mu_0} - \frac{\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{3}\frac{1}{\mu_i^2}}{2\left(\sum_{i=1}^{3}\frac{1}{\mu_i}\right)}. \end{cases} \tag{3.19}$$

The second inequality in Relation (3.19) implies $w^* < \frac{1}{\mu_0}$. Since at least the expected waiting time in Model 2 is strictly higher than $\frac{1}{\mu_0}$ (any new call has at least to wait for the residual time of an email treatment), this second inequality is impossible. Roughly speaking, the condition for the optimality of Model 2 (for some values of λ) holds when the service level on the call waiting is higher than the expected email service time.

In what follows, we numerically illustrate the analysis above. For various system parameters, Figure 3.6 gives the optimal model choice as a function of the mean arrival rate of calls, λ. An

intuitive reasoning of a manager would choose the ordering Model 4 (emails between calls and inside a call), then 2 (emails only between calls), then 3 (emails only inside a call), then 1 (no emails) as λ increases.



(a) $\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$, $w^* = 1$

(b) $\mu_i = 2$ for $i = 0, ..., 3$, $w^* = 1$

(c) $\mu_0 = 1$, $\mu_1 = \mu_3 = 5$, $\mu_2 = 0.8$, $w^* = 5$

(d) $\mu_0 = 0.8$, $\mu_1 = \mu_3 = 10$, $\mu_2 = 0.5$, $w^* = 5$

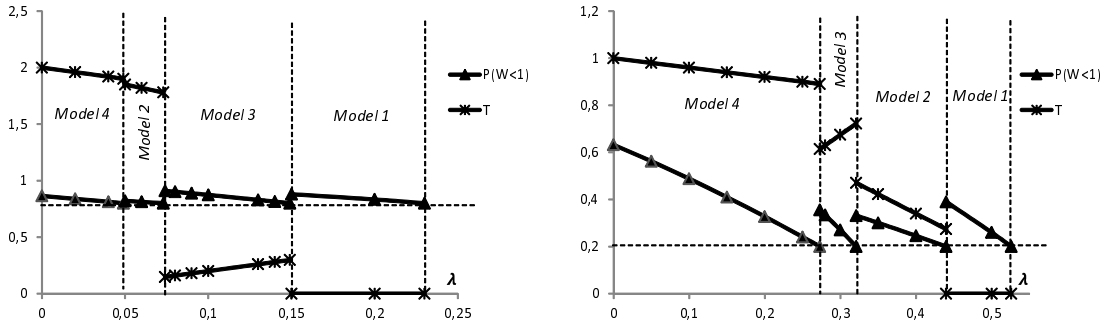
Figure 3.6: Comparison between Models 1 to 4 with a constraint on $E(W)$

The ordering Model 2 then Model 3 is not always appropriate, and some situations may require to consider some counterintuitive ordering. For instance, Model 3 is better than Model 2 for small values of λ if $R \leq \overline{\lambda}_4$ and $\overline{\lambda}_3 < \overline{\lambda}_2$, see Figure 3.6(c). In other words, this happens when the constraint on $E(W)$ is not too restrictive and when the expected second stage service duration is long. Another more surprising ordering, as λ increases, is Model 2, then Model 3, then again Model 2 (see Figure 3.6(d)) which happens for system parameters such that $\overline{\lambda}_4 < R < \overline{\lambda}_3 < \overline{\lambda}_2$.

Comparison with a constraint on $P(W < AWT)$

In the constraint of Problem (3.1), we want that the probability that a call waits less than a given threshold, defined as AWT is at least a given service level, defined as SL, i.e., $\mathrm{P}(W < AWT) \geq SL$. Note that a special case of this constraint is that on P_D, the call probability of waiting. The

expressions involved in the analysis of $P(W < AWT)$ are quite complicate to allow an analytical comparison between the models as we have done for a constraint on $E(W)$. We have then conducted a numerical comparison analysis (not totaly illustrated here). The main qualitative conclusions are similar to those for the case of a constraint on $E(W)$. As λ increases, it is not always true as one would intuitively expect that a manager should choose first Model 2 and then at some point of λ she shifts to Model 3 (Figure 3.7(a)). The optimal choice may change with the system parameters and we may have the ordering Model 3 then Model 2 (Figure 3.7(b)).



(a) $\mu_i = 2$ for $i = 0, ..., 3$ and $\alpha^* = 80\%$ (b) $\mu_0 = 1$, $\mu_1 = 5 = \mu_3$, $\mu_2 = 0.8$ and $\alpha^* = 20\%$

Figure 3.7: Comparison between Models 1 to 4 with the constraint on P$(W < 1)) \geq \alpha^*$

3.5.2 Optimization of Model PM

In this section we focus on the general case, Model PM. We are interested in the optimization of the parameters p and q in Model PM for Problem (3.1). Concretely, we want to find the optimal routing parameters of Model PM that allows the manager to maximize the email throughput while respecting a call service level constraint. Recall that the stability condition of Model PM is $\lambda < \frac{q}{\mu_0} + \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}$.

The expression of the email throughput T for Model PM is given in Proposition 2. It is straightforward to prove that for $p, q \in [0, 1]$ the maximum of T (best situation) is reached for $p = q = 1$. The proof is then omitted. Also, the expected call waiting time of Model PM (given in Equation (3.10)) is maximized (worst) for $p = q = 1$. Therefore in order to solve Problem (3.1), one would be interested analyzing the sensitivity of T with respect to p and q. In Lemma 1 we

prove a sensitivity result for T. The result will be used later in our analysis.

Lemma 1 *We have* $\frac{\partial T}{\partial p} > \frac{\partial T}{\partial q}$ *if and only if* $0 < \rho_0 < \overline{\rho}_0$, *where*

$$\overline{\rho}_0 = \frac{\sqrt{(p-q-\rho_1-\rho_2-\rho_3)^2 - 4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)} - q + p - (\rho_1+\rho_2+\rho_3)}{2(q-p^2+p)}.$$

Proof. See Section B.6 of the appendix. □

In what follows we address the question: starting from $p = q = 1$, in which direction should we decrease T? Should we decrease p or q first?

For $p = q = 1$, we have $\overline{\rho}_0 = \frac{1}{2}\left(\sqrt{(\rho_1+\rho_2+\rho_3)^2 + 4(1-\rho_1-\rho_2-\rho_3)} - (\rho_1+\rho_2+\rho_3)\right)$. Let us now prove (for $p = q = 1$) that $\overline{\rho}_0 > \rho_0$. From the one hand, proving $\overline{\rho}_0 > \rho_0$ is equivalent to proving $\sqrt{(\rho_1+\rho_2+\rho_3)^2 + 4(1-\rho_1-\rho_2-\rho_3)} > 2\rho_0 + (\rho_1+\rho_2+\rho_3)$ or equivalently $\rho_0^2 + \rho_0(\rho_1+\rho_2+\rho_3) - (1-(\rho_1+\rho_2+\rho_3)) < 0$ or also $(\rho_0+1)(\rho_0 - (1-(\rho_1+\rho_2+\rho_3)) < 0$. From the other hand, we have $\rho_0 + 1 > 0$. Also, the stability condition of Model 4 (Model PM with $p = q = 1$) is $\rho_0 + \rho_1 + \rho_2 + \rho_3 < 1$. Then $\rho_0 < 1 - (\rho_1+\rho_2+\rho_3)$. As a conclusion the inequality $\overline{\rho}_0 > \rho_0$ is true, for $p = q = 1$. Using Lemma 1, this means that starting from $p = q = 1$, we have $\frac{\partial T}{\partial p} > \frac{\partial T}{\partial q} > 0$. As a consequence, when we need to modify the values of p and q in order to decrease the expected call waiting time (the constraint in Problem (3.1)), the maximum of T is guarantied by first decreasing q (the email throughput is less sensitive to the variation of q than that of p). The question now is: which direction to use next? In other words when $p = 1$ and some value of q such that $0 < q < 1$, is it possible that it is better to decrease p instead of q? The answer is no and the proof is as follows. For $p = 1$, let us try to find a value of q for which we have $\overline{\rho}_0 \leq \rho_0$. This is equivalent to $\frac{\sqrt{(1-q-\rho_1-\rho_2-\rho_3)^2+4q(1-\rho_1-\rho_2-\rho_3)}-q+1-\rho_1-\rho_2-\rho_3}{2q} \leq \rho_0$. Thus, $q^2\rho_0^2 + q\rho_0 - (1-\rho_1-\rho_2-\rho_3)(1+\rho_0) > 0$. This trinomial in q has two real solutions; $q_1 = -\frac{1+\sqrt{4\rho_0+5-4(\rho_1+\rho_2+\rho_3)(\rho_0+1)}}{2\rho_0}$ and $q_2 = \frac{-1+\sqrt{4\rho_0+5-4(\rho_1+\rho_2+\rho_3)(\rho_0+1)}}{2\rho_0}$. Obviously $q_1 < 0$. We also have $q_2 > 1$ because: proving $q_2 - 1 > 0$ is equivalent to proving $\rho_0^2 + (\rho_1+\rho_2+\rho_3)\rho_0 + 1 > 0$.

The discriminant of this latter trinomial in ρ_0 is negative and it is equal to $(\rho_1+\rho_2+\rho_3)^2-4$. So $q_2>1$ for any $\rho_0>0$. Therefore it is impossible to find a value of q between 0 and 1 for which $0<\frac{\partial T}{\partial p}<\frac{\partial T}{\partial q}$. In conclusion starting from $p=q=1$, when we need to change the values of p and q, the best direction to maximize T is to first decrease q until $q=0$ and only then start to decrease p from $p=1$.

Consider now Problem (3.1) with a constraint on $E(W)$. From the one hand, the constraint $E(W)\leq w^*$ implies

$$\frac{p}{\mu_0}+\frac{(\rho_1+\rho_2+\rho_3)^2+\rho_1^2+\rho_2^2+\rho_3^2+2q\rho_0(\rho_0+\rho_1+\rho_2+\rho_3)}{2\lambda(1-(\rho_1+\rho_2+q\rho_0+\rho_3))}\leq w^*,$$

for $p,q\in[0,1]$, or equivalently

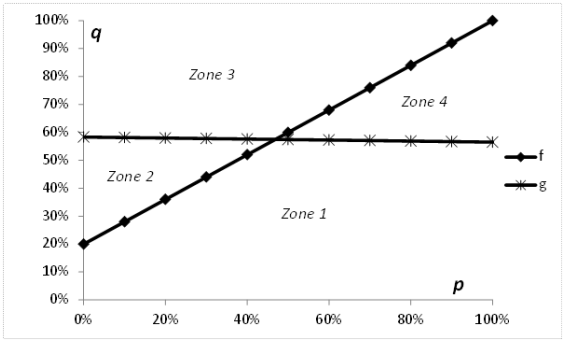
$$q\leq\frac{2\lambda(1-\rho_1-\rho_2-\rho_3)(w^*-p/\mu_0)-(\rho_1+\rho_2+\rho_3)^2-\rho_1^2-\rho_2^2-\rho_3^2}{2\rho_0(\rho_0+\rho_1+\rho_2+\rho_3+\lambda(w^*-p/\mu_0))},\tag{3.20}$$

for $p,q\in[0,1]$. From the other hand, the condition in Lemma 1, $0<\rho_0<\overline{\rho}_0$, is equivalent to

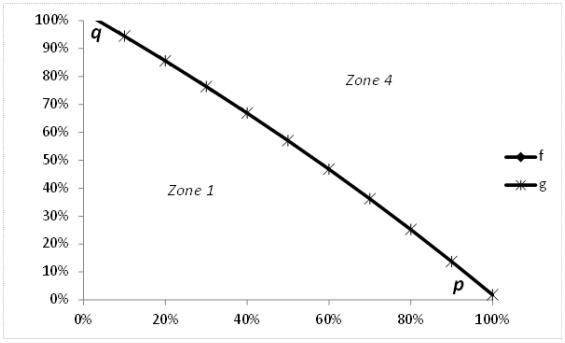
$$q<\frac{1-(\rho_1+\rho_2+\rho_3)(1+\rho_0)}{\rho_0(1+\rho_0)}+\frac{1-\rho_0}{1+\rho_0}p+\frac{\rho_0}{1+\rho_0}p^2,\tag{3.21}$$

for $p,q\in[0,1]$. Let us denote the right hand sides of Inequalities (3.20) and (3.21) by the functions in $p\in[0,1]$ $f(p)$ and $g(p)$, respectively. Illustrations of these functions are given in Figure 3.8.

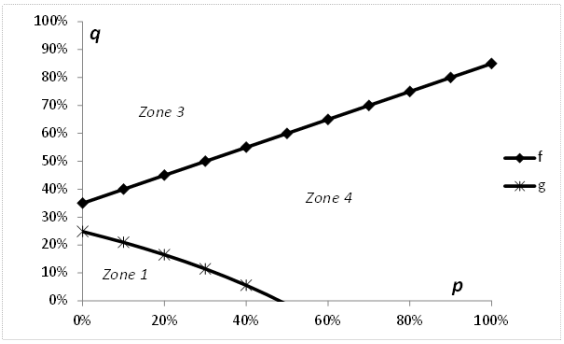
In what follows we prove an interesting result on the optimal values of p and q. Consider for example Figure 3.8(a) and assume that the agent is in a situation such that (p,q) belongs to Zone 1 or 2. Then the constraint on $E(W)$ is respected, but T can be improved. Using Lemma 1, we should increase p first (q first) for Zone 1 (Zone 2). From Figure 3.8(a), we also see that we should decrease p first (q first) for Zone 3 (Zone 4). It is clear that the optimal couple (p,q) will be on the curve of f. Moreover, we prove in Theorem 1 that the optimal point is such that $p\in\{0,1\}$ or



(a) $\lambda = 0.5$, $\mu_0 = \mu_1 = \mu_2 = \mu_3 = 2$, $w^* = 5$



(b) $\lambda = 0.22$, $\mu_0 = \mu_1 = \mu_2 = \mu_3 = 2$, $w^* = 1$



(c) $\lambda = 0.1$, $\mu_0 = 0.1$, $\mu_1 = \mu_2 = \mu_3 = 2$, $w^* = 5$

Figure 3.8: Behavior of $f(p)$ and $g(p)$

$q \in \{0, 1\}$.

Theorem 1 *For $p, q \in [0, 1]$, the optimal values of p and q of the optimization problem*

$$\begin{cases} \textit{Maximize } T \\ \textit{subject to } E(W) \leq w^*, \end{cases} \tag{3.22}$$

are always extreme values (0 or 1) for at least p or q.

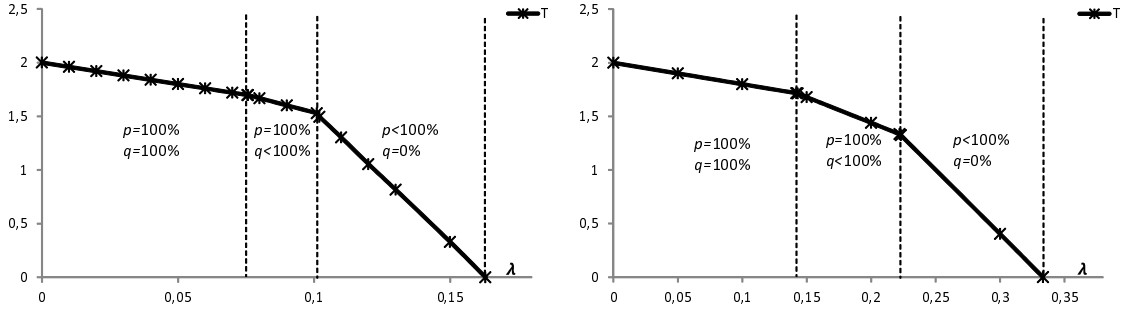
Proof. We want to maximize the email throughput $T(p, q)$ while respecting a constraint on the expected call waiting time ($E(W)(p, q) \leq w^*$). We use the method of Lagrange multipliers to find the optimal point (p, q). Let us denote the Lagrange multiplier by α (α is real). Then α and the extremum (p, q) of our optimization problem are solutions of the set of the 3 equations $D(T + \alpha(W - w^*)) = 0$, where D is the differential applicator in α, p and q. These 3 equations are

$$\frac{\partial(T + \alpha(W - w^*))}{\partial p} = \mu_0 \frac{(1 - \rho_1 - \rho_2 - q\rho_0 - \rho_3)(1 + \rho_0)}{(1 + p\rho_0)^2} + \alpha \frac{1}{\mu_0} = 0,$$

$$\frac{\partial(T+\alpha(W-w^*))}{\partial q} = \mu_0 \frac{(1-p)\rho_0}{1+p\rho_0}$$
$$+\alpha\frac{1}{2\lambda}\frac{\rho_0(2(\rho_0+\rho_1+\rho_2+\rho_3)(1-(\rho_1+\rho_2+\rho_3))+(\rho_1+\rho_2+\rho_3)^2+\rho_1^2+\rho_2^2+\rho_3^2}{(1-(\rho_1+\rho_2+\rho_3+q\rho_0))^2} = 0,$$

$$\frac{\partial(T+\alpha(W-w^*))}{\partial \alpha} = \frac{p}{\mu_0} + \frac{(\rho_1+\rho_2+\rho_3)^2+\rho_1^2+\rho_2^2+\rho_3^2+2q\rho_0(\rho_0+\rho_1+\rho_2+\rho_3)}{2\lambda(1-(\rho_1+\rho_2+q\rho_0+\rho_3))} - w^* = 0.$$

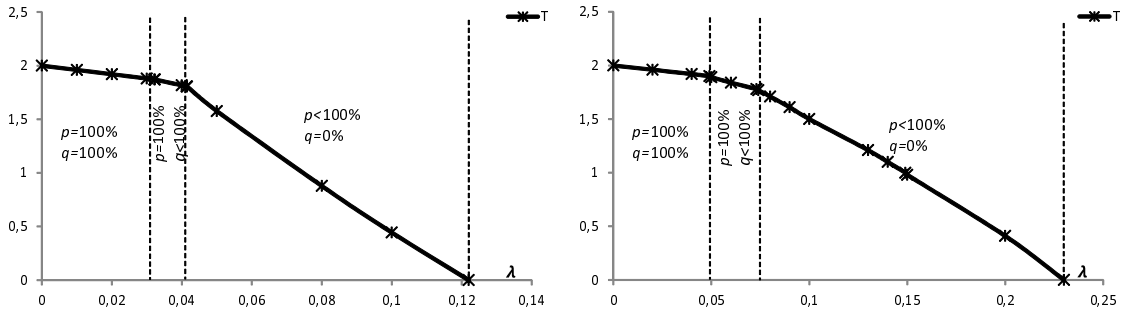
Solving this set of 3 equations leads to two couples of solutions (p_1, q_1) and (p_2, q_2). The expressions of these solutions are too long (see for example Section B.7 of the appendix for the expression of q_2), but we show for any case of system parameters ρ_i $(i = 0, ..., 3)$ under the condition of stability that all the values of p_1, q_1, p_2 and q_2 are out of the interval $[0, 1]$. We then deduce for the optimal values of p and q that at least one them has an extremum value (0 or 1). This finishes the proof of the theorem. □



(a) $\mu_0 = 2$, $\mu_1 = 1 = \mu_3$, $\mu_2 = 3$ (b) $\mu_i = 2$ for $i = 0, ..., 3$

Figure 3.9: Optimal p and q with $w^* = 1$

Figure 3.9 provides a numerical illustration of Theorem 1. We observe as a function of the system parameters that at least one of the routing parameters is either 0 or 1. This gives the merit to the study of the extreme cases Models 1,...,4. Moreover they are easy to implement and easy to understand for both managers and agents. Note that the same observation holds also for Problem (3.1) with a constraint on $P(W < AWT)$. This is based on several numerical examples (a rigorous proof as that for $E(W)$ is too complex to develop). An illustration is given in Figure 3.10.



(a) $\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$ (b) $\mu_i = 2$ for $i = 0, ..., 3$

Figure 3.10: Optimal p and q with $\mathrm{P}(W < 1) \geq 80\%$)

3.6 Approximations

In this section, we develop two approximation methods. In Section 3.6.1 we focus on a light-traffic approximation for Model PM in the case of a single server. In Section 3.6.2, we propose an approximation method to extend our email routing optimization problem to the multi-server case.

3.6.1 Light-Traffic Approximation for Model PM

Although the analysis in Section 3.4.1 provides an exact method to derive the system steady-state probabilities of Model PM, the expressions involved are cumbersome. In this section, we develop an approximate analysis that allows to obtain simpler expressions for those quantities under the light-traffic regime.

Under the light-traffic regime, we have $\rho_i << 1$ and also $\frac{\rho_i}{1+\rho_i} \simeq \rho_i$, for $i = 0, ..., 3$. Equation (3.9) can be then rewritten as $x^4 - \sigma_1 x^3 + \sigma_2 x^2 - \sigma_3 x + \sigma_4 = 0$, or equivalently

$$(x - \rho_0)(x - \rho_1)(x - \rho_2)(x - \rho_3) = 0.$$

Also, the matrix $\frac{1}{\lambda}A$ becomes

$$\frac{1}{\lambda}A = \begin{pmatrix} 1/\rho_1 & 0 & 0 & 0 & 0 \\ -1/\rho_1 & 1/\rho_2 & 0 & 0 & 0 \\ 0 & -q/\rho_2 & 1/\rho_0 & 0 & 0 \\ 0 & -(1-q)/\rho_2 & -1/\rho_0 & 1/\rho_3 & 0 \\ 0 & 0 & 0 & 0 & 1/\rho_0 \end{pmatrix},$$

which easily leads to

$$\lambda A^{-1} = \begin{pmatrix} \rho_1 & 0 & 0 & 0 & 0 \\ \rho_2 & \rho_2 & 0 & 0 & 0 \\ q\rho_0 & q\rho_0 & \rho_0 & 0 & 0 \\ \rho_3 & \rho_3 & \rho_3 & \rho_3 & 0 \\ 0 & 0 & 0 & 0 & \rho_0 \end{pmatrix}.$$

The computation of the steady-state probabilities simplifies as follows. We have $P_0 = 1-p$, $m_0 = p$, $a_0 = \rho_1$, $b_0 = \rho_2$, $b'_0 = q\rho_0$ and $c_0 = \rho_3$. Using Equation (3.7), we obtain $a_n = \rho_1^{n+1}$, for $n \geq 0$. Also, $b_n = \rho_2 a_{n-1} + \rho_2 b_{n-1}$ for $n \geq 1$. Therefore, if $\mu_1 \neq \mu_2$, we have $b_n = \rho_2 \sum_{k=0}^{n} \rho_1^k \rho_2^{n-k} = \rho_2 \frac{\rho_2^{n+1} - \rho_1^{n+1}}{\rho_2 - \rho_1}$ for $n \geq 0$, and if $\mu_1 = \mu_2$, we have $b_n = (n+1)\rho_1^{n+1}$ for $n \geq 0$. From Equation (3.7), we may write $b'_n = q\rho_0 a_{n-1} + q\rho_0 b_{n-1} + \rho_0 b'_{n-1}$, for $n \geq 1$. Thus if $\mu_1 \neq \mu_2$, we have $b'_n = q\rho_0 \frac{\rho_2^{n+1} - \rho_1^{n+1}}{\rho_2 - \rho_1} + \rho_0 b'_{n-1}$, for $n \geq 1$. Therefore if $\mu_i \neq \mu_j$ for $i \neq j \in \{0,1,2\}$ we have $b'_n = \frac{q\rho_0\rho_2}{(\rho_2-\rho_1)(\rho_2-\rho_0)}(\rho_2^{n+1} - \rho_0^{n+1}) - \frac{q\rho_0\rho_1}{(\rho_2-\rho_1)(\rho_1-\rho_0)}(\rho_1^{n+1} - \rho_0^{n+1})$, or equivalently

$$b'_n = q\rho_0 \left(\frac{\rho_0^{n+2}}{(\rho_0-\rho_1)(\rho_0-\rho_2)} + \frac{\rho_1^{n+2}}{(\rho_1-\rho_0)(\rho_1-\rho_2)} + \frac{\rho_2^{n+2}}{(\rho_2-\rho_0)(\rho_2-\rho_1)} \right),$$

for $n \geq 0$. If $\mu_i = \mu_j$ and $\mu_i \neq \mu_k$ for distinctive $i, j, k \in \{0,1,2\}$, then

$$b'_n = q\rho_0 \left(\frac{(n+2)\rho_i^{n+1}}{\rho_i - \rho_k} + \frac{\rho_k^{n+2}}{(\rho_k - \rho_i)^2} \right),$$

for $n \geq 0$. If $\mu_1 = \mu_2 = \mu_0$, we have

$$b_n^{'} = q\rho_0^{n+1}\frac{(n+1)(n+2)}{2},$$

for $n \geq 0$. From Equation (3.7), we also may write $c_n = \rho_3(a_{n-1} + b_{n-1} + b'_{n-1} + c_{n-1})$, for $n \geq 1$. Thus, if $\mu_i \neq \mu_j$ for $i \neq j \in \{0, 1, 2, 3\}$, we have

$$\begin{aligned} c_n =& \rho_3\left(\frac{\rho_0^{n+2}(\rho_0 - \rho_0(1-q))}{(\rho_0-\rho_1)(\rho_0-\rho_2)(\rho_0-\rho_3)} + \frac{\rho_1^{n+2}(\rho_1 - \rho_0(1-q))}{(\rho_1-\rho_0)(\rho_1-\rho_2)(\rho_1-\rho_3)}\right. \\ &\left. + \frac{\rho_2^{n+2}(\rho_2 - \rho_0(1-q))}{(\rho_2-\rho_0)(\rho_2-\rho_1)(\rho_2-\rho_3)} + \frac{\rho_3^{n+2}(\rho_3 - \rho_0(1-q))}{(\rho_3-\rho_0)(\rho_3-\rho_1)(\rho_3-\rho_2)}\right), \end{aligned}$$

for $n \geq 0$. The expression of c_n, for $n \geq 0$, can be also easily derived in closed form in the other remaining cases for μ_i and μ_j for $i \neq j \in \{0, 1, 2, 3\}$. Finally, we deduce from Equation (3.7) that $m_n = p\rho_0^n$, for $n \geq 0$. Because of the light-traffic approximation in the above computations, the system steady-state probabilities do not sum up to one. We then normalize them by dividing each one of them by the summation of all the steady-state probabilities. In Table 3.3 we evaluate the light-traffic approximation. The comparison of the approximate results with those from the exact numerical analysis show that the approximation is appropriate under the light-traffic regime. Note that the probabilities for $n \geq 4$ are too close to zero for both the exact and the approximation methods.

3.6.2 Multi-Server Case

In this section, we focus on the email routing problem for the multi-server case. The modeling is the same as described in Section 3.3, expect that instead of one server, there are s identical, parallel servers. As previously, we consider a call center manager that wants to optimize the email routing as a function of the system parameters (Problem (3.1)). In other words, we want to either optimize p and q for Model PM, or give the ordering of the extreme cases Models 1 to 4 (easier to use in

Table 3.3: Validation of the light-traffic approximation

	$\frac{\lambda}{\mu}$=28%, $\lambda = 0.2$, $\mu_1 = \mu_2 = 5$, $\mu_3 = \mu_0 = 2$, $p = 10\%$, $q = 80\%$		$\frac{\lambda}{\mu}$=20.83%, $\lambda = 0.1$, $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_0 = 4$, $p = q = 1$		$\frac{\lambda}{\mu}$=2.00%, $\lambda = 0.01$ $\mu_i = 2$ for $i = 0, \cdots, 3$ $p = q = 50\%$	
	Exact	Approximation	Exact	Approximation	Exact	Approximation
P_0	65.9406%	67.9873%	0,0000%	0,0000%	49,0025%	49.0093%
a_0	3.6157%	3.0217%	8,5896%	7,8955%	0,4962%	0.4901%
b_0	3.4766%	3.0217%	4,0903%	3,9478%	0,4937%	0.4901%
b'_0	6.3211%	6.0433%	1,9953%	1,9739%	0,2456%	0.2450%
c_0	7.3267%	7.5541%	2,5745%	2,6318%	0,4900%	0.4901%
m_0	7.3267%	7.5541%	77,2358%	78,9551%	49,0025%	49.0093%
a_1	0.3308%	0.1209%	1,2323%	0,7896%	0,0038%	0.0025%
b_1	0.4518%	0.2417%	0,7816%	0,5922%	0,0062%	0.0049%
b'_1	1.3961%	1.0878%	0,4299%	0,3454%	0,0043%	0.0037%
c_1	2.1406%	1.9641%	0,6378%	0,5483%	0.0098%	0.0086%
m_1	0.6661%	0.7554%	1,8838%	1,9739%	0.2438%	0.2450%
a_2	0.0451%	0.0048%	0,1559%	0,0790%	0.0000%	0.0000%
b_2	0.0607%	0.0145%	0,1114%	0,0691%	0,0001%	0.0000%
b'_2	0.2374%	0.1378%	0,0648%	0,0432%	0.0000%	0.0000%
c_2	0.4380%	0.3414%	0,1042%	0,0758%	0,0001%	0.0001%
m_2	0.0606%	0.0755%	0,0459%	0,0493%	0,0012%	0.0012%
a_3	0.0071%	0.0002%	0,0195%	0,0079%	0.0000%	0.0000%
b_3	0.0091%	0.0008%	0,0146%	0,0074%	0.0000%	0.0000%
b'_3	0.0382%	0.0153%	0,0087%	0,0048%	0.0000%	0.0000%
c_3	0.0787%	0.0499%	0,0146%	0,0089%	0.0000%	0.0000%
m_3	0.0055%	0.0076%	0,0011%	0,0012%	0.0000%	0.0000%

Table 3.4: Comparison between approximation and simulation

		Interval of the call arrival rate λ			
		$s = 10$, $\mu_0 = 2$, $\mu_1 = \mu_3 = 1$, $\mu_2 = 3$		$s = 10$, $\mu_0 = \mu_1 = \mu_2 = \mu_3 = 2$	
Constraint on calls	Model	Simulation	Approximation	Simulation	Approximation
$E(W) \leq 1$	Model 4	$0 - 3.04$	$0 - 2.96$	$0 - 4.44$	$0 - 4.42$
	Model 3	$3.04 - 3.74$	$2.96 - 2.97$	$4.44 - 4.48$	$4.42 - 4.45$
	Model 2	$3.04 - 3.74$	$2.97 - 3.66$	$4.48 - 6.05$	$4.45 - 6.04$
	Model 1	$3.74 - 3.85$	$3.66 - 3.69$	$6.05 - 6.07$	$6.04 - 6.06$
$P(W < 1) \geq 0.8$	Model 4	$0 - 1.2$	$0 - 1.1$	$0 - 1.3$	$0 - 1.2$
	Model 2	$1.2 - 1.9$	$1.1 - 1.7$	$1.3 - 2.0$	$1.2 - 1.8$
	Model 1	$1.9 - 2.3$	$1.7 - 2.1$	$2.0 - 2.5$	$1.8 - 2.4$

practice and also efficient).

An exact analysis as that done for the single server case is too complex. We propose an approximation based on the single server results. It consists on replacing the s servers by a single *super* server. The service rates become $s\mu_0$ (for emails), $s\mu_1$ (first stage of call service), $s\mu_2$ (second stage of call service), $s\mu_3$ (third stage of call service). We have conducted an extensive simulation study in order to assess the quality of this approximation. Some of the comparison results between approximation and simulation are given in Table 3.4.

In Table 3.4 we give as a function of the interval of the call arrival rate value the ordering of Models 1 to 4 with respect to the optimization problem. The intervals are given using the single server approximation and also using a combined simulation and optimization of the multi-server system. The same intervals hold also when considering Model PM. We observe from Table 3.4 that the approximation provides an appropriate solution for the email routing optimization.

3.7 Conclusions

We considered a blended call center with calls and emails. The call service is characterized by successive stages where one of them is a break for the agent in the sense that inside the conversation there is no required interaction during a non-negligible time between the two parties. We addressed an important question in the call center practice: how should managers make use of this opportunity

in order to better improve performance? We focused on the optimization of the email routing given that calls have a non-preemptive priority over emails. Our objective was to maximize the throughput of emails subject to a constraint on the call waiting time.

We developed a general framework (Model PM) with two probabilistic parameters for the email routing to agents. One parameter controls the routing between calls, and the other does the control inside a call conversation. We have also considered 4 particular cases corresponding to the extreme values of the probabilistic parameters. For these routing models, we have derived various structural results with regard to the optimization problem. We have also numerically illustrated and discussed the theoretical results in order to provide guidelines to call center managers. In particular, we proved for the optimal routing that all the time at least one of the two email routing parameters has an extreme value.

There are several avenues for future research. It would be interesting to extend the structural results to the multi-server case. It would also be useful but challenging to extend the analysis to cases with an additional channel, in particular the chat which is increasingly used in call centers. Using the chat channel, an agent may handle many customers at the same time, which represent an additional opportunity to efficiently use the agent time.

Chapter 4

Adaptive Threshold Policies for Multi-Channel Call Centers

In the context of multi-channel call centers with inbound and outbound calls, we consider a threshold policy on the reservation of agents for the inbound calls. We study a general non-stationary model where the call arrival follows a non-homogeneous Poisson process. The optimization problem consists on maximizing the throughput of outbound calls under a constraint on the waiting time of inbound calls. We propose an efficient adaptive threshold policy easy to implement in practice in the Automatic Call Distribution (ACD). This scheduling policy is evaluated through a comparison with the optimal performance measures found in the case of a constant stationary arrival rate, and also a comparison with other intuitive adaptive threshold policies in the general non-stationary case.

4.1 Introduction

Call centers require a very accurate match of demand and supply. The delay of the treatment of a call, its waiting time, is usually not allowed to exceed 20 seconds. Thus a very accurate prediction of the demand is required. However, this can rarely be obtained, because of the volatility of call arrival patterns. Therefore there is often a mismatch between demand and the scheduled supply, consisting of rostered call center employees (usually called agents). Moreover, even if the demand is accurately forecasted, a considerable overcapacity should be scheduled to be able to deal with the random Poisson fluctuations of the demand. Usually queueing models are used to quantify this overcapacity, most often Erlang C.

To prevent idle overcapacity and to limit the necessity to have extremely accurate forecast inbound calls are sometimes mixed with other types of customer contacts which have a less strict allowable delay, such as emails or outbound calls. This is called *(call) blending*. The amount of capacity assigned to the other channels is supposed to adapt to the number of inbound calls, giving at the same time a good service level for the inbound calls and a good occupancy of the call center agents.

Because of the strict waiting time requirement on inbound calls it is best to give them priority over the other channels. To maximize agent productivity it would be best to assign an email to every idle agent where there are no inbound calls in queue. This would lead to a 100% productivity. There are two objections against this policy: a 100% occupancy is not sustainable over a longer period because of agent fatigue, and it would lead to long waiting times for inbound calls because never an agent is waiting for an inbound call to arrive. For these reasons a more sophisticated assignment policy is required.

In Bhulai and Koole (2003) and Gans and Zhou (2003b) it is shown that the optimal assignment policy is of the following form: outbound calls should only be scheduled when there are no waiting inbound calls and when the number of idle agents exceeds a certain threshold. Thus the problem of

controlling our blended call center reduces to determining the right threshold level. This threshold however depends on all the parameters of the system. But these parameters, especially the arrival rate, are often hard to determine. This calls for a policy in which the threshold is adapted to the current situation without using explicitly the parameters of the system. In this chapter such adaptive policies are studied, both for systems with a constant (but unknown) arrival rate and for the more realistic situation of a fluctuating arrival rate. The parameter that is used to update the threshold is the service level up to that moment, a number which is always available in call centers. The overall objective is to reach a certain service level by the end of the day, while maximizing the number of emails or outbound calls that are done.

We discuss the relevant literature. There is a rich literature on planning and scheduling in call centers, see Gans et al. (2003); Akşin et al. (2007). However, few papers focus on blending. The general context of multi-channel call centers is described in Koole (2013), Chapter 7.

Deslauriers et al. (2007) extend the earlier mentioned papers by having different types of agents. outbound calls are served only by blend agents, whereas inbound calls can be served by either inbound-only or blend agents. Inbound callers may balk or abandon. They evaluate several performance measures of interest, including the rate of outbound calls and the proportion of inbound calls waiting more than some fixed number of seconds. A collection of continuous-time Markov chain (CTMC) models which capture many real-world characteristics while maintaining parsimony that results in fast computation are presented. They discuss and explore the tradeoffs between model fidelity and efficacy and compare different CTMC models with a realistic simulation model of a Bell Canada call center.

Armony and Ward (2010) present an optimization problem; the objective is to minimize steady-state expected customer wait time subject to a fairness constraint on the workload division. They show in such a problem, which is close to ours, that a threshold policy outperforms a common routing policy used in call centers (that routes to the agent that has been idle the longest).

Milner and Olsen (2008) consider a call center with contract and non contract customers. They explore the use to give priority to contract customers only in off peaks. They show that this choice is a good one under classical assumptions (steady-states) and the same performance measure as ours and present also examples when it is not. This result is important since we found an insight arguing that the service level for inbound calls has to be very strictly respected during off peaks.

This chapter is organized as follows. Section 4.2 presents our model. Sections 4.3 and 4.4 contain our results, first for a constant arrival rate in Section 4.3 and then in Section 4.4 with a fluctuating arrival rate. We end with a short conclusion.

4.2 Model

We consider a call center modeled as a multi-server queuing system with two types of jobs, foreground jobs (inbound calls) and background jobs (outbound calls, emails, etc.). We use the terms foreground jobs and calls (background jobs and emails) interchangeably. The arrival process of calls is assumed to be a non-homogeneous Poisson process with rate $\lambda(t)$, for $t \geq 0$. Calls arrive at a dedicated first come, first served (FCFS) queue with infinite capacity. There is an infinite supply of background jobs, waiting for treatment in a dedicated FCFS queue. There are s identical, parallel servers (agents in call center parlance). Each agent can handle both types of jobs. We assume that the service times of foreground and background jobs are exponentially distributed with rates μ and μ_0, respectively. Neither abandonment nor retrials are modeled.

Foreground jobs are more important than background ones in the sense that the former request a quasi-instantaneous answer (waiting time in the order of seconds or minutes), while the latter are more flexible and could be delayed for several (tens of) hours. The objective of the call center manager over a working day is to maximize the email throughput while satisfying a constraint on the call waiting time in the queue.

Since the model is transient, we can not define the waiting time of an arbitrary customer as a

unique random variable. There is a random number of served customers during the working period S, if $S > 0$ we define the random variable for the waiting time of customer n for $n \in \{1, \ldots, S\}$, W_n. We want that the expected proportion of calls that wait less than a predefined threshold τ is at least equal to α, i.e., $S^{-1}E\sum_{n=1}^{S} I\{W_n \leq \tau\} \geq \alpha$, for $\tau \geq 0$ and $0 \leq \alpha \leq 1$. Note that we do not consider arriving customers at the end of the working period which can not be served.

We then aim to find the best routing rules in terms of efficiency for the considered problem and easiness of implementation in call center software. We assume that preemption of jobs in service is not allowed. This is a quite natural assumption. An agent usually prefers to finish answering an underway email rather than starting it over later on. This is also preferred from an efficiency perspective. Evidently, when the background jobs are outbound calls, then it is neither acceptable to preempt.

For a similar model but with a constant arrival rate and equal service requirements for the two job types, Bhulai and Koole (2003) prove that the optimal policy is a threshold policy with the priority given to calls (some servers reserved for calls). Their result is mainly based on the fact that it is optimal to handle calls as long as the queue of calls is not empty. For our general modeling, the analysis is more complicated. Even for a constant arrival rate but different service requirements, the optimal policy is hard to obtain, and might not be useful in practice (for software implementation for example). For simplicity and usefulness of the results in practice, we then restrict ourselves to the case of threshold policies. Moreover, Bhulai and Koole (2003) numerically show, for more general cases, that the appealing threshold policies are good approximations of the optimal ones. More concretely, the functioning of the call center under a threshold policy is as follows. Let us denote the threshold by u, $0 \leq u \leq s$. Upon arrival, a call is immediately handled by an available agent, if any. If not, the call waits in the queue. When an agent becomes idle, she handles the call at the head of the queue with calls, if any. If not, the agent may either handle an email at the head of the email queue, or she remains idle. If the number of idle agents (excluding her) is

at least $s - u$, then the agent in question handles an email. Otherwise, she remains idle. In other words, there are $s - u$ agents that are reserved for calls, u is the number of agents that are always working.

In this chapter, we propose an adaptive threshold policy which adjusts the threshold as a function of the non-stationary arrival process of calls. We divide the working day into N identical intervals, each with length θ. The total working duration in a day is L, $L = N\theta$. At the beginning of each interval i ($i = 1, ..., N$), we define the threshold u_i, $0 \leq u_i \leq s$, under which the job routing policy works during interval i. Let T denote the expected throughput of emails over the whole day, i.e., the ratio between the number of treated emails and L. Let also SL be the proportion, for the whole day, of calls that have waited less than τ, $SL = S^{-1}E\sum_{n=1}^{S} I\{W_n \leq \tau\}$. In summary, our optimization problem can be formulated as

$$\begin{cases} \text{Maximize } T \\ \text{subject to } SL \geq \alpha, \end{cases} \tag{4.1}$$

where the decision variables are u_i with $0 \leq u_i \leq s$, for $i = 1, ..., N$. It is clear that the best case for calls is such that $u_i = 0$ for all i, which means that no email is treated and SL is maximized (case of an M(t)/M/s with only calls). We therefore assume from now on that the parameters $\lambda(t)$ for $t \geq 0$, μ and s are such that $SL \geq \alpha$ for $u_i = 0$ ($i = 1, ..., N$).

4.3 Constant Arrival Rate

We consider a basic case with a constant arrival rate, $\lambda(t) = \lambda$ for $t \geq 0$ and a constant threshold, $u_i = u$ for $i = 1, ..., N$ and $0 \leq u \leq s$. The purpose of the analysis in this section is to understand the behavior of the performance measures as a function of the threshold in order to build an efficient method for the threshold adaptation rule (u_i for $i = 1, ..., N$) in the case of a non-constant arrival rate. In Section 4.3.1 we propose a method to compute the performance measures, then in Section

4.3.2 we use them to provide a useful insight to construct our adaptive policy.

4.3.1 Performance Measures

In Section 4.3.1 we provide close form formula of the performance measures in the case of equal service rates and study the form of these measures as a function of the threshold. Then in Section 4.3.1 we propose a numerical method to compute the performance measures in the case of unequal service rates. Since we consider a stationary model we can define a unique random variable for the waiting time of an arbitrary customer W, and denote by $P(W < \tau)$ the probability that an arbitrary customer waits less than τ ($\tau > 0$).

Equal Service Rates

We consider the case $\mu = \mu_0$. First, we compute the performance measures of interest for calls and emails for a given constant reservation threshold, denoted by u, $0 \leq u \leq s$. We then develop some structural results that will be used in Section 4.3.2.

Let us define the stochastic process $\{x(t), t \geq 0\}$, where $x(t) \in \mathbb{N}$ is the number of jobs in service plus the number of jobs in the queue of calls. Since $\mu = \mu_0$, we need not distinguish between the two job types in service. The process $\{x(t), t \geq 0\}$ is a birth-death process. The transition rate from state x to state $x-1$ is $\min\{x, s\}\mu$, for $x \geq 1$, and that from state x to state $x+1$ is λ, for $x \geq 0$. Under the stability condition $\frac{\lambda}{s\mu}$, we denote by p_x the steady-state probability to be in state $x \in \mathbb{N}$, and by a the ratio $\frac{\lambda}{\mu}$. In Theorem 2, we give the expression of the email throughput, $T(s, u, a)$, and that of the probability that the call waiting time is less than τ, $SL = P(W < \tau)$.

Theorem 2 *For $0 \leq u \leq s$, we have*

$$T(s,u,a) = \mu \left(\sum_{k=0}^{s-u} \frac{a^k u!}{(u+k)!} + \frac{a^{s-u} u!}{s!} \frac{a}{s-a} \right)^{-1} \left(u + \sum_{k=1}^{s-u} \frac{a^k u!}{(u+k-1)!} + \frac{a^{s-u+1} u!}{(s-1)!(s-a)} \right) - \lambda, \tag{4.2}$$

$$P(W < \tau) = 1 - C(s,u,a)e^{-\tau(s\mu-\lambda)}, \tag{4.3}$$

with

$$C(s,u,a) = \frac{a^{s-u}u!}{s!(1-a/s)}\left(\sum_{k=0}^{s-u}\frac{a^k u!}{(u+k)!} + \frac{a^{s-u}u!}{s!}\frac{a}{s-a}\right)^{-1}. \tag{4.4}$$

Proof. For $0 \leq x < u$, we have $p_x = 0$. For $0 \leq k \leq s-u$, we have $p_{u+k} = \frac{a^k u!}{(u+k)!}p_u$. For $k \geq 0$, we have $p_{s+k} = \frac{a^k}{s^k}p_s$. Since all probabilities sum up to one, we obtain

$$p_u = \left(\sum_{k=0}^{s-u}\frac{a^k u!}{(u+k)!} + \frac{a^{s-u}u!}{s!}\frac{a}{s-a}\right)^{-1}. \tag{4.5}$$

The email throughput can be seen as the overall throughput (of calls and emails) minus the call throughput. Thus

$$T(s,u,a) = \sum_{k=0}^{s-u}(u+k)\mu p_{u+k} + s\mu\sum_{k=1}^{\infty}p_{s+k} - \lambda.$$

After some algebra, we deduce that

$$T(s,u,a) = \mu p_u\left(u + \sum_{k=1}^{s-u}\frac{a^k u!}{(u+k-1)!} + \frac{a^{s-u+1}u!}{(s-1)!(s-a)}\right) - \lambda.$$

Note that the lower bound of $T(s,u,a)$ is $T(s,0,a) = 0$, which corresponds to the case when all servers are reserved to calls. As for the upper bound, it is $T(s,s,a) = s\mu - \lambda$, which corresponds to the case of no server reservation for calls (the infinite amount of emails leads to $s\mu$ as a total throughput for the two job types).

The call service level, $P(W > \tau)$, is obtained using the PASTA property. We have $P(W > \tau) = \sum_{n=0}^{\infty} p_{s+n}P(W > \tau | x = n+s)$, where $P(W > \tau | x = s+n)$ is the conditional probability that the waiting time of a new call exceeds τ, given that it finds all servers busy and n calls waiting ahead in

the queue, $n \geq 0$. It is easy to see that this conditional waiting time follows an Erlang distribution with $n+1$ stages and a rate of $s\mu$ per stage. Then, $P(W > \tau | x = s+n) = \sum_{k=0}^{n} e^{-s\mu\tau} \frac{(s\mu\tau)^k}{k!}$, which leads to

$$\begin{aligned} P(W > \tau) &= \sum_{n=0}^{\infty} p_s \frac{a^n}{s^n} \sum_{k=0}^{n} e^{-s\mu\tau} \frac{(s\mu\tau)^k}{k!} \\ &= \lim_{n\to\infty} \left(p_s e^{-s\mu\tau} \sum_{k=0}^{n} \sum_{n=k}^{\infty} \frac{(s\mu\tau)^k}{k!} \left(\frac{a}{s}\right)^n \right). \end{aligned}$$

Observing that $\sum_{n=k}^{\infty} \left(\frac{a}{s}\right)^n = \left(\frac{a}{s}\right)^k \frac{1}{1-a/s}$ implies

$$P(W > \tau) = C(s,u,a) e^{-\tau(s\mu-\lambda)}, \tag{4.6}$$

with

$$C(s,u,a) = \frac{p_s}{1-a/s} = \frac{a^{s-u}u!}{s!(1-a/s)} \left(\sum_{k=0}^{s-u} \frac{a^k u!}{(u+k)!} + \frac{a^{s-u}u!}{s!} \frac{a}{s-a} \right)^{-1}. \tag{4.7}$$

Note that the upper bound of $C(s,u,a)$ is $C(s,s,a) = 1$ (no server reservation for calls, then, any arriving call has to wait for service), and its lower bound is $C(s,0,a) = \frac{a^s}{s!(1-a/s)} \left(\sum_{k=0}^{s} \frac{a^k}{(k)!} + \frac{a^s}{s!(1-a/s)} \right)^{-1}$ (all servers are reserved to calls, which corresponds for calls to the case of an M/M/s queue with no emails). This finishes the proof of the theorem. □

In Section C.1 of the appendix, we prove some monotonicity results of the performance measures in the threshold. We prove that the email throughput is strictly increasing and neither convex nor concave in u, for $0 \leq u \leq s$. However the end of the email throughput, for $s-2 \leq u \leq s$ and $s \geq 2$, is concave in u. The inbound service level $P(W < \tau)$ is strictly decreasing in u, for $0 \leq u \leq s$. We prove that it is concave in u, for the cases $a < 1$ and also for $a \geq u+1$ $(u < s)$. An extensive numerical study for the remaining cases shows that the concavity still holds.

Unequal Service Rates

In this section we focus on the performance evaluation (email throughput and call waiting time distribution) for the case of unequal service rates, $\mu \neq \mu_0$. In contrast to the case of equal service rates, the performance expressions are here too cumbersome to allow the development of useful structural results. The results of this section are however still useful for the numerical experiments in Section 4.3.2 in order to build the insights on the threshold policy for the more general case with a non-constant call arrival rate.

As in Bhulai and Koole (2003), our approach consists on using a Markov chain analysis to derive the steady-state probabilities of the system, from which the performance measures are characterized thereafter. To simplify the presentation, we focus on the particular case $u = s$. The analysis for the case $u = 0$ is obvious, and that of the remaining cases, $0 < u < s$, is done similarly to the case $u = s$. It simply adds a finite number of additional equations but does not impact the general form of the steady-state probabilities. Consider the stochastic process $\{(x(t), y(t)), t \geq 0\}$, where $x(t)$ is the number of waiting calls in the queue and $y(t)$ is the number of emails being in service, $(x, y) \in \mathbb{N}^2$. This process is a Markov chain. For $x \geq 0$ and $y \geq 0$, the transition rate from (x, y) to $(x + 1, y)$ is λ. For $x \geq 1$ and $y \geq 0$, the transition rate from (x, y) to $(x - 1, y)$ is $(s - y)\mu$. For $x \geq 1$ and $y \geq 1$ the transition rate from (x, y) to $(x - 1, y - 1)$ is $y\mu_0$. For $y \geq 0$, the transition rate from $(0, y)$ to $(0, y + 1)$ is $(s - y)\mu$. Under the stability condition $\frac{\lambda}{s\mu} < 1$, we denote by $p_{x,y}$ the steady-state probability that the system is in state (x, y).

For $y = s$ and $x > 0$, we have $p_{x,s}(\lambda + s\mu_0) = \lambda p_{x-1,s}$, then $p_{x,s} = \left(\frac{\lambda}{\lambda+s\mu_0}\right)^x p_{0,s}$. Using $\lambda p_{0,s} = \mu p_{0,s-1}$, we deduce that $p_{x,s} = \frac{\mu}{\lambda}\left(\frac{\lambda}{\lambda+s\mu_0}\right)^x p_{0,s-1}$ for $x \geq 0$. For $y = s - 1$ and $x > 0$, we may write $(\lambda+\mu+(s-1)\mu_0)p_{x,s-1} = \lambda p_{x-1,s-1}+\mu p_{x+1,s-1}+s\mu_0 p_{x+1,s}$. The associated homogeneous equation in the variable z is $\mu z^2 - (\lambda + \mu + (s - 1)\mu_0)z + \lambda = 0$, for $z \in \mathbb{C}$. It has two solutions denoted by z_1 and z_2 and are given by $z_1 = \frac{1}{2\mu}\left(\lambda + \mu + (s - 1)\mu_0 + \sqrt{(\lambda + \mu + (s - 1)\mu_0)^2 - 4\lambda\mu}\right)$

and $z_2 = \frac{1}{2\mu}\left(\lambda+\mu+(s-1)\mu_0 - \sqrt{(\lambda+\mu+(s-1)\mu_0)^2 - 4\lambda\mu}\right)$. We deduce, for $x \geq 0$, that

$$p_{x,s-1} = \beta_1(z_1)^x + \beta_2(z_2)^x + \beta_3\left(\frac{\lambda}{\lambda+s\mu_0}\right)^x,$$

with $\beta_3 = -\frac{\mu s}{\lambda+(\mu_0-\mu)s}p_{0,s-1}$. From the boundaries $x=0$ and $x=1$, we obtain $p_{0,s-1} = \beta_1+\beta_2 - \frac{\mu s}{\lambda+(\mu_0-\mu)s}p_{0,s-1}$ and $p_{1,s-1} = \beta_1 z_1 + \beta_2 z_2 - \frac{\mu s}{\lambda+(\mu_0-\mu)s}\left(\frac{\lambda}{\lambda+s\mu_0}\right)p_{0,s-1}$, respectively, which implies

$$\beta_1 = \frac{2\mu p_{0,s-2}(\lambda+s(\mu_0-\mu)) + p_{0,s-1}(z_2\mu\mu_0 s - \lambda\mu_0 s + \lambda\mu s + z_2\lambda\mu - \lambda\mu - \lambda^2}{(z_2-z_1)\ \mu\ ((\mu_0-\mu)s+\lambda)},$$

$$\beta_2 = \frac{2\mu p_{0,s-2}(\lambda+s(\mu_0-\mu)) + p_{0,s-1}(z_1\mu\mu_0 s - \lambda\mu_0 s + \lambda\mu s + z_1\lambda\mu - \lambda\mu - \lambda^2}{(z_2-z_1)\ \mu\ ((\mu_0-\mu)s+\lambda)}.$$

For $y=k$, $0 \leq k < s$, and $x > 0$ we have

$$(\lambda+(s-k)\mu+k\mu_0)p_{x,k} = \lambda p_{x-1,k} + (s-k)\mu p_{x+1,k} + (k+1)\mu_0 p_{x+1,k+1}. \tag{4.8}$$

The homogeneous equation associated to Equation (4.8) is

$$(s-k)\mu z^2 - (\lambda+(s-k)\mu+k\mu_0)z + \lambda = 0,$$

with z as a variable, for $z \in \mathbb{C}$. It has two solutions denoted by $z_{1,k}$ and $z_{2,k}$ and are given by

$$z_{1,k} = \frac{1}{2(s-k)\mu}\left(\lambda+(s-k)\mu+k\mu_0 - \sqrt{(\lambda+(s-k)\mu+k\mu_0)^2 - 4(s-k)\lambda\mu}\right),$$

$$z_{2,k} = \frac{1}{2(s-k)\mu}\left(\lambda+(s-k)\mu+k\mu_0 + \sqrt{(\lambda+(s-k)\mu+k\mu_0)^2 - 4(s-k)\lambda\mu}\right),$$

for $0 \leq k < s$. Because of the last term in the right hand side of Equation (4.8), one may write,

for $0 \leq k \leq s$ and $x > 0$,

$$p_{x,k} = \sum_{i=k}^{s} A_{i,k} z_{1,i}^{x} + B_{i,k} z_{2,i}^{x},$$

with $z_{1,s}$ and $z_{2,s}$ defined as $z_{1,s} = \frac{\lambda}{\lambda + s\mu_0}$ and $z_{2,s} = 0$, respectively, and $A_{i,k}, B_{i,k} \in \mathbb{R}$ for $0 \leq k < s$ and $k \leq i \leq s$. Using Equation (4.8), we can prove that

$$A_{i,k+1} = A_{i,k} \frac{-(s-k)\mu z_{1,i}^2 + (\lambda + (s-k)\mu + k\mu_0) z_{1,i} - \lambda}{(k+1)\mu_0 z_{1,i}^2},$$

for $0 \leq k < s$ and $k < i \leq s$. Similarly, we have

$$B_{i,k+1} = B_{i,k} \frac{-(s-k)\mu z_{2,i}^2 + (\lambda + (s-k)\mu + k\mu_0) z_{2,i} - \lambda}{(k+1)\mu_0 z_{2,i}^2},$$

for $0 \leq k < s$ and $k < i \leq s$. For $i = k$, we can easily derive $A_{k,k}$ and $B_{k,k}$ as a function of $p_{0,k}$ and $p_{1,k}$, for $0 \leq k < s$. The relation between $p_{1,k}$ and $p_{0,k}$ is given by

$$(\lambda + (s-k)\mu) p_{0,k} = (s-k)\mu p_{1,k} + (k+1)\mu_0 p_{1,k+1} + (s-k-1) p_{0,k-1},$$

for $0 \leq k < s$. We also have $z_{1,0} = 1$ and $z_{2,0} = \frac{\lambda}{s\mu}$. Since all probabilities sum up to one, we obtain $A_{0,0} = 0$. In conclusion through the above analysis, we have characterized all steady-state probabilities, $p_{i,k}$, for $i \geq 0$ and $0 \leq k \leq s$. The email throughput $T(\lambda, \mu, \mu_0, s)$ may be written as

$$T(\lambda, \mu, \mu_0, s) = \mu_0 \sum_{k=1}^{s} \sum_{i=0}^{\infty} k p_{i,k},$$

or equivalently

$$T(\lambda, \mu, \mu_0, s) = \lambda + \sum_{k=1}^{s} k\mu_0 p_{0,k},$$

for $s \geq 1$. As for the call waiting performance, it is given by

$$P(W > \tau) = \sum_{k=1}^{s} \sum_{i=0}^{\infty} p_{i,k} P(W > \tau | (x, y) = (i, k)),$$

where $P(W > \tau | (x, y) = (i, k))$ is the conditional probability that the waiting time of a new call exceeds τ, given that it finds i emails in service, $s - i$ calls in service, and k calls waiting ahead in the queue, for $0 \leq i \leq s$ and $k \geq 0$. The computation of $P(W > \tau | (x, y) = (i, k))$, for $0 \leq i \leq s$ and $k \geq 0$, is as follows. For $k = 0$ and $0 \leq i \leq s$, the new call has to wait for a service completion of one of the i emails, or one of the $s-i$ calls, so, $P(W > \tau | (x, y) = (i, 0)) = e^{-\tau(i\mu_0 + (s-i)\mu)}$. For $k = 1$ and $0 \leq i \leq s$, the probability that the next service completion is that of an email is $\frac{i\mu_0}{i\mu_0 + (s-i)\mu}$. Thus, the waiting time of the new call follows a hypoexponential distribution consisting of the summation of two exponential random variables with rates $i\mu_0 + (s-i)\mu$ and $\max(0, i-1)\mu_0 + \min(s, s-i+1)\mu$ with probability $\frac{i\mu_0}{i\mu_0 + (s-i)\mu}$, and it follows an Erlang distribution with 2 phases and $i\mu_0 + (s-i)\mu$ as a rate per stage with probability $1 - \frac{i\mu_0}{i\mu_0 + (s-i)\mu}$. This leads to

$$\begin{aligned} P(W > \tau | (x, y) = (i, 1)) = {} & \frac{i\mu_0}{i\mu_0 + (s-i)\mu} \\ & \times \frac{((i-1)\mu_0 + (s-i)\mu) e^{-\tau(i\mu_0 + (s-i)\mu)} - (i\mu_0 + (s-i)\mu) e^{-\tau((i-1)\mu_0 + (s-i)\mu)}}{\mu - \mu_0} \\ & + \frac{(s-i)\mu}{i\mu_0 + (s-i)\mu} e^{-\tau(i\mu_0 + (s-i+1)\mu)} (1 + \tau(i\mu_0 + (s-i)\mu)), \end{aligned}$$

for $0 \leq i \leq s$. One can continue in the same way to derive all the conditional waiting time probabilities for $k > 1$, which finishes the characterization of the performance measures (email throughput and call waiting time distribution) in the case of unequal service rates.

4.3.2 Construction of the Adaptative Threshold Policy

In this section, we use the previous results to find an insight on how we should adapt the threshold as a function of the intensity of the call arrivals. The objective is to maximize the throughput

of emails while reaching the constraint on the call waiting times for the whole day. We find that during the periods with low demand, the need of having a good service level is more important than during the periods with high demand. On the basis of this observation, we build a method for adapting the threshold. We then evaluate this method by comparing it with the optimal threshold policy.

Numerical Observations

For a given time interval long enough to reach the stationary regime, one can use the results of Section 4.3.1 to obtain the optimal threshold, denoted by u^*, for Problem (4.1). Consider now a working day with two time intervals, each with a different call arrival rate, and on each of which the stationary regime is reached. We want to find the optimal couple of thresholds that answers our optimization problem, where the call service level constraint is for the whole day. We denote the first (second) time interval by I_1 (I_2) and its arrival rate by λ_1 (λ_2). Without loss of generality, we consider cases where $\lambda_1 \leq \lambda_2$.

In Table 4.1, we consider various scenarios of arrival rates, service rates, and relative durations between the two time intervals. We give the optimal threshold of each interval in isolation (i.e. the highest threshold which verifies the service level constraint). They are denoted by u_1^* and u_2^* for I_1 and I_2, respectively. We also evaluate the couple of thresholds which answers Problem (4.1) on the set of the two intervals. This couple is found by an exhaustive test of all the possible values for the couple (u_1,u_2). We denote by $(u_1, u_2)^*$ this optimal couple. Remark that for this couple, Problem (4.1) does not have to be answered on each interval but on the set of the two intervals. Finally, we give the performance measures for each interval and for the set of the two intervals for the couple $(u_1, u_2)^*$.

We observe that u_1^* (respectively u_2^*) is always higher or equal to u_1 (respectively lower or equal to u_2) in the optimal couple $(u_1, u_2)^*$. An interesting insight here is that, while respecting the global call service level, we should strictly respect the service level during the interval with a small

Table 4.1: Optimal couples of thresholds ($s = 10$, $\tau = 30$ seconds, $\alpha = 80\%$)

λ_1	λ_2	μ	μ_0	I_1	I_2	u_1^*	u_2^*	$(u_1, u_2)^*$	$P(W_1 < \tau)$	$P(W_2 < \tau)$	$P(W < \tau)$	T_1	T_2	T
1	1	0.2	0.2	50%	50%	8	8	(8,8)	84.04%	84.04%	84.04%	0.758	0.758	0.758
1	1.3	0.2	0.2	50%	50%	8	6	(8,7)	84.04%	77.99%	80.62%	0.758	0.401	0.580
0.5	1.5	0.2	0.2	50%	50%	9	–	(8,4)	96.81%	74.79%	80.30%	1.169	0.055	0.611
1	1.3	0.2	0.2	67%	33%	8	6	(8,7)	84.04%	77.99%	81.66%	0.758	0.401	0.639
1	1.3	0.2	0.2	80%	20%	8	6	(8,8)	84.04%	69.15%	80.39%	0.758	0.552	0.711
0.5	1.5	0.2	0.2	90%	10%	9	–	(9,7)	88.19%	63.94%	82.13%	1.350	0.277	1.243
1	1.5	0.2	0.2	50%	50%	8	–	(7,5)	90.92%	72.93%	80.13%	0.604	0.111	0.357
1	1.5	0.2	1	50%	50%	10	–	(10,7)	89.34%	74.94%	80.70%	5.191	0.961	3.076
1	1.5	0.2	1	80%	20%	10	–	(10,10)	89.34%	67.56%	83.40%	5.191	2.908	4.734
1.3	1.4	0.2	1	50%	50%	9	8	(9,9)	83.51%	77.09%	80.18%	2.863	2.440	2.652
1.3	1.4	0.2	1	80%	20%	9	8	(9,10)	83.51%	68.19%	80.26%	2.863	3.621	3.014
1.3	1.4	1	0.2	50%	50%	9	9	(9,9)	88.63%	87.77%	88.18%	1.616	1.598	1.601
1.3	1.4	1	0.2	80%	20%	9	9	(9,10)	88.63%	60.45%	82.10%	1.616	1.794	1.742

arrival rate (I_1), and violate the constraint when the arrival rate is high (I_2). The reason is related to the sensitivity of the service level to an increase of the threshold. When the workload increases the sensitivity of the service level for a given threshold ($\Delta SL(u) = SL(u+1) - SL(u)$ for $0 \leq u < s$) first increases and then decreases (except for $u = s - 1$, the sensitivity only decreases). This can be seen in Figure 4.1.

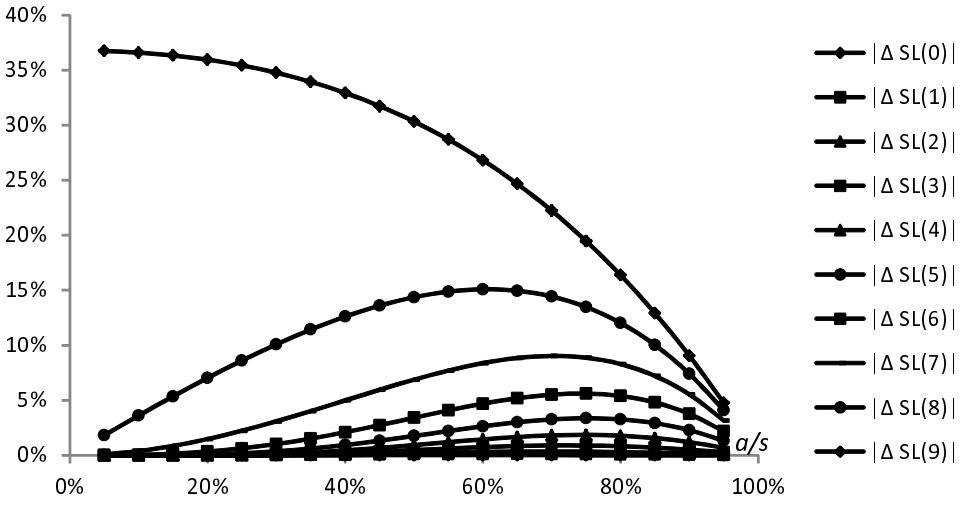


Figure 4.1: Evolution of the Sensitivity of the Service Level function of the Threshold and the Workload ($s = 10$, $\tau = 30$ seconds, $\mu = \mu_0 = 0.2$)

In practice the workload in call centers are usually higher than 80%. If a situation with a low workload happens, the threshold would increase and reach its maximal values ($u = s - 1$ or $u = s$). Since the last part of the curves ($|\Delta SL(u)|$ in function of the workload) and the whole curve for $u = s - 1$ are decreasing (see Figure 4.1) we mainly consider in practice situations where

the sensitivity of $SL(u)$ is decreasing in the workload.

Proposition 4 proves that there is less waste for the service level, when increasing the threshold in situations for which the sensitivity of the service level is decreasing in the workload.

Proposition 4 *If $\lambda_1 < \lambda_2$ and $\Delta SL(u)$ is decreasing in the workload then $|\Delta SL_{\lambda_1}(u_1^*)| \geq |\Delta SL_{\lambda_2}(u_2^*)|$.*

Proof. If $u_1^* = u_2^*$ then increasing u is less sensitive in SL_{l_2} than in SL_{l_1} since the sensitivity of SL is decreasing in the workload. The other case is $u_1^* > u_2^*$ because $a_1 < a_2$. Since SL is decreasing and concave in u, SL_{l_1} is more sensitive to the increasing of u from u_1^* than from u_2^*. From u_2^*, SL_{l_1} is more sensitive to the increasing of u than SL_{l_2}, then SL_{l_2} is less sensitive to the increasing of u from u_2^* than SL_{l_1} would be from u_1^*. □

Yet the opposite seems to be more intuitive, since it would be hard to compensate a very bad service level during an interval with a high number of calls. Counterexamples can be found when $\lambda_1 << \lambda_2$. For example for $\lambda_1 = 0.1$, $\lambda_2 = 1$, $\mu_0 = \mu = 0.2$, $s = 10$, $I_1 = I_2 = 50\%$, $\alpha = 80\%$, we have $u_1^* = 9$, $u_2^* = 8$ and $(u_1, u_2)^* = (10, 8)$.

Our Adaptive Threshold Policy (ATP)

We propose for Problem (4.1) an adaptive threshold policy which adjusts the threshold as a function of the call workload. The idea of the policy comes from the numerical observations in Section 4.3.2. As mentioned in Section 4.2, the threshold is reevaluated at the beginning of each interval i $(i = 1, ..., N)$. The threshold associated to interval i is denoted by u_i. The global service level for the whole day (all N intervals) is denoted by SL, and the global one from interval 1 to interval i is denoted by SL_i, for $i = 1, ..., N$.

If SL_i is higher (lower) than α at the beginning of an interval i $(i = 2, ..., N)$ then the policy increases (decreases) the threshold. To change the threshold, we use a real parameter denoted by c_i $(i = 1, ..., N)$. The threshold u_i is defined as the closest integer to c_i, for $i = 1, ..., N$. Note that the parameter c_i is chosen to be real in order to smooth the change in the threshold u_i. We start with

$u_1 = c_1 = s$. For $i \geq 2$, if we need to increase the threshold, then we take $c_i = c_{i-1} + 1 - c_{i-1}/s$. If not, then $c_i = c_{i-1} - c_{i-1}/s$. This policy is refereed to as ATP.

As the workload of calls decreases, ATP increases the threshold with a decreasing speed. This decreasing speed allows a slow increase in the threshold and then gives advantage to calls, which is coherent with the insight of Section 4.3.2. The opposite is also true and coherent with the insight. The advantages of ATP is that it is simple, easy to implement, and at the same time efficient as we will show later.

Evaluation of the Adaptative Threshold Policy

In this section, we evaluate the quality of the ATP policy by comparing it with the optimal one. First we provide the optimal threshold policy. Because of the discrete nature of the threshold, one may intuitively see that the threshold should vary between two or more values. The reason is that we need to satisfy exactly the constraint on calls in Problem (4.1) in order to maximize the email throughput. Both for cases $\mu_0 = \mu$ and $\mu_0 \neq \mu$, Theorem 3 provides a weak condition under which the optimal policy is a randomization of the threshold between two values.

Theorem 3 *Consider $0 \leq u_1, u_2 \leq s$ such that $SL(u_1) \leq \alpha \leq SL(u_2)$. If there exists $\gamma \in \mathbb{R}$ for which $(u_1, u_2) \in \arg\max_u T(u) + \gamma SL(u)$, then randomizing between u_1 and u_2 is optimal.*

Proof. Let $p \in [0, 1]$ be the parameter of randomization between u_1 and u_2. Assume that we can find a couple $(u_3, u_4) \neq (u_1, u_2)$ and a parameter of randomization $q \in [0, 1]$ such that the constraint on calls is also saturated and $SL(u_3) \leq \alpha \leq SL(u_4)$. We have $pT(u_1) + (1-p)T(u_2) + \gamma pSL(u_1) + \gamma(1-p)SL(u_2) \geq qT(u_3) + (1-q)T(u_4) + \gamma qSL(u_3) + \gamma(1-q)SL(u_4)$. Since $\gamma pSL(u_1) + \gamma(1-p)SL(u_2) = \gamma qSL(u_3) + \gamma(1-q)SL(u_4) = \gamma\alpha$, we deduce that $pT(u_1) + (1-p)T(u_2) \geq qT(u_3) + (1-q)T(u_4)$. Then the couple (u_1, u_2) is optimal, which completes the proof. □

The randomization between two thresholds allows for the constraint on calls to be met exactly. For our system with constant parameters, we believe that the randomization is between two suc-

Table 4.2: Comparison under steady-states assumption (θ=15min)

	Optimal c	Optimal T	ATP T	Difference
Scenario 1 ($\lambda = 4$, $\mu = \mu_0 = 0.2$, $s = 28$)	25.49	1.39	1.37	1.46%
Scenario 2 ($\lambda = 0.02$, $\mu = \mu_0 = 0.2$, $s = 1$)	0.13	0.02	0.02	0.00%
Scenario 3 ($\lambda = 18$, $\mu = \mu_0 = 0.2$, $s = 100$)	93.91	1.65	1.58	4.43%
Scenario 4 ($\lambda = 4$, $\mu = 0.27$, $\mu_0 = 0.15$, $s = 28$)	26.63	1.89	1.89	0.00%
Scenario 5 ($\lambda = 4$, $\mu = 0.17$, $\mu_0 = 1$, $s = 28$)	23.21	2.00	1.79	11.73%

cessive thresholds. Since the throughput is neither convex nor concave it is difficult to rigorously prove this result. However, if we denote by u^* ($0 \leq u^* \leq s$) the highest threshold that verifies $SL(u^*) > \alpha$, we numerically checked that with $\gamma = -\frac{T(u^*+1)-T(u^*)}{SL(u^*+1)-SL(u^*)}$ (for $0 \leq u^* < s$), the expression $T(u)+\gamma \times SL(u)$ is strictly increasing from $u = 0$ to $u = u^*$, strictly decreasing from $u = u^*+1$ to $u = s$ and $T(u^*) + \gamma SL(u^*) = T(u^* + 1) + \gamma SL(u^* + 1)$. Then for all the considered numerical situations the optimal policy is a randomization between two adjacent values when $0 \leq u^* < s$. When $u^* = s$, the optimal policy is to keep the threshold constant and equal to s.

In Table 4.2, we propose 5 representative scenarios with constant arrival rates and compare the optimal throughput with the one found with under ATP. A comprehensive numerical study can be found in Section C.2 of the appendix. Although the ATP method is not optimal, the difference with the optimum is quite small. This shows the advantage of ATP in the case of constant arrival rates. Recall that our main purpose in this chapter is the analysis of the case with a fluctuating arrival rate. In the next section, we consider the case of a fluctuating arrival rate and evaluate the performance of ATP through a comparison with other intuitive methods.

4.4 Non-Constant Arrival Rates

In Section 4.4.1 we compare ATP with methods that use constant step sizes. Then in Section 4.4.2 we analyze the impact of the parameters on the choice of the method. In Section 4.4.3 we propose some other intuitive adaptive methods.

We consider cases where the length of the working day equals eight hours (L = 8h) and a frequent possibility of reevaluating the real threshold c, at the beginning of each time interval with length $\theta = 1$ min, 5 min or 15 min. We use simulation to obtain the performance measures. For each scenario, we run n replications. We then introduce a measure of the bias after the n simulations, denoted by $\overline{r}_n$ and calculated as $\overline{r}_n = \frac{\sum_{k=1}^{n} \text{Max}(\alpha - \overline{SL}_k, 0)}{n}$, where $\overline{SL}_k$ is the expected service level of simulation k ($1 \leq k \leq n$). Since the value of $\overline{r}_n$ should be as small as possible, we introduce a coefficient A which would be the aversion of the call center manager to the risk and introduce an utility indicator denoted by U_n and given by $\overline{T}_n - A \times \overline{r}_n$, where $\overline{T}_n$ is the average throughput after n simulations. The confidence intervals are a safe way to evaluate the required number of equivalent simulations, n. The confidence interval for a proportion p and a risk of 5% is $\left(p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}\right)$ in which n is the number of terms used to calculate the proportion p. If we want a precision of one decimal we need $2 \times 1.96\sqrt{\frac{0.8(1-0.8)}{n}} < 0.001$ then $n >$2 458 624. In order to have safe results we run each simulation 3 000 000 times.

4.4.1 Comparison with Constant Step Methods

We propose different scenarios to compare ATP with constant step size methods. We denote by h the step size ($0 < h \leq 1$). When we need to increase (respectively decrease) the real threshold c_i after i intervals ($1 \leq i < N$) under the case $SL_i > \alpha$ (respectively $SL_i < \alpha$) we add h to c_i (respectively we add $-h$ to c_i). In each scenario we use an aversion of risk equal to 100 and initialize the system with $c_0 = u_0 = s$. In some scenarios the number of agents varies over the day. When the number of agents decreases, we could be in a situation in which $c > s$, i.e., the number of busy agents becomes higher than the new value for s. To avoid such a situation, we force in the simulation the change of c to the new smaller value of s. Any undertaken task by a removed agent is lost. In all scenarios the constraint on calls is such that the proportion of calls that wait less than 30 seconds is at least 80%, τ = 30s and α =80%. We consider the following scenarios:

- **Scenario 1:** $\lambda = 4$, $\mu = \mu_0 = 0.2$, $s = 28$ and $N = 480$ ($\theta = 1$ min),
- **Scenario 2:** $\lambda = 4$, $\mu = \mu_0 = 0.2$, $s = 28$ and $N = 32$ ($\theta = 15$ min),
- **Scenario 3:** $\lambda = 4$, $\mu = 0.27$, $\mu_0 = 0.15$, $s = 28$ and $N = 480$,
- **Scenario 4:** $\lambda = 4$, $\mu = 0.17$, $\mu_0 = 1$, $s = 28$ and $N = 480$,
- **Scenario 5:** λ linearly decreasing from 5 to 3 , $\mu = \mu_0 = 0.2$, $s = 34$ if $\lambda > 4.5$, $s = 28$ if $4.5 > \lambda > 3.5$, $s = 23$ in the remaining cases, and $N = 480$,
- **Scenario 6:** λ linearly increasing from 3 to 5 , $\mu = \mu_0 = 0.2$, $s = 34$ if $\lambda > 4.5$, $s = 28$ if $4.5 > \lambda > 3.5$, $s = 23$ in the remaining cases, and $N = 480$,
- **Scenario 7:** During the first quarter of the period λ is linearly increasing from 1 to 5, during the second quarter λ is linearly decreasing from 5 to 3, during the third quarter λ is linearly increasing from 3 to 5 and during the last quarter λ is linearly decreasing from 5 to 1, $\mu = \mu_0 = 0.2$, $s = 34$ if $\lambda > 4.5$, $s = 28$ if $4.5 > \lambda > 3.5$, $s = 23$ in the remaining cases, and $N = 480$,
- **Scenario 8:** The period T is divided into 10 sub-periods and the value of λ alternates between the values 5 and 0.5, i.e., it is 5 in the first sub-period, 0.5 in the second one, again 5 in the third one, and so on, $\mu = \mu_0 = 0.2$, $s = 28$ and $N = 480$.

The results are presented in Table 4.3. We consider values of h equal to 0.1, 0.2, 0.5 and 1. We observe that ATP performs better or at least similarly to the constant step methods with an aversion of risk equal to 100.

In Figures 4.2(a), 4.2(b) and 4.2(c), we present the evolution of the threshold, the proportion of customers that wait less than 30 seconds and the email throughput as a function of time in one simulation of scenario 2. This is an illustration that could help to understand why ATP is efficient. With a small value of h ($h = 0.2$), the initialization has an important impact on the evolution of

Table 4.3: Comparison between ATP and constant step methods

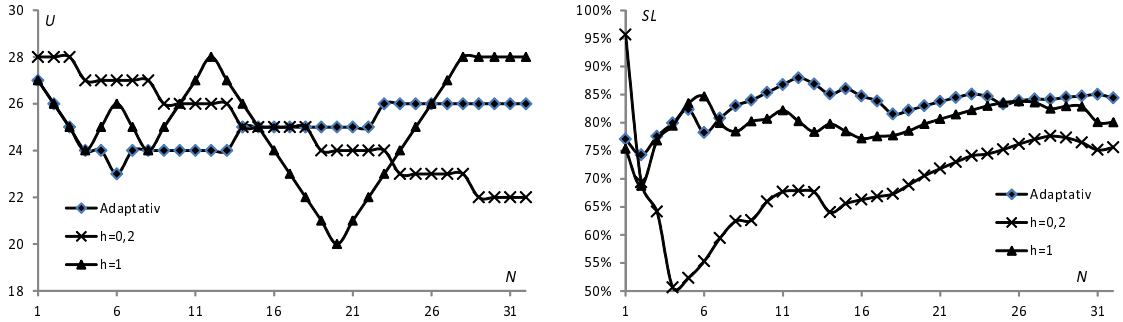
	h	$\overline{T}$	$\overline{SL}$	$\overline{r}$	U		h	$\overline{T}$	$\overline{SL}$	$\overline{r}$	U
Sc 1	0.1	1.17	80.6%	0.0046	0.71	Sc 2	0.1	1.53	72.15%	0.0782	-6.3
	0.2	1.12	80.5%	0.0036	0.77		0.2	1.38	78.7%	0.0201	-0.63
	0.5	1.04	80.1%	0.0032	0.72		0.5	1.23	81.4%	0.0063	0.60
	1	0.98	80.0%	0.0035	0.63		1	1.19	80.7%	0.0062	0.57
	ATP	1.09	80.7%	0.0027	0.82		ATP	1.12	85.6%	0.0008	1.04
Sc 3	0.1	1,85	80,3%	0,0023	1.62	Sc 4	0.1	3.20	78.9%	0.0314	0.06
	0.2	1,80	80,3%	0,0017	1.63		0.2	3.07	79.5%	0.0277	0.30
	0.5	1,68	80,3%	0,0013	1.55		0.5	3.05	79.2%	0.0278	0.27
	1	1,57	80,2%	0,0014	1.43		1	3.14	79.2%	0.0281	0.33
	ATP	1,72	81,0%	0,0003	1.68		ATP	2.95	78.9%	0.0264	0.31
Sc 5	0.1	1.19	79.9%	0.0067	0.52	Sc 6	0.1	1.05	83.2%	0.0014	0.91
	0.2	1.13	80.1%	0.0037	0.76		0.2	1.04	81.7%	0.0021	0.83
	0.5	1.08	80.0%	0.0033	0.75		0.5	1.04	80.8%	0.0025	0.79
	1	1.01	79.9%	0.0033	0.68		1	1.04	80.2%	0.0032	0.72
	ATP	1.12	80.4%	0.0018	0.93		ATP	1.09	82.1%	0.0007	1.02
Sc 7	0.1	1.38	81.6%	0.0010	1.28	Sc 8	0.1	3,04	78,8%	0,0246	0,59
	0.2	1.37	81.2%	0.0015	1.21		0.2	2,84	79,5%	0,0201	0,83
	0.5	1.27	80.4%	0.0017	1.10		0.5	2,72	79,3%	0,0178	0,94
	1	1.24	80.3%	0.0011	1.14		1	2,76	78,9%	0,0188	0,88
	ATP	1.38	81.2%	0.0005	1.33		ATP	2,83	79,7%	0,0172	1,11

the threshold. At the beginning with $u_0 = c_0 = s = 28$, there is a need to decrease the threshold. A small value of h does not allow to do this decreasing quickly enough. Then there is a need to keep on decreasing the threshold in order to have a chance to reach the service level on calls over the whole day. On the other hand a high value of h ($h = 1$) goes with a fluctuation of the threshold, with sometimes bad call service levels and other times bad email throughput. Note that the higher is h, the faster the service level converges its target. In what follows we go further in analyzing the impact of the main parameters on the choice of h.

4.4.2 Impact of the Parameters

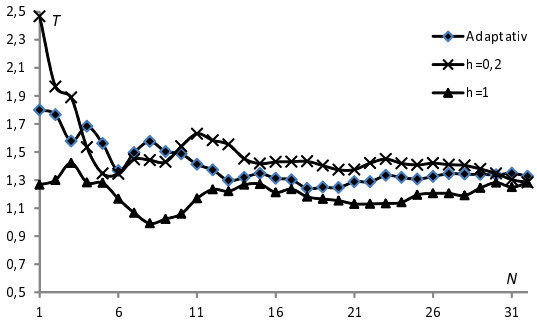
In this section, we analyze the impact of the parameters in the choice of a constant value for h. This might permit opportunities for providing new methods for adapting h and also help to understand the performance of ATP.

Impact of the Number of Intervals, N: The comparison between scenarios 1 and 2 in Table 4.3 indicates that there is a link between h and N. In scenario 2 with only 32 intervals, a small



(a) Threshold

(b) Service Level



(c) Email Throughput

Figure 4.2: Evolution of the threshold, the service level and the throughput (Scenario 2)

Table 4.4: Optimal h and N ($\lambda = 4$, $s = 28$, $\mu = \mu_0 = 0.2$, $\alpha = 80\%$, $\tau = 30$s, $N = 480$ and $L = 480$ min)

h	0.1	0.2	0.5	1
N	82	53	40	9
T	1,34	1,33	1,31	1,30
SL	80,03%	80,01%	80,03%	80,04%

value of h does not allow to reach the call service level constraint. In scenario 1, a large number of intervals and a high value for h lead to an important fluctuation from $u = 0$ to $u = s$. For a given value of h there should be an optimal number of intervals. In Table 4.4 we present an example of optimal values of N for a given h. We observe that increasing N decreases in h.

An advantage of ATP is its ability of adaptation to the number of intervals. A high number of intervals could lead to a high probability of reaching extreme and inefficient states ($u = 0$ and $u = s$ usually). Thanks to the slowing in the speed of the threshold reduction when c is small and the slowing in the threshold increasing when c is high, this is unlikely to happen. ATP also provides a high capacity of reaction when the threshold is too low or too high, which is important

in the case of a small number of intervals.

Impact of the Aversion of Risk A: The choice of the method for adapting the threshold depends on the risk aversion of the manager. In our simulations we considered that $A = 100$ and showed the efficiency of ATP. This value provides a balance between the performance ($\overline{T}$) and the risk ($\overline{r}$). One further question would be about the choice of the manager if her aversion to the risk is different. If we consider the extreme case of an infinite aversion to the risk ($A = \infty$), the choice will be made for the smallest value of $\overline{r}$. We observe that the choice will still be for ATP rather than the other methods even more than with $A = 100$. This is an important advantage of ATP; it is a safe method (i.e., the probability the reach the service level constraint at the end of the working period is high). Since the threshold c is usually closer to s than to 0 because of the concavity in the service level we usually have a higher speed in decreasing the threshold than in increasing it, which is safe and explains the small values for $\overline{r}$. On the other hand if the manager has no risk aversion ($A = 0$) then the choice will be made for the highest average throughput ($\overline{T}$). ATP is then not the best one but it still provides results close to the best ones in Table 4.3.

Impact of the Arrival Process: From Table 4.3 we observe that the choice of a constant step h is influenced by the volatility of the arrival process (comparison between scenarios 1 and 8). When the arrival rate is constant the optimal solution may be an oscillation between a threshold u and a threshold $u+1$. Then, in order to reach this solution, we need small values of h in order not to leave too often those two states. When the arrival rate has a high volatility, there is a need to change the threshold very often in order to reach a good value of u in a short time. This explains why $h = 0.2$ is better than $h = 0.5$ in scenario 1 and the opposite is true in scenario 8. The comparison between scenarios 5 and 6 is also interesting; in scenario 5 the arrival rate decreases over the day and in scenario 6, it increases. We observe that small values of h are better in scenario 6 because of the initial conditions; $u_0 = c_0 = s$. In scenario 5 those initial conditions with a small value of

h lead to a difficulty in reaching a good threshold. We observe that ATP performs well in various scenarios of volatility in the arrival process thanks to its capacity of creating slow or fast changes in the threshold.

Impact of the Emails Service Rate: Consider scenarios 3 and 4. We observe that ATP performs better when the emails are served slower (scenario 3) than when they are served faster (scenario 4) than the calls. Because of the concavity of the service level, the threshold is usually closer to s than to 0. This implies a higher speed in decreasing than in increasing the threshold. When the emails are served faster than the calls, the need to increase c is more important because an email does not occupy an agent for a long period of time but with our method this increasing might be too slow. However, we notice that this case has less meaning for our study since the problem of reserving agents is interesting in the case of long service times for background jobs (relatively to calls).

4.4.3 Comparison with other Intuitive Methods

In this section, we compare ATP with other intuitive adaptive methods. We propose the following ones based on the reevaluation of the step h_i after each intervals i ($i = 1, ..., N$).

Method 1: The first intuitive idea is to propose a decision based on the distance from the achieved service level and the target after each interval. The intuition is that the need to change the threshold increases with this distance. We initialize with $h_0 = 0$, $c_0 = u_0 = s$ and $SL_0 = 100\%$. After each interval we reevaluate h_i according to the relation:

$$h_{i+1} = \text{Min}\left\{1, |SL_i - \alpha|\right\},$$

for $i \in \{0, ..., N-1\}$.

Method 2: Method 2 is a variation of Method 1. We propose a decision based on the cumulative distance with the service level target, α. The intuition is that the need to change the threshold not only increases with the distance to the target service level, but also increases with the time spent above or under this target. More precisely, we initialize by $h_0 = 0$, $c_0 = u_0 = s$ and $SL_{-1} = SL_0 = 100\%$. After each interval we reevaluate h_i according to the relation:

$$h_{i+1} = \text{Min}\left\{1, h_i + |SL_i - \alpha|\right\} \times \mathbb{I}_{(SL_i-\alpha)(SL_{i-1}-\alpha)>0},$$

for $i \in \{0, ..., N-1\}$.

Method 3: We propose the same evaluation of h_i as in Method 2 but instead of using the service level SL_i of the last i intervals ($i = 1, \cdots, N$), we use the service level measured only on the last interval i ($i = 1, \cdots, N$). This method is made to correct a too important weight that could be given to the past in the previous method.

Methods 4a and 4b: Methods 4a and 4b are not really intuitive. The idea behind them is the question of when the strongest decisions in the change of the threshold should be taken. If we choose the strongest changes in the threshold at the beginning of the period we could quickly reach the service level constraint (Method 4a). If we choose the strongest changes at the end of the period we could maximize the email throughput at the beginning and do an efficient correction at the end of the working period to reach the service level constraint (Method 4b). More precisely, in Method 4a we propose after i intervals to choose

$$h_i = 1 - \frac{i}{N},$$

and in Method 4b we choose

$$h_i = \frac{i}{N},$$

Table 4.5: Comparison of the methods

	h	$\overline{T}$	$\overline{SL}$	$\overline{r}$	U		h	$\overline{T}$	$\overline{SL}$	$\overline{r}$	U
	$M1$	1.01	79.9%	0.0038	0.62		$M1$	1.25	80.45%	0.0080	0.44
	$M2$	0.99	79.9%	0.0037	0.62		$M2$	1.13	81.2%	0.0068	0.45
	$M3$	1.45	72.6%	0.0743	-5.98		$M3$	1.59	68.0 %	0.0639	-4.79
	$M4a$	1.01	80.9%	0.0041	0.59		$M4a$	1.24	80.7%	0.0090	0.34
Sc 1	$M4b$	1.06	80.0%	0.0029	0.76	Sc 2	$M4b$	1.20	80.2%	0.0099	0.22
	0.1	1.17	80.6%	0.0046	0.71		0.1	1.53	72.15%	0.0782	-6.3
	0.2	1.12	80.5%	0.0036	0.77		0.2	1.38	78.7%	0.0201	-0.63
	0.5	1.04	80.1%	0.0032	0.72		0.5	1.23	81.4%	0.0063	0.60
	1	0.98	80.0%	0.0035	0.63		1	1.19	80.7%	0.0062	0.57
	ATP	1.09	80.7%	0.0027	0.82		ATP	1.12	85.6%	0.0008	1.04

for $i = 1, \cdots, N$.

We compare the proposed methods in Table 4.5 with the constant step sizes methods and ATP under scenarios 1 and 2. We observe that those methods are not as good as ATP and even sometimes not as good as the constant step size methods. Methods 4a and 4b are not efficient for a simple reason; the choices in changing the threshold mainly depend on the demand and not on the closeness to the end of the working day. We observe on other simulations that Method 4a is efficient when the variability in the demand is high at the beginning of the working period and the opposite is true for Method 4b. Although Methods 1 and 2 are the most intuitive, we observe that they are not efficient. The weight of the past is too heavy and entails extreme choices in the threshold (which are often inefficient) so as to compensate the past values. Method 3 is often more efficient in terms of the email throughput, however it converges very slowly. We observe on other simulations that Method 3 could be a good proposition only if a working day is long enough (at least 1000 hours). An intermediate solution between Methods 2 and 3 would be to propose a decision in the changes of the threshold based on the average of the service levels measured on all past intervals weighted by coefficients which are increasing with the proximity to the last interval. Many solutions can be proposed in that direction but none of them seems to be efficient for a representative number of scenarios.

4.5 Conclusion and Future Research

We considered call centers with inbound calls and an infinite supply of non-preemtable outbound jobs. We proposed a scheduling policy, refereed to as ATP, where the objective is to do as much outbound as possible while satisfying a service level constraint on the call waiting time. In the real-life call center context with fluctuating call arrival rate, the assignment policy for outbound adapts itself to the current service level. We showed the efficiency of ATP by comparing its performance with those of other policies. One of the main advantage of ATP is its ability to quickly react when an important change in the arrival process happens and also its ability to avoid inefficient states when the arrival rate remains constant.

Future research on this subject may follow two directions. First, a theoretical modeling for the adaptive blending might be useful to better understand ATP. This is now hindered by the fact that no theory seems to exist on this type of control problems. One of the difficulties to build a Markov chain is the non-exponentiality of the decision interval length defined in the ACD. Another difficulty is the lack of transient results for the performance measures of queueing call center models. Second, the complexity of a real-life call center has been partly avoided in our study. Features such as abandonments, retrials, different types of the inbound calls, switching times between different tasks, and the finite number of back office tasks, are important but including them considerably complicate the analysis.

Chapter 5

Call Centers with a Callback Option

We consider a call center model with a callback option, which allows to transform an inbound call into an outbound one. The optimization problem consists on minimizing the expected waiting time of the outbound calls while respecting a service level constraint on the inbound ones. We propose a routing policy with two thresholds, one on the reservation of agents for inbound calls, and another on the number of waiting outbound calls. The purpose of this study is to determine a curve relating the two thresholds. The paper version of this chapter is the ongoing paper Legros et al. (2013b).

5.1 Introduction

In the context of highly congested call centers, the use of an alternative service channel can be proposed to customers so as to balance workload and avoid excessive abandonments. Such an alternative channel could be a Web site, an e-mail service, a proposition to the customer to call at a less busy time, a possibility of routing to another team of agents (see Manitz and Stolletz (2013)) or to leave a number and be called back later. We focus on this last alternative. Armony and Maglaras (2004a,b) propose a model in which customers are given a choice of whether to wait online for their call to be answered or to leave a number and be called back within a specified time. They show that this callback scheme allows the system to increase its performance.

Combining inbound and outbound calls leads to a call center with blended operations. The key distinction of problems with blending comes from the fact that callbacks have less urgency relative to inbound calls. The call blending problem has led to research on performance evaluation (see Bernett et al. (2002), Pichitlamken et al. (2003) and Deslauriers et al. (2007)) and analysis of blending policies (see Gans et al. (2003), Bhulai and Koole (2003), Armony and Maglaras (2004a) and Legros et al. (2013a)). Because of the strict waiting time requirement on inbound calls, it is best to give them priority over the outbound calls. To maximize agent productivity it would be best to assign a call (inbound or outbound) to every idle agent when it is possible. The objection against this policy is that it could lead to long waiting times for future arrivals from inbound calls. In Bhulai and Koole (2003) and Gans and Zhou (2003b), it is shown under certain assumptions that the optimal assignment policy is of the following form: outbound calls should only be scheduled when there are no waiting inbound calls and when the number of idle agents exceeds a certain threshold.

In the case of a callback option, this policy has to be completed. The blending models in the literature usually assume an infinite number of outbound tasks. With a callback option this number is finite and depends on all the parameters of the system. Thus this number should be considered

as another threshold for the choice of the threshold in the agents reservation. Otherwise, it could lead to long waiting times for outbound calls.

In this chapter, we consider a call center modeling with a single customer type and a callback option. We develop a method based on Markov chains to evaluate its performance measures. We also provide an efficient routing policy based on the number of reserved agents for the inbound calls and the number of outbound calls. The overall objective is to reach a certain long term service level for the inbound calls, while minimizing the expected waiting time of the outbound ones. Few of our key findings related to the performance measures are highlighted next. We prove that the expected waiting time for the outbound calls is decreasing in the proportion of outbound calls and in the threshold of reserved agents. We also show that sensitivity of the performance measures is higher on the threshold of reserved agents than on a limit in the number of outbound calls. These findings provide an intuitive justification of the optimal curve relating the two thresholds.

The rest of the chapter is organized as follows. In Section 5.2 we present the modeling of a call center with a callback option. In Section 5.3 we evaluate the performance measures and study the impact of the main parameters. In Section 5.4, we numerically study the general form of the optimal curve. We finally give some concluding remarks and highlight future research.

5.2 Model

We consider a call center modeled as a multi-server queueing system with two types of jobs, inbound and outbound calls. The arrival process of inbound calls is assumed to be a homogeneous Poisson process with rate λ. Inbound calls arrive at a dedicated first come, first served (FCFS) queue with infinite capacity (queue 1). There are s identical, parallel servers (agents in call center parlance). Each agent can handle both types of jobs. We assume that the service times of inbound or outbound calls are exponentially distributed with the same rate μ. Neither abandonment nor retrials are modeled.

Inbound calls are more important than outbound ones in the sense that the former request a quasi-instantaneous answer (waiting time in the order of seconds or minutes), while the latter are more flexible and could be delayed for several hours. The possibility of the callback option is given in this purpose. The objective of the call center manager is to minimize the expected waiting time for the outbound calls while satisfying a strict constraint on the inbound calls expected waiting time.

The functioning we consider for the callback option and the related customers behavior is inspired by Armony and Maglaras (2004a) and described as follow. This option is proposed to a newly arrival only when her expected waiting time is too long (i.e. too many waiting calls in queue 1). We define a limit k ($k \in \mathbb{N}$) in the number of waiting calls in queue 1. Upon her arrival, a customer can find three situations. If at least one agent is available then she is routed to one of the idle agents. Otherwise if the number of waiting calls in queue 1 is strictly lower than k, the callback option is not proposed and she waits in queue 1. If this number is higher than or equal to k the callback option is proposed. The customer reaction is assumed to be probabilistic. Directly upon arrival, she decides to accept the callback with a probability q or she decides to wait in queue 1 with a probability $1 - q$ ($q \in [0, 1]$). Note that the callback option is only proposed upon arrival. If a callback is asked, the call arrives at a dedicated FCFS queue with infinite capacity (queue 2).

The calls in queue 1 (inbound calls) have a non preemptive priority over the calls in queue 2 (outbound calls). As mentioned in Chapter 4, Bhulai and Koole (2003) prove for a similar model that the optimal policy is a threshold policy with the priority given to inbound calls (some servers reserved for inbound calls). More concretely, the functioning of the call center under a threshold policy is as follows. Let us denote the threshold by c, $1 \leq c \leq s$ (Note that the case $c = 0$ can not be considered for stability reasons). When an agent becomes idle, she handles the call at the head of queue 1, if any. If not, the agent may either handle a call at the head of queue 2 if any, or she remains idle. If the number of idle agents (excluding her) is at least $s - c$, then the agent

in question handles an outbound call (from queue 2). Otherwise, she remains idle. In other words, there are $s - c$ agents that are reserved for inbound calls.

In this chapter, we propose a threshold policy which adjusts the threshold as a function of the number of outbound calls in queue 2, denoted by y ($y \in \mathbb{N}$). We want to minimize the expected waiting time of the outbound calls, denoted by $E(W_2)$ and respect a constraint of service level, denoted by w_1^* on the expected waiting time of the inbound calls, denoted by $E(W_1)$. We are also interested for the analysis on the performance in terms of proportion of customers who asks for a callback, denoted by π. In summary, our optimization problem can be formulated as

$$\begin{cases} \text{Minimize } E(W_2) \\ \text{subject to } E(W_1) \leq w_1^*, \end{cases} \tag{5.1}$$

where the decision variable is $c(y)$. We provide a relation between the threshold c and the number of calls in queue 2 so as to build the optimal curve that answer Problem (5.1).

5.3 Performance Measures Results in the Case $c(y) = c$

In this section we simplify the model by assuming that $c(y) = c$ in order to evaluate the main performance measures ($E(W_1)$, $E(W_2)$ and π). We first analyze the underlying Markov chain. Then we determine closed form expressions for $E(W_1)$ and π and a numerical method to exactly evaluate $E(W_2)$. Second we examine the behavior of the performance measures as a function of k, q and c. Finally, we reconsider the modeling by adding a limit in the capacity of the second queue.

Let us define the random process $\{(x(t), y(t)), t \geq 0\}$ where $x(t)$ and $y(t)$ denote the number of calls in queue 1 or in service and the number of outbound calls at a given time $t \geq 0$, respectively. We have $x(t), y(t) \in \{0, 1, 2, ...\}$, for $t \geq 0$. Since call inter-arrival and service times are exponentially distributed, $\{(x(t), y(t)), t \geq 0\}$ is a Markov chain. We denote by $p_{x,y}$ the steady-state probability to be in state (x, y) ($x, y \in \mathbb{N}$), and by a the ratio $\frac{\lambda}{\mu}$. Because of the priority for inbound calls we

can find $y > 0$ if and only if $x \geq c$. For $0 \leq x < s + k$ and $y \geq 0$ the transition rate from state (x, y) to state $(x + 1, y)$ is λ. For $x \geq s + k$ and $y \geq 0$ the transition rate from state (x, y) to state $(x + 1, y)$ is $(1 - q)\lambda$. For $x \geq 1$, $x \neq c$ and $y \geq 0$ the transition rate from state (x, y) to state $(x - 1, y)$ is $\min(x, s)\mu$. For $x \geq s + k$ and $y \geq 0$ the transition rate from state (x, y) to state $(x, y + 1)$ is $q\lambda$. For $x = c$ and $y \geq 1$ the transition rate from state (c, y) to state $(c, y - 1)$ is $c\mu$. The stability condition is $\frac{a}{s} < 1$. Figure 5.1 illustrates the Markov chain.

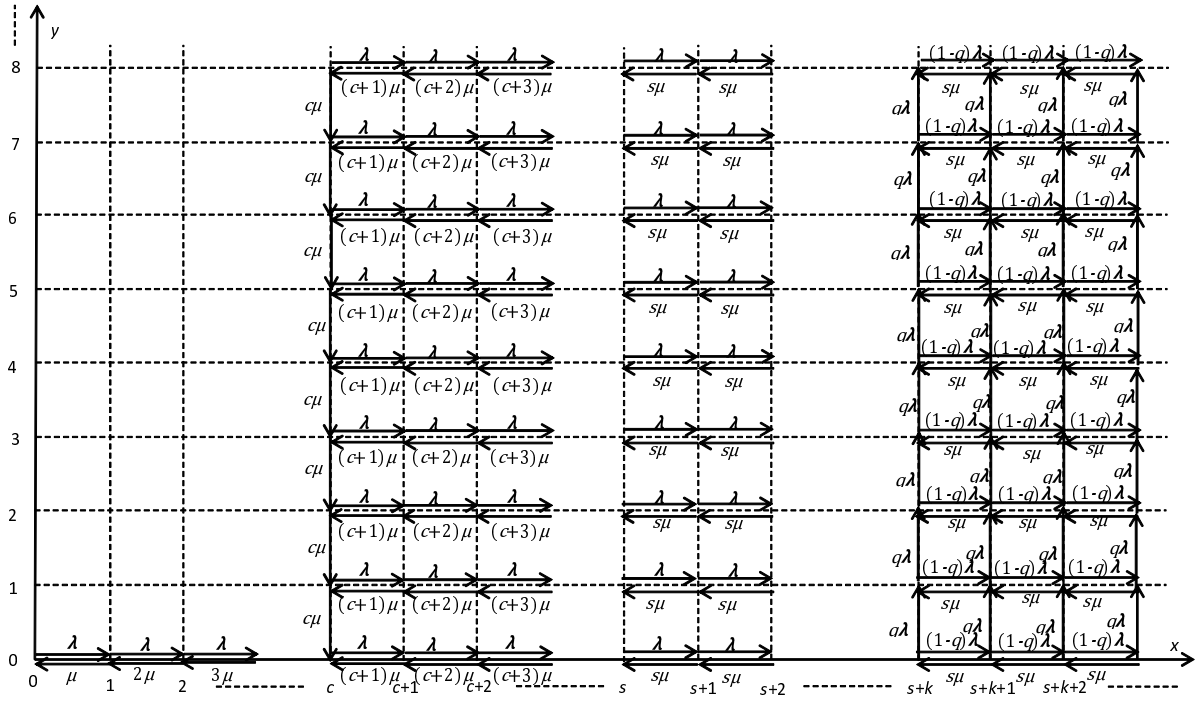


Figure 5.1: Markov chain for the callback option, $c(y) = c$

For $0 \leq x \leq c$ we have

$$\lambda p_{x,0} = (x + 1)\mu p_{x+1,0}.$$

Thus

$$p_{c,0} = \frac{a^c}{c!} p_{0,0}.$$

We denote by $P_x = \sum_{i=0}^{\infty} p_{x,i}$ for $x \geq c$. Thus for $c \leq x < s$ we have

$$\lambda P_x = (x+1)\mu P_{x+1},$$

then

$$P_{c+x} = \frac{a^x c!}{(c+x)!} P_c.$$

We also have

$$P_{s+x} = \frac{a^x}{s^x} P_s,$$

for $0 \leq x \leq k$ and

$$P_{s+k+x} = \frac{((1-q)a)^x}{s^x} P_{s+k},$$

for $x \geq 0$. With $c\mu(P_c - p_{c,0}) = \lambda q \sum_{x=0}^{\infty} P_{s+k+x}$, we obtain

$$P_c = \frac{\frac{a^c}{c!} p_{0,0}}{1 - \frac{q}{c} \frac{a^{s-c} c!}{s!} \frac{a^k}{s^k} \frac{1}{1 - \frac{a(1-q)}{s}}}.$$

Since the overall sum of the probabilities equals to one, we have

$$p_{0,0} = \left[\sum_{x=0}^{c-1} \frac{a^x}{x!} + \frac{\frac{a^c}{c!}}{1 - \frac{q}{c} \frac{a^{s-c} c!}{s!} \frac{a^k}{s^k} \frac{1}{1 - \frac{a(1-q)}{s}}} \left(\sum_{x=0}^{s-c-1} \frac{a^x c!}{(c+x)!} + \frac{a^{s-c} c!}{s!} \sum_{x=0}^{k-1} \frac{a^x}{s^x} + \frac{a^{s-c} c!}{s!} \frac{a^k}{s^k} \frac{1}{1 - \frac{a(1-q)}{s}} \right) \right]^{-1}. \tag{5.2}$$

Thus we have a close form formula for $p_{x,0}$ for $0 \leq x \leq c$. For $x \geq 0$, we have $(1-q)\lambda p_{s+k+x,0} + s\mu p_{s+k+x+2,0} = (\lambda + s\mu) p_{s+k+x+1,0}$. The associated homogeneous equation is $s\mu x^2 - (\lambda + s\mu)x + (1-q)\lambda = 0$. The solutions of this equation are $x_1 = \frac{\lambda + s\mu + \sqrt{(\lambda+s\mu)^2 - 4s\mu(1-q)\lambda}}{2s\mu} = \frac{1}{2}\left(1 + \frac{a}{s} + \sqrt{(1+\frac{a}{s})^2 - \frac{4(1-q)a}{s}}\right)$ and $x_2 = \frac{1}{2}\left(1 + \frac{a}{s} - \sqrt{(1+\frac{a}{s})^2 - \frac{4(1-q)a}{s}}\right)$. Then $p_{s+k+x,0} = \alpha_1 x_1^x + \alpha_2 x_2^x$. We have $p_{s+k,0} = \alpha_1 + \alpha_2$ and because $c\mu p_{c,1} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,0}$ we also have $c\mu p_{c,1} = q\lambda \left(\frac{\alpha_1}{1-x_1} + \frac{\alpha_2}{1-x_2} \right)$. Thus we define α_1 and α_2 as a function of $p_{s+k,0}$ and $p_{c,1}$. Since we have $c\mu p_{c,1} + (c+1)\mu p_{c+1,0} =$

$\lambda p_{c,0},\ p_{c+1,0},\ \lambda p_{c+x,0} + (c+x+2)\mu p_{c+x+2,0} = (\lambda + (c+x+1)\mu)p_{c+x+1,0}$ for $0 \leq x \leq s-c-2$ and $\lambda p_{s+x,0} + s\mu p_{s+x+2,0} = (\lambda + s\mu)p_{s+x+1,0}$ for $0 \leq x \leq s-2$, we evaluate $p_{c+x,0}$ as a function of $p_{c,1}$ for $x > 0$. Finally with the expression of $p_{s+k+x,0}$, we can compute $p_{c,1}$ as a function of $p_{s+k,0}$. The others probabilities $p_{s+k+x,y}$ for $x \geq 0$ and $y > 0$ are solutions of $(1-q)\lambda p_{s+k+x,y} + s\mu p_{s+k+x+2,y} + q\lambda p_{s+k+x+1,y-1} = (\lambda + s\mu)p_{s+k+x+1,y}$. Consequently, we have $p_{s+k+x,y} = Q_{1,y}(x)x_1^x + Q_{2,y}(x)x_2^x$ in which $Q_{1,y}(x)$ and $Q_{2,y}(x)$ are polynomials in the variable x with a degree equals to y. We evaluate the first coefficients of those polynomials with the ones of $Q_{1,y-1}(x)$ and $Q_{2,y-2}(x)$. The constant term in $Q_{1,y}(x)$ and in $Q_{1,y}(x)$ is found as a function of $p_{c,y+1}$ with the relation $c\mu p_{c,y+1} = q\lambda \sum_{x=0}^{\infty} p_{s+k+x,y}$. Then we compute $p_{c,y+1}$ using the relations $\lambda p_{c+x,y} + (c+x+2)\mu p_{c+x+2,y} = (\lambda + (c+x+1)\mu)p_{c+x+1,y}$ for $0 \leq x \leq s-c-2$ and $\lambda p_{s+x,y} + s\mu p_{s+x+2,y} = (\lambda + s\mu)p_{s+x+1,y}$ for $0 \leq x \leq s-2$.

Proportion of called back. We have

$$\pi = q\sum_{x=0}^{\infty} P_{s+k+x} = qP_c \frac{a^{s-c}c!}{s!}\frac{a^k}{s^k}\frac{1}{1-\frac{a(1-q)}{s}} = \frac{qP_{s+k}}{1-\frac{a(1-q)}{s}}.$$

Expected waiting time for the first queue. We have $\lambda E(W_1) = \sum_{x=0}^{\infty} xP_{s+x}$. Then

$$\lambda E(W_1) = \sum_{x=0}^{k-1} xP_{s+x} + \sum_{x=0}^{\infty}(x+k)P_{s+k+x} = P_s\sum_{x=0}^{k-1} x\frac{a^x}{s^x} + P_{s+k}\left(\frac{k\left(1-\frac{a(1-q)}{s}\right)+\frac{a(1-q)}{s}}{\left(1-\frac{a(1-q)}{s}\right)^2}\right).$$

Expected waiting time for the second queue. For $c = s$, the system is work conserving. Since the overall system is an M/M/s queue, we have a closed form expression for the expected waiting time of the overall customers; $E(W)$. We evaluate the expected waiting time for the second queue with the relation;

$$\pi E(W_2) + (1-\pi)E(W_1) = E(W).$$

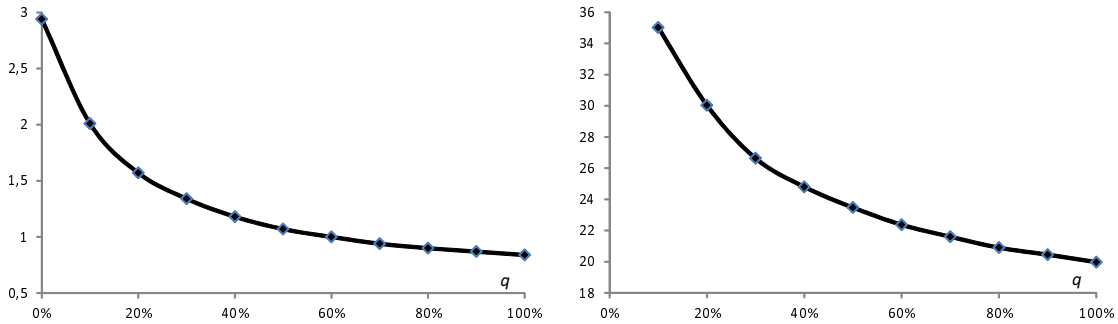
When $c < s$ the system is not work-conserving. We can evaluate $E(W_2)$ with the steady states

Table 5.1: $E(W_1)$, $E(W_2)$, $E(W)$ and π as a function of c with $\lambda = 1$, $s = 6$, $k = 3$ and $q = 80\%$

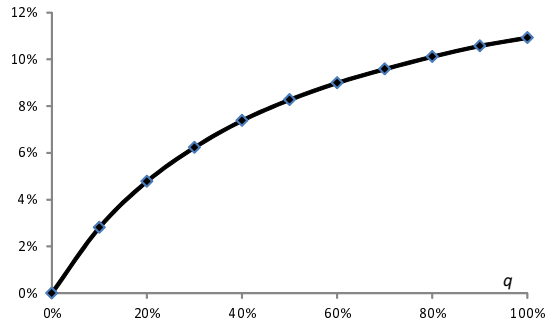
c	1	2	3	4	5	6
$E(W_1)$	0.65	0.67	0.7	0.75	0.81	0.9
$E(W_2)$	∞	∞	1698	68.33	32.85	20.85
$E(W)$	∞	∞	142.83	6.63	3.75	2.94
π	7.51%	7.76%	8.18%	8.62%	9.20%	10.11%

probabilities and the relation $\pi\lambda E(W_2) = \sum_{y=0}^{\infty}\sum_{x=0}^{\infty} y p_{x,y}$. Since the sums of this expression are infinite we can find a significant numerical value by space state truncation.

Figures 5.2, 5.3, 5.4 and Table 5.1 illustrate the evolution of the performance measures as a function of k, q and c. As a function of q, we observe that $E(W_1)$ and $E(W_2)$ are decreasing and convex and π is increasing and concave. As a function of k, $E(W_1)$ and $E(W_2)$ are increasing and concave and π is decreasing and convex. As a function of c, $E(W_1)$ is increasing and convex, $E(W_2)$ is decreasing and concave and π is increasing and convex. We observe a negative correlation between π and $E(W_2)$. It means that the smallest the proportion of called backs is, the more they will have to wait. When the proportion of called back is small, the proportion of inbound calls is important. Those inbound calls benefit from reservation and priority. Thus they impact more negatively the performance of the outbound calls. Increasing q is similar to decreasing k, an increasing in the proportion of outbound calls and a decreasing in the expected waiting time of both inbound and outbound calls. We note for $E(W_1)$ and $E(W_2)$ that the convexity is more important in q than the concavity in k. Recall that q only depends on the behavior of the customers. This implies that a little increasing in q from a low value will impact more than from a high value. The parameter k is defined by the manager of the call center but can not be too high since the manager usually wants to avoid abandonment. The only parameter of control is c. We observe that $E(W_1)$ and π are increasing in c. This is not surprising. As c increases the reservation for inbound calls decreases which implies an increasing of $E(W_1)$. Thus the proportion of callbacks also increases and $E(W_2)$ decreases.



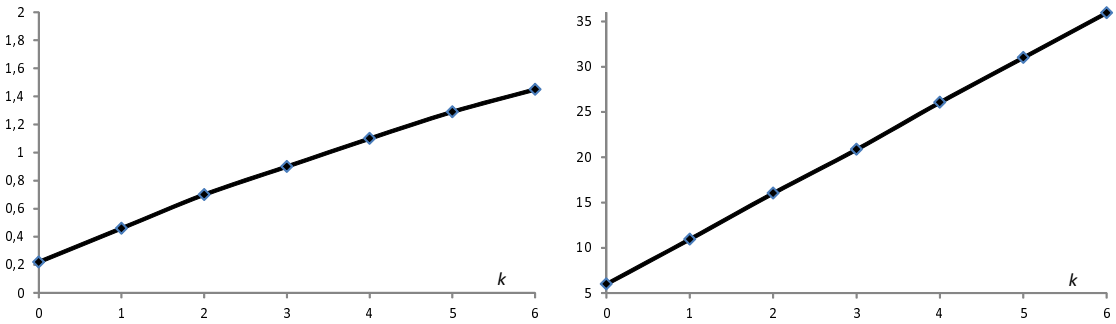
(a) $E(W_1)$ (b) $E(W_2)$



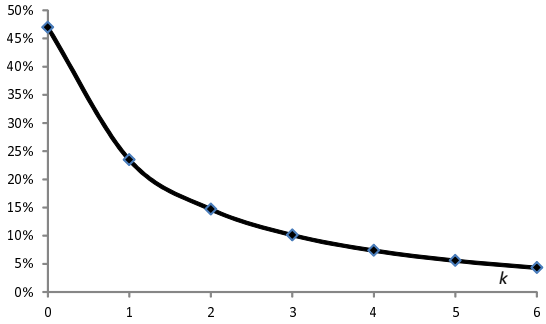
(c) π

Figure 5.2: $E(W_1)$, $E(W_2)$ and π as a function of q with $\lambda = 1$, $s = 6$, $c = 6$ and $k = 3$

Control on the number of outbound tasks. In what follows we limit the number of outbound waiting tasks to K (limited capacity for queue 2), the parameter K could be another parameter of control for the manager. In other words the callback option is not proposed when the number of outbound calls is higher or equal than K. We propose a method to answer the optimization problem by choosing the optimal couple (c,K). The Markov chain and the evaluation of the performance measures are similar to the case with infinite capacity in queue 2. The difference is that the variable y is limited to values between 0 and K. Table 5.2 illustrates the numerical results. The intuitive observation in Table 5.2 is that the increasing of K induces a decreasing of $E(W_1)$, an increasing of $E(W_2)$ and an increasing of π because increasing K allows more space for the callbacks. We find values of K for which $E(W_1)$ and $E(W_2)$ are both decreasing in c (example with $K = 1$). Since the number of outbound calls is very limited, the reservation does not impact much, an increasing of c increases the overall performance of the system and both $E(W_1)$ and $E(W_2)$ benefit from the increasing of c. Thus the best solution when K is small is to have $c = s$.



(a) $E(W_1)$ (b) $E(W_2)$



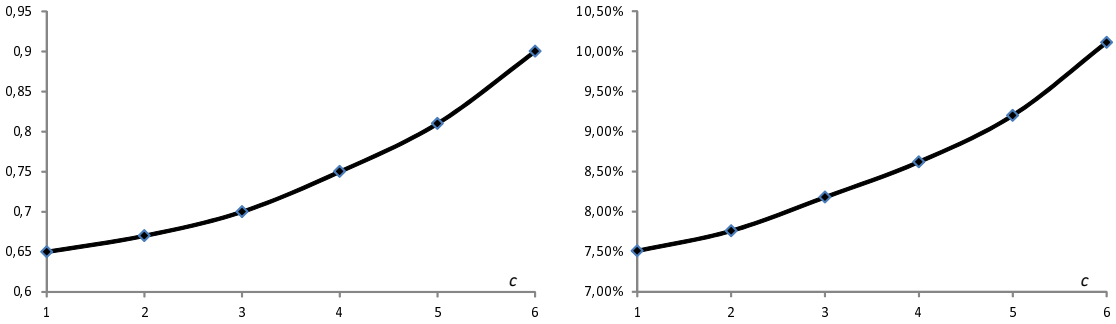
(c) π

Figure 5.3: $E(W_1)$, $E(W_2)$ and π as a function of k with $\lambda = 1$, $s = 6$, $c = 6$ and $q = 80\%$

The limit in this model is the risk of abandon when the arrival rate increases.

5.4 General Case

In this section we consider the general case, using simulation experiments. The problem remains the same, but we do the optimization in finding the optimal curve, $c(y)$. In Figure 5.5 we present the Markov chain associated to the general model. First we consider a simple curve to understand



(a) $E(W_1)$ (b) π

Figure 5.4: $E(W_1)$ and π as a function of c with $\lambda = 1$, $s = 6$, $k = 3$ and $q = 80\%$

Table 5.2: Impact of c and K ($\lambda = 1$, $\mu = 0.2$, $s = 6$, $k = 3$, $q = 80\%$)

	K	0	1	2	3	4	5	10	100	∞
$c = 3$	$E(W_1)$	2.94	2.69	2.49	2.33	2.19	2.06	1.68	0.84	0.7
	$E(W_2)$		51.71	57.30	62.84	68.57	73.98	101.57	541.8	1698
	$E(W)$	2.94	3.20	3.51	3.85	4.20	4.57	6.64	44.28	142.83
	π	0.00%	1.04%	1.85%	2.51%	3.04%	3.49%	4.96%	8.03%	8.18%
$c = 4$	$E(W_1)$	2.94	2.66	2.39	2.19	2.02	1.88	1.36	0.76	0.75
	$E(W_2)$		35.44	37.11	38.90	40.57	42.33	49.13	68.28	68.33
	$E(W)$	2.94	3.08	3.2	3.36	3.52	3.7	4.45	6.61	6.63
	π	0.00%	1.28%	3.32%	3.18%	3.90%	4.50%	6.48%	8.56%	8.62%
$c = 5$	$E(W_1)$	2.94	2.60	2.36	2.11	1.94	1.78	1.24	0.82	0.81
	$E(W_2)$		26.36	27.15	27.48	28	28.62	30.03	32.58	32.85
	$E(W)$	2.94	2.96	3.03	3.06	3.13	3.2	3.39	3.72	3.75
	π	0.00%	1.50%	2.72%	3.74%	4.59%	5.29%	7.46%	9.17%	9.20%

the impact of the number of outbound calls in the choice of the threshold c. Second, we propose an intuitive construction of the optimal curve $c(y)$ in some simple cases. Finally, we present the problem of the construction of the optimal curve $c(y)$ via dynamic programming.

We examine the impact of the number of outbound calls (y) on the choice of the threshold (c) by considering a simple curve $y(c)$. We define a limit in the number of outbound calls denoted by y^*. We reserve $s - c$ agents for inbound calls only if we have at most y^* outbound tasks, otherwise we do not reserve any agent ($c = s$). Note that with $y^* = \infty$ we have the same model as in Section 5.3 and with $y^* = 0$ we have no reservation ($c = s$). Although this method is not optimal, it helps to understand the impact of a decision made on the number of outbound calls. In Table 5.3 we present simulation results for different values of c and y^*. The number of states for which it is possible to have waiting outbound calls increases in y^*. Thus we observe that for a given value of c, $E(W_1)$ decreases and $E(W_2)$ increases in y^*. Since the number of counterproductive states (with $y > 0$ and $x < s$) also increases in y^*, we observe that $E(W)$ (expected average waiting time of both inbound and outbound calls) increases in y^*. When c increases the performance measures are less sensitive to the increasing of y^*. In particular, the extreme case $y^* = 0$ is equivalent to $c = s$. This result is important, it means that for a given service level w_1^* for the inbound calls we should choose the highest possible value for c and adapt y^*. For example if $w_1^* = 0.83$, we find in Table

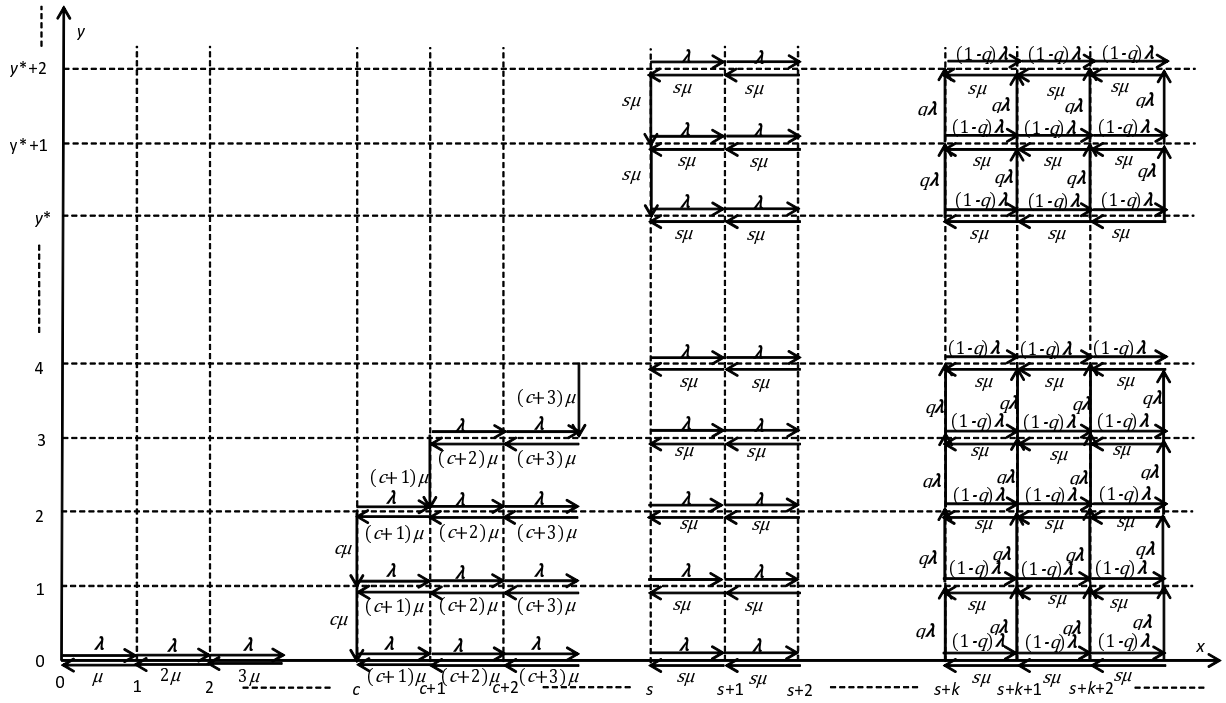


Figure 5.5: Markov chain with a general curve $c(y)$

5.3 three couples (c,y^*) that answer the problem: (3,5), (4,5) and (5,10). The best couple is the last one because it goes with the smallest value for $E(W_2)$. In what follows, we use this insight to propose an intuitive construction of the curve $c(y)$.

We start the construction with $c = s$, then we choose to decrease c to $s-1$ and start with $y^* = 1$ and increase y^* so as to decrease $E(W_1)$. If for any values of y^* (even $y^* = \infty$) we have $E(W_1) > w_1^*$, hence we choose to have $y^* = \infty$ when $c = s-1$ and therefore we decrease c to $s-2$. If we find a value of y^* for which $E(W_1) > w_1^*$ and for y^*+1 we stop there and choose this value of y^* and $c = s-1$. We continue on the same way until we reach $c = 1$.

Example: Curve $c(y)$ for $w_1^* = 0.78$ with $\lambda = 1$, $\mu = 0.2$, $s = 6$, $k = 3$ and $q = 80\%$. We have the following simulated results. For $c = 5$ and $y^* = \infty$ we have $E(W_1) = 0.81$, then we need to decrease c to $c = 4$, then we have for $y^* = 7$, $E(W_1) = 0.7801$ and $E(W_2) = 43.15$ and for $y^* = 8$, $E(W_1) = 0.7777$ and $E(W_2) = 44.57$. Hence the best solution to reach $w_1^* = 0.78$ is to have $y^* = 0$ for $c < 4$, $y^* = 8$ when $c = 4$ and $y^* = \infty$ when $c > 4$. With this method, the best couple would

Table 5.3: Impact of c and y^* ($\lambda = 1$, $\mu = 0.2$, $s = 6$, $k = 3$, $q = 80\%$)

	y^*	0	1	2	3	4	5	10	100	∞
$c = 3$	$E(W_1)$	0.9	0.88	0.87	0.85	0.84	0.83	0.8	0.74	0.7
	$E(W_2)$	20.85	24.38	27.81	32	36.23	40.72	65.01	515.97	1698
	$E(W)$	2.94	3.22	3.49	3.85	4.19	4.6	6.64	44.5	142.83
	π	10.11%	9.93%	9.75%	9.64%	9.48%	9.44%	9.10%	8.49%	8.18%
$c = 4$	$E(W_1)$	0.9	0.88	0.86	0.85	0.84	0.83	0.8	0.76	0.75
	$E(W_2)$	20.85	22.78	24.83	27.09	29.05	31.13	41.06	68.38	68.33
	$E(W)$	2.94	3.04	3.19	3.37	3.51	3.65	4.45	6.64	6.63
	π	10.11%	9.87%	9.71%	9.59%	9.46%	9.33%	9.06%	8.69%	8.62%
$c = 5$	$E(W_1)$	0.9	0.89	0.87	0.86	0.86	0.85	0.83	0.81	0.81
	$E(W_2)$	20.85	21.88	22.77	23.54	24.42	24.99	28.31	32.57	32.85
	$E(W)$	2.94	2.98	3.03	3.07	3.13	3.16	3.40	3.72	3.75
	π	10.11%	9.98%	9.84%	9.73%	9.67%	9.57%	9.35%	9.18%	9.20%

be $c = 4$ and $y^* = 16$ which leads to $E(W_1) = 0.7799$ and $E(W_2) = 50.63$.

Using dynamic programming we define the problem of the search of the optimal curve. We denote by c_1 and c_2 the cost of a waiting customer in queue 1 and queue 2, respectively. We denote by $V_n(x, y)$ the expected costs over n steps. We have

$$\begin{aligned} V'_{n+1}(x,y) &= c_1(x-s)^+ + c_2 y \\ &+ \frac{\lambda}{\lambda + s\mu}\left(1_{x<k+s} V_n(x+1,y) + 1_{x\geq k+s}(qV_n(x,y+1) + (1-q)V_n(x+1,y))\right) \\ &+ \frac{\mu}{\lambda + s\mu}\left(\min(x,s)V_n(x-1,y) + (s - \min(x,s))V_n(x,y)\right), \end{aligned} \tag{5.3}$$

and

$$V_{n+1}(x,y) = \begin{cases} V'_{n+1}(x,y) \text{ if } x \geq s \\ \min(V'_{n+1}(x,y), V'_{n+1}(x+1,y-1)) \text{ if } x < s \end{cases} \tag{5.4}$$

The long-term average optimal actions are a solution of the optimal equation (in vector notation) $TV = g + V$. Another way of obtaining them is through value iteration, by recursively defining $V_{n+1} = TV_n$, for arbitrary V_0. In the expression we see that it is optimal to schedule a type 2 task only if $V(x+1, y-1) < V(x, y)$. With this relation we can numerically build the optimal curve

$y(c)$.

In summary, we numerically find an efficient curve $y(c)$ that answers Problem (5.1). This curve is defined by two parameters: c^* ($0 < c^* \leq s$) and y^* ($y^* \geq 0$). For $c < c^*$, we do not do any reservation and $y(c) = 0$. For $c = c^*$, $y(c^*) = y^*$. For $c > c^*$, $y(c) = +\infty$. The main reason is that the sensitivity of the performance measures is more important in c than on the number of outbound calls. We still have to compare this intuitive result with the optimal curve that can be found through a dynamic programming approach.

5.5 Conclusions and Future Research

In this chapter we considered simple call center modeling with a callback option. We assumed a threshold policy on the reservation of agents for inbound calls and derived the main performance measures. The optimization problem is defined as minimizing the expected waiting time of the outbound calls while respecting a service level constraint on the inbound ones. We answered this problem by building a curve representing the relation between the threshold and the number of outbound tasks.

For future research, it might be interesting to evaluate via closed form expressions, the performance measures on the optimal curve. Moreover, the complexity of the customer behavior has been partly avoided. It is also interesting to include the feature of customer abandonment for the outbound calls. Another extension is to consider different service time distribution for inbound and outbound calls.

Chapter 6

Conclusions 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.

6.1 Conclusions

This thesis focused on operations management issues for SBR and multi-channel call centers. We investigated an important problem in the design and management of SBR call centers. The purpose of this study was to propose an intelligent architecture. We considered the context of call centers with unbalanced parameters. Under most cases of asymmetry, the well known existing architectures such as chaining lose their robustness. Thus we proposed a new call center architecture (single pooling) and demonstrated its efficiency. We showed that SP behaves well in most cases of asymmetry in the parameters. There are opportunities for managers of call centers to improve performance using this new architecture.

Next, we considered a blended call center with calls and emails. The call service is characterized by successive stages where one of them is a break for the agent. We focused on the optimization of the email routing given that calls have a non-preemptive priority over emails. Our objective was to maximize the throughput of emails subject to a constraint on the call waiting time. We

developed a general framework with two probabilistic parameters for the email routing to agents. One parameter controls the routing between calls, and the other does the control inside a call conversation. We have also considered 4 particular cases corresponding to the extreme values of the probabilistic parameters. For these routing models, we have derived various structural results and discussed the theoretical results in order to provide guidelines to call center managers. In particular, we proved for the optimal routing that all the time at least one of the two email routing parameters has an extreme value.

Next, we considered call centers with inbound calls and an infinite supply of non-preemptable outbound jobs. We proposed a scheduling policy, refereed to as ATP, where the objective is to do as much outbound as possible while satisfying a service level constraint on the call waiting time. In the real-life call center context with fluctuating call arrival rate, the assignment policy for outbound adapts itself to the current service level. We showed the efficiency of ATP by comparing its performance with those of other policies. One of the main advantage of ATP is its ability to quickly react when an important change in the arrival process happens and its ability to avoid inefficient states when the arrival rate remains constant.

Finally, we considered a call center modeling with a callback option. We assumed a threshold policy on the reservation of agents for inbound calls and derived the main performance measures. The optimization problem is defined as minimizing the expected waiting time of the outbound calls while respecting a service level constraint on the inbound ones. We answered this problem by building the optimal curve representing the relation between the threshold and the number of outbound tasks.

6.2 Future Research

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.

For the architectures of chaining as well as for single pooling, there is a need to provide and analyse a much general model. Another interesting work is to find an architecture which resists better than chaining and single pooling to blocking, especially in small call centers.

In the modeling of Chapter 3, it would be interesting to extend the structural results to the multi-server case. It would also be useful but challenging to extend the analysis to cases with an additional channel, in particular the chat which is increasingly used in call centers. Using the chat channel, an agent may handle many customers at the same time, which represent an additional opportunity to efficiently use the agent time.

Future research on the subject of adapting blending in Chapter 4 may follow two directions. First, further analytical analysis of the adaptive blending might be useful to better understand ATP and maybe to propose a better adaptive policy. Second, the complexity of call centers has been partly avoided in our study. The customers abandonment, the callback option, the different types of the inbound calls, the switching times between different tasks, the tiredness of the agent, a finite number of back office tasks could interfere in our results and might introduce other interesting performance measures in the optimization problem.

For future research on the callback option in Chapter 5, it may be useful to derive the optimal performance measures. Moreover, again, the complexity of the customer behavior has been partly avoided. The customers abandonment could occur even when the callback option is proposed. The service of the outbound calls could be done with a part of already informed customers which induces very short service times duration.

Appendix A

Appendix of Chapter 2

A.1 Optimization Heuristic

A coherent comparison between chaining and single pooling requires first the optimization of their total cost. In this section, we develop a greedy heuristic for the optimization step of the two models, and prove their efficiency. The heuristic is a simulation based optimization method. Recall that for each model, we optimize the total staffing cost under the constraints $W_i \leq W_i^*$, for $i = 0, 1, ..., n$. In the numerical examples below, we consider Markovian assumptions for inter-arrival and service times. The analysis can be applied in a similar way to any other assumption.

The question addressed here is how can we compute the optimal number of agents in each team? In some particular cases the answer is simple. For example when the arrival and service rates are identical for all skills and when all skills have the same costs, we would create teams with the same number of agents. A more difficult situation is in the case of asymmetric arrival or service rates. For the simulation based optimization considered here, some information about the simulation process are as follows. We use C++ program codes. For a given simulation with a given set of parameters, we consider a single replication that we run for a sufficiently long time. The lengths of the confidence intervals for the different performance measures derived by simulation are in the order of 10^{-4}. To obtain such confidence intervals, we simply gradually increase the replication

length up to the point that ensures the accuracy objective. This implies that the simulation length may vary from one set of parameters to another. The total number of generated calls varies and is in the order of tens of millions. The delay to run a simulation also varies and is in the order of several minutes.

A.1.1 Single Pooling

In what follows we present three staffing heuristics, and then select the best one. The heuristics consist on adaptations of greedy and local search algorithms.

Algorithm 1: Decreasing Greedy in Team 0

Without customers 0, single pooling is simply an FD model. A first idea of staffing is then to use a decreasing greedy algorithm as follows. We start such that we have a collection of $n+1$ independent M/M/s queues. In each team i, the number of agents is the minimum required one to reach $W_i \leq W_i^* = 0.2$, for $i = 0, 1, ..., n$. In each iteration, we decrement the number of agents in team 0 by one, and evaluate all W_i, for $i = 0, 1, ..., n$. We stop the algorithm when all the service levels are no longer reached for the first time. We then consider the results of the before last iteration. Table A.1 presents the results of the decreasing greedy algorithm in a single pooling model with 3 customer types, and compare it with the those of FF and FD models. The staffing level in team i is denoted by s_i, $i = 0, 1, 2$.

Table A.1: Decreasing greedy in team 0 for single pooling ($n = 2$, $\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$, $i = 1, 2$)

λ_1	λ_2	λ_0	FF	FD	Single pooling s_1	s_2	s_0	Total $s_0 + s_1 + s_2$
1	0.5	0.2	13	19	9	6	0	15
0.2	0.5	1	13	19	4	6	5	15
1	1	1	20	27	9	9	5	23
3	2	1	36	44	20	15	5	40
2	1	3	36	44	15	9	18	42
0.5	0.2	0.1	8	13	6	4	0	10
0.1	0.2	0.5	8	13	3	4	2	9
10	5	15	151	171	57	31	83	171
10	15	5	151	171	57	83	30	170
10	10	10	151	171	57	57	57	171

Algorithm 2: Increasing Greedy

Another idea is to proceed by introducing customers 0 in the system step by step. We start from an FD model with no customers 0, and we define the staffing level in team i such that $W_i \leq 0.2$, $i = 1, ..., n$ (team 0 is being empty). In each iteration, we increase λ_0 by a given small step value (we have chosen in the experiments a sufficiently small step of $\lambda_0/100$). If $W_0 \geq W_0^*$, we add one agent in team 0. If $W_0 \leq W_0^*$ and $W_i > W_i^*$ for some customer types, we then add an agent in the team with the highest W_i. We stop the algorithm once λ_0 reaches its value and the constraints $W_i \leq W_i^*$ are all satisfied, for $i = 0, 1, ..., n$. Table A.2 provides the simulation results of this algorithm.

Table A.2: Increasing greedy for single pooling ($n = 2$, $\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$, $i = 1, 2$)

λ_1	λ_2	λ_0	FF	FD	Single pooling s_1	s_2	s_0	Total $s_0 + s_1 + s_2$
1	0.5	0.2	13	19	9	6	0	15
0.2	0.5	1	13	19	6	7	0	15
1	1	1	20	27	10	10	1	21
3	2	1	36	44	21	16	1	38
2	1	3	36	44	17	12	8	37
0.5	0.2	0.1	8	13	6	4	0	10
0.1	0.2	0.5	8	13	4	5	0	9
10	5	15	151	171	61	36	64	161
10	15	5	151	171	60	85	18	163
10	10	10	151	171	60	61	41	162

Algorithm 3: Increasing Greedy with No Agents in Team 0

The algorithm is identical to the previous one, expect that we force team 0 to be empty. Table A.3 presents the simulated results for this algorithm.

From the results of all algorithms, we observe that the decreasing greedy algorithm (algorithm 1) is the worst. The reason is that it is not possible to increase or decrease the number of agents in a regular team i ($i = 1, ..., n$). Many effective configurations could not then be reached under this algorithm. The other two algorithms are equivalent in terms of the total number of agents in our simulation experiments. We have chosen to use algorithm 2 in the experiments of Section 5 of the main paper.

Table A.3: Increasing greedy with no Agents in team 0 for single pooling ($n = 2$, $\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$, $i = 1, 2$)

λ_1	λ_2	λ_0	FF	FD	Single pooling s_1	s_2	s_0	Total $s_0 + s_1 + s_2$
1	0.5	0.2	13	19	9	6	0	15
0.2	0.5	1	13	19	6	7	0	15
1	1	1	20	27	11	10	0	21
3	2	1	36	44	22	16	0	38
2	1	3	36	44	21	16	0	37
0.5	0.2	0.1	8	13	6	4	0	10
0.1	0.2	0.5	8	13	4	5	0	9
10	5	15	151	171	93	68	0	161
10	15	5	151	171	70	93	0	163
10	10	10	151	171	81	81	0	162

We go further in order to check the quality of algorithm 2. In Table A.4, we provide optimization results using algorithm 2 and also using other configurations with one agent in less. We observe that these other configurations do not allow to satisfy all the constraints $W_i \leq W_i^* = 0.2$, for $i = 0, 1, ..., n$, which proves the efficiency of algorithm 2.

Table A.4: Efficiency of algorithm 2 ($n = 2$, $\mu_i = \mu_0 = 0.2$, $W_i^* = W_0^* = 0.2$, $i = 1, 2$)

	s_1	s_2	s_0	W_1	W_2	W_0	Constraints
$\lambda_1 = 1$,	9	6	0	0.123	0.126	0.019	Satisfied (algorithm 2)
$\lambda_2 = 0.5$,	8	6	0	0.143	0.407	0.057	Not satisfied (one agent in less)
$\lambda_0 = 0.2$	7	7	0	0.898	0.049	0.036	Not satisfied (one agent in less)
	9	5	0	0.338	0.141	0.051	Not satisfied (one agent in less)
$\lambda_1 = 3$,	21	16	1	0.174	0.154	0.110	Satisfied (algorithm 2)
$\lambda_2 = 2$,	21	16	0	0.204	0.191	0.205	Not satisfied (one agent in less)
$\lambda_0 = 1$	22	15	0	0.145	0.281	0.209	Not satisfied (one agent in less)
	20	16	1	0.290	0.177	0.183	Not satisfied (one agent in less)
	21	15	1	0.195	0.269	0.195	Not satisfied (one agent in less)
$\lambda_1 = 10$,	60	85	18	0.152	0.171	0.167	Satisfied (algorithm 2)
$\lambda_2 = 15$,	60	85	17	0.166	0.207	0.177	Not satisfied (one agent in less)
$\lambda_0 = 5$	60	84	18	0.171	0.218	0.212	Not satisfied (one agent in less)
	59	85	18	0.205	0.172	0.180	Not satisfied (one agent in less)
	59	86	17	0.210	0.168	0.178	Not satisfied (one agent in less)

A.1.2 Chaining

We also use a greedy algorithm to optimize the staffing of chaining. The simulation results reveal that increasing and decreasing greedy algorithms are efficient and very similar if we start the optimization heuristic with a good initialization of the team sizes. We choose to use a decreasing greedy algorithm since it is faster than an increasing greedy one (no need to increase the λ_0 with

a high number of small steps).

The method is as follows. In each team i ($i = 0, 1, ..., n$), we start with the worst (over-estimated) staffing level s_i computed from an FD model. In order to take into account the chaining configuration, i.e., the fact that customers type $i-1$ can be routed to team i and customers type i can be routed to team $i+1$, we adjust the initial staffing levels from s_i to s'_i for team i ($i = 0, 1, ..., n$). We use the method suggested by Wallace and Whitt (2005). The corrected staffing level s'_i is given by

$$s'_i = s_i - R_{i,i+1} + R_{i-1,i}, \tag{A.1}$$

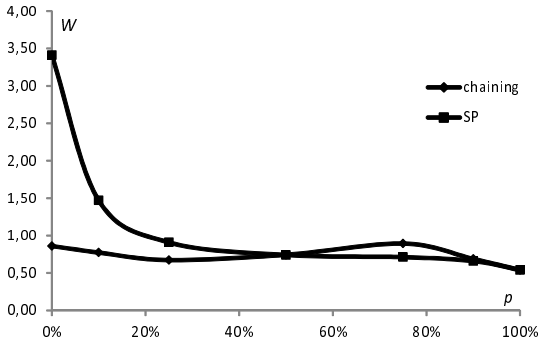
for $i = 0, 1, ..., n$, where $R_{i,j} = \frac{s_i s_j}{s - s_i}$. This number is that of agents of team i who could go to team j, $i, j = 0, 1, ..., n$ and $i \neq j$. Using Equation (A.1), s'_i may not be an integer. We then round it to the nearest integer above.

A.2 Metric Comparison

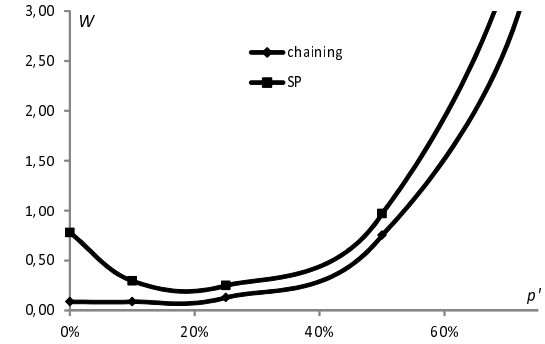
In Table A.5 and Figures A.1(a)-A.1(d), we compare between the expected waiting times of SP and chaining for a given total staffing level. We optimize the staffing of the various teams in the two models for the case $t = 0$, i.e., no incremental cost for regular skills. The results show the same qualitative conclusions as those in Chapter 2.

Table A.5: Performance measures of SP and chaining

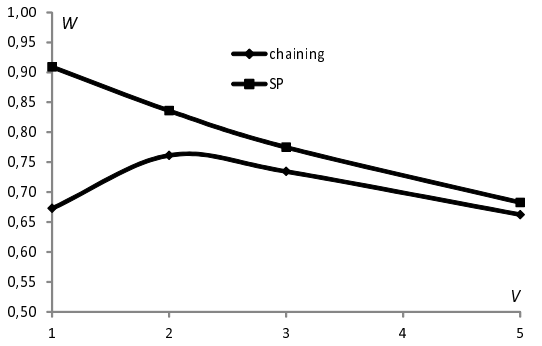
Impact of p			Impact of p'			Impact of V			Impact of U		
p	SP	Chaining	p'	SP	Chaining	V	SP	Chaining	U	SP	Chaining
0%	3.41	0.86	0%	0.78	0.09	1	0.91	0.67	1	0.33	0.17
10%	1.47	0.77	10%	0.30	0.09	2	0.84	0.76	2	0.37	0.19
25%	0.91	0.67	25%	0.25	0.13	3	0.77	0.73	3	0.39	0.34
50%	0.74	0.74	50%	0.97	0.76	5	0.68	0.66	5	0.39	0.38
75%	0.71	0.89	75%	5.11	4.10						
90%	0.66	0.69	90%	139	102						
100%	0.54	0.54									



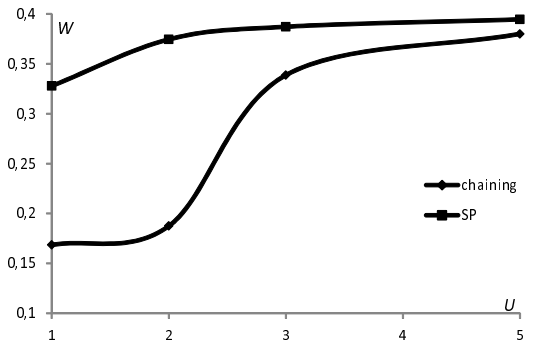
(a) Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} s_i = 44$, $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$)



(b) Impact of p' ($\lambda_0 = 2$, $\lambda_i = 1.5$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} s_i = 48$, $i = 1, ..., 4$, $p = 25\%$, $U = V = 1$)



(c) Impact of V ($\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} s_i = 44$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$)



(d) Impact of U ($\lambda_0 = 2$, $\lambda_i = 1.5$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} s_i = 48$, $\mu_0 = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $V = 1$)

Figure A.1: Performance measures of SP and chaining

A.3 Impact of Abandonment

The experiments of Tables A.6-A.9 are associated to Figures 2.10(a)-2.10(d) of Chapter 2, respectively. The experiments of Tables A.10-A.13 are associated to Figures 2.11(a)-2.11(d) of Chapter 2, respectively.

Table A.6: Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $\gamma_i = \gamma_0 = \gamma$, $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$\gamma = 0$	0%	49	50.95	52.9	58.75	68.5	60	t=28.21%
	10%	49	50.7	52.4	57.5	66	56	t=20.58%
	25%	48	49.3	50.6	54.5	61	52	t=15.38%
	50%	49	49.9	50.8	53.5	58	52	t=16.67%
	75%	51	51.55	52.1	53.75	56.5	51	t=0%
	90%	51	51.3	51.6	52.5	54	51	t=0%
$\gamma = 0.1$	0%	47	48.55	50.1	54.75	62.5	56	t=29.03%
	10%	46	47.5	49	53.5	61	52	t=20.00%
	25%	46	47.3	48.6	52.5	59	52	t=23.08%
	50%	46	46.9	47.8	50.5	55	51	t=27.78%
	75%	48	48.5	49	50.5	53	50	t=20.00%
	90%	49	49.35	49.7	50.75	52.5	49	t=0.00%
$\gamma = 0.2$	0%	44	45.5	47	51.5	59	52	t=26.67%
	10%	42	43.35	44.7	48.75	55.5	48	t=22.22%
	25%	44	45.25	46.5	50.25	56.5	48	t=16.00%
	50%	44	44.8	45.6	48	52	48	t=25.00%
	75%	45	45.4	45.8	47	49	48	t=37.50%
	90%	45	45.2	45.4	46	47	48	t=75.00%

Table A.7: Impact of p' ($\lambda_i = \lambda_0 = 2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$)

		Chaining					SP	Crossing value
	p'	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$\gamma = 0$	0%	60	62.45	64.9	72.25	84.5	72	t=24.49%
	10%	59	60.95	62.9	68.75	78.5	67	t=20.51%
	25%	58	59.65	61.3	66.25	74.5	62	t= 12.12%
	50%	60	61.05	62.1	65.25	70.5	65	t=23.81%
	75%	61	61.6	62.2	64	67	68	t=58.33%
	90%	65	65.25	65.5	66.25	67.5	69	t=80.00%
$\gamma = 0.1$	0%	57	59.1	61.2	67.5	78	67	t=23.81%
	10%	57	59.05	61.1	67.25	77.5	65	t=19.51%
	25%	57	58.75	60.5	65.75	74.5	61	t=11.43%
	50%	59	60.2	61.4	65	71	64	t=20.83%
	75%	60	60.75	61.5	63.75	67.5	64	t=26.67%
	90%	61	61.25	61.5	62.25	63.5	62	t=20.00%
$\gamma = 0.2$	0%	55	57.05	59.1	65.25	75.5	60	t=12.20%
	10%	55	56.95	58.9	64.75	74.5	60	t=12.82%
	25%	55	56.75	58.5	63.75	72.5	58	t=8.57%
	50%	55	56.1	57.2	60.5	66	59	t=18.18%
	75%	57	57.7	58.4	60.5	64	60	t=21.43%
	90%	59	59.25	59.5	60.25	61.5	59	t=0.00%

Table A.8: Impact of V ($\lambda_0 = 2$, $\mu_0 = \mu_i = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$)

	V	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
$\gamma = 0$	1	48	49.3	50.6	54.5	61	52	t=15.38%
	2	49	50.3	51.6	55.5	62	53	t=15.38%
	3	49	50.25	51.5	55.25	61.5	52	t=12.00%
	5	50	51.25	52.5	56.25	62.5	52	t=8.00%
$\gamma = 0.1$	1	46	47.3	48.6	52.5	59	52	t=23.08%
	2	46	47.25	48.5	52.25	58.5	51	t=20.00%
	3	46	47.15	48.3	51.75	57.5	51	t=21.74%
	5	46	47.1	48.2	51.5	57	51	t=22.73%
$\gamma = 0.2$	1	44	45.25	46.5	50.25	56.5	48	t=16.00%
	2	45	46.25	47.5	51.25	57.5	51	t=24.00%
	3	45	46.15	47.3	50.75	56.5	51	t=26.09%
	5	46	47.05	48.1	51.25	56.5	52	t=28.57%

Table A.9: Impact of U ($\mu_0 = 0.2$, $\lambda_0 = 4$, $\lambda_i = 1$, $W_0 = W_i^* = 0.2$, $i = 1, ..., 4$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $p' = 20\%$, $p = 50\%$, $V = 1$)

	U	Chaining t=0%	t=5%	t=10%	t=25%	t=50%	SP	Crossing value (Chaining = SP)
$\gamma = 0$	1	49	50.25	51.5	55.25	61.5	52	t=12.00%
	2	49	49.75	50.5	52.75	56.5	53	t=26.67%
	3	50	51.65	52.3	54.25	57.5	52	t=7.69%
	5	52	52.65	53.3	55.25	58.5	52	t=0.00%
$\gamma = 0.1$	1	46	46.9	47.8	50.5	55	51	t=27.78%
	2	48	48.9	49.8	52.5	57	51	t=23.53%
	3	50	50.85	51.7	54.25	58.5	51	t=5.88%
	5	51	51.75	52.5	54.75	58.5	51	t=0.00%
$\gamma = 0.2$	1	44	44.8	45.6	48	52	48	t=25.00%
	2	47	47.85	48.7	51.25	55.5	48	t=5.88%
	3	49	49.8	50.6	53	57	49	t=0.00%
	5	49	49.65	50.3	52.25	55.5	49	t=0.00%

Table A.10: Impact of p ($\mu_i = \mu_0 = 0.2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \lambda_i = 8$, $i = 1, ..., 4$, $p' = 20\%$, $U = V = 1$)

		Chaining					SP	Crossing value
	p	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$\gamma_0 = 0.1$ $\gamma_i = 0$	0%	49	50.95	52.9	58.75	68.5	60	t=28.21%
	10%	48	49.6	51.2	56	64	58	t=31.25%
	25%	48	49.25	50.5	54.25	60.5	56	t=32.00%
	50%	48	48.9	49.8	52.5	57	54	t=33.33%
	75%	48	48.6	49.2	51	54	52	t=33.33%
	90%	49	49.3	49.6	50.5	52	51	t=33.33%
$\gamma_i = \gamma_0 = 0.1$	0%	47	48.55	50.1	54.75	62.5	56	t=29.03%
	10%	46	47.5	49	53.5	61	52	t=20.00%
	25%	46	47.3	48.6	52.5	59	52	t=23.08%
	50%	46	46.9	47.8	50.5	55	51	t=27.78%
	75%	48	48.5	49	50.5	53	50	t=20.00%
	90%	49	49.35	49.7	50.75	52.5	49	t=0.00%
$\gamma_0 = 0$ $\gamma_i = 0.1$	0%	47	48.55	50.1	54.75	62.5	56	t=29.03%
	10%	48	49.3	50.6	54.5	61	54	t=23.08%
	25%	48	49.25	50.5	54.25	60.5	52	t=16.00%
	50%	48	48.85	49.7	52.25	56.5	52	t=23.53%
	75%	49	49.75	50.5	52.75	56.5	51	t=13.33%
	90%	49	49,3	49,6	50,5	52	49	t=0,00%

Table A.11: Impact of p' ($\lambda_0 = \lambda_i = 2$, $W_0^* = W_i^* = 0.2$, $\sum_{i=0}^{4} \frac{1}{\mu_i} = 25$, $i = 1, ..., 4$, $p = 20\%$, $U = V = 1$)

		Chaining					SP	Crossing value
	p'	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$\gamma_0 = 0.1$ $\gamma_i = 0$	0%	60	62.25	64.5	71.25	82.5	72	t=26.67%
	10%	58	60	62	68	78	66	t=20.00%
	25%	58	59.5	61	65.5	73	65	t=23.33%
	50%	58	59.2	60.4	64	70	65	t=29.17%
	75%	58	58.6	59.2	61	64	66	t=66.67%
	90%	59	59.25	59.5	60.25	61.5	68	t=180.00%
$\gamma_0 = \gamma_i = 0.1$	0%	57	59.1	61.2	67.5	78	67	t=23.81%
	10%	57	59.05	61.1	67.25	77.5	65	t=19.51%
	25%	57	58.75	60.5	65.75	74.5	61	t=11.43%
	50%	59	60.2	61.4	65	71	64	t=20.83%
	75%	60	60.75	61.5	63.75	67.5	64	t=26.67%
	90%	61	61.25	61.5	62.25	63.5	62	t=20.00%
$\gamma_0 = 0$ $\gamma_i = 0.1$	0%	57	59.15	61.3	67.75	78.5	66	t=20.93%
	10%	56	57.85	59.7	65.25	74.5	61	t=13.51%
	25%	57	58.35	59.7	63.75	70.5	61	t=14.81%
	50%	58	59.05	60.1	63.25	68.5	63	t=23.81%
	75%	60	60.75	61.5	63.75	67.5	65	t=33.33%
	90%	63	63.25	63.5	64.25	65.5	65	t=40.00%

Table A.12: Impact of V ($\lambda_0 = 2$, $W_0^* = W_i^* = 0.2$, $\mu_i = \mu_0 = 0.2$, $i = 1, ..., 4$, $p = 25\%$, $p' = 20\%$, $U = 1$)

		Chaining					SP	Crossing value
	V	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$\gamma_0 = 0.1$ $\gamma_i = 0$	1	48	49.4	50.8	55	62	53	t=17.86%
	2	49	50.3	51.6	55.5	62	53	t=15.38%
	3	50	51.15	52.3	55.75	61.5	53	t=13.04%
	5	52	53.05	54.1	57.25	62.5	54	t=9.52%
$\gamma_0 = \gamma_i = 0.1$	1	46	47.3	48.6	52.5	59	52	t=23.08%
	2	46	47.25	48.5	52.25	58.5	51	t=20.00%
	3	46	47.15	48.3	51.75	57.5	51	t=21.74%
	5	46	47.1	48.2	51.5	57	51	t=22.73%
$\gamma_0 = 0$ $\gamma_i = 0.1$	1	47	48.35	49.7	53.75	60.5	52	t=18.52%
	2	47	48.15	49.3	52.75	58.5	51	t=17.39%
	3	47	48	49	52	57	51	t=20.00%
	5	47	47.85	48.7	51.25	55.5	52	t=29.41%

Table A.13: Impact of U ($\lambda_0 = 4$, $W_0^* = W_i^* = 0.2$, $\mu_0 = 0.2$, $i = 1, ..., 4$, $p = 50\%$, $p' = 20\%$, $V = 1$)

		Chaining					SP	Crossing value
	U	t=0%	t=5%	t=10%	t=25%	t=50%		(Chaining = SP)
$\gamma_0 = 0.1$ $\gamma_i = 0$	1	47	47.95	48.9	51.75	56.5	52	t=26.32%
	2	48	48.8	49.6	52	56	53	t=31.25%
	3	48	48.85	49.7	52.25	56.5	52	t=23.53%
	5	48	48.9	49.8	52.5	57	52	t=22.22%
$\gamma_i = \gamma_0 = 0.1$	1	46	46.9	47.8	50.5	55	51	t=27.78%
	2	48	48.9	49.8	52.5	57	51	t=23.53%
	3	50	50.85	51.7	54.25	58.5	51	t=5.88%
	5	51	51.75	52.5	54.75	58.5	51	t=0.00%
$\gamma_0 = 0$ $\gamma_i = 0.1$	1	48	48.85	49.7	52.25	56.5	52	t=23.53%
	2	48	48.7	49.4	51.5	55	49	t=7.14%
	3	49	49.9	50.8	53.5	58	50	t=5.56%
	5	50	51	52	55	60	51	t=5.00%

Appendix B

Appendix of Chapter 3

B.1 Simultaneous Treatment of Back Office Tasks

In this section we focus on the possibility to work simultaneously on different tasks (usually back office tasks). This way of working is not common in call centers as we know that a human resource is limited in the simultaneous treatment. However, we wonder if the possibility of simultaneity goes with an increasing of the performance measures. We consider a model with no increasing in productivity when an agent treats tasks simultaneously. A simple classical modeling of this situation is to suppose that an agent can treat at most k tasks simultaneously. The tasks arrive one by one according to a Poisson arrival process of rate λ. We also suppose that the service time distribution follows an exponential distribution of rate μ/n for $1 \leq n \leq k$ when n tasks are done simultaneously by an agent. When an agent treats more than one task at the same time we suppose that the service rate changes automatically in function of the number of tasks in service when one task finishes or starts the service. We denote by s the number of agents in the call center. When a task arrives, if all agents are busy with k tasks, the new task waits in the queue. If not this new task is routed to the less busy agent. If two or more agents are the least busy then the new task is routed to one of these agents with the same probability. When an agent finishes a task she starts a new one if there is one in the queue, if not she finishes her remaining tasks in service if any.

We consider the performance measures of the average waiting time (denoted by W_q), the average service time (denoted by W_s) and the overall time spent in the call center ($W_q + W_s$) for a task.

In the case $s = 1$ we can find formulas for the performance measures. The Markov chain in this case is the one of a M/M/1 queue. Only the performance measures are evaluated differently. If we denote by p_x ($x \in \mathbb{N}$) the probability to have x tasks in the call center, we have $p_x = a^x(1-a)$ with $a = \lambda/\mu$. Thus with Little law and a capacity of simultaneity of k, we have

$$W_q = \frac{1}{\lambda}\sum_{x=0}^{+\infty} x p_{x+k} = \frac{1}{\lambda}(1-a)a^{k+1}\frac{1}{(1-a)^2} = \frac{1}{\lambda}\frac{a^{k+1}}{1-a},$$

and

$$W_s = \frac{1}{\lambda}\left(\sum_{x=0}^{k-1} x p_x + k\sum_{x=0}^{k-1} p_{x+k}\right) = \frac{1}{\lambda}\frac{a(1-a^k)}{1-a}.$$

We note that $W_q + W_s = \frac{1}{\lambda}\frac{a}{1-a}$ is the overall spent time in a classical M/M/1 queue. Thus is we increase the level of simultaneity k, we decrease the expected waiting time but we increase the service time. There is no benefit in the overall spent time in the call center.

In Table B.1 we present numerical results for the evolution of the performance measures in function of the level of simultaneity. We use simulation for $s > 1$. When $s > 1$ we still observe a decreasing of the average waiting time and an increasing of the service time in function of k. Although the overall spent time increases with k, then the call center is less efficient when the level of simultaneity increases. The reason is the possibility of inefficient states in the system when $s > 1$; for example an agent with two tasks and one idle agent at the same time.

These observations reduce the possible interest for a simultaneous treatment. Note that the conclusions are based on the assumption that there is no increasing in productivity when an agent treats tasks simultaneously. That is why we also consider another simple modeling similar as the previous one. The only difference is when an agent treats n tasks simultaneously, the service rate is $\frac{\mu}{(1+t)^{n-1}}$ for $1 \leq n \leq k$ and $t \geq 0$. This model is more compatible with human behavior. In

Table B.1: Performance Measures in function of the level of simultaneity ($\mu = 0.2$)

	k	1	2	3	5	10	$+\infty$
($\lambda = 0.1$, $s = 1$)	W_q	5	2.5	1.25	0.31	0.01	0
	W_s	5	7.5	8.75	9.69	9.99	10
	$W_q + W_s$	10	10	10	10	10	10
($\lambda = 0.19$, $s = 1$)	W_q	95	90.25	85.74	77.38	59.87	0
	W_s	5	9.75	14.26	22.62	40.13	100
	$W_q + W_s$	100	100	100	100	100	100
($\lambda = 0.5$, $s = 6$)	W_q	0.068	0.000	0.000	0.000	0.000	0.000
	W_s	5.000	5.311	5.315	5.317	5.318	5.318
	$W_q + W_s$	5.068	5.311	5.315	5.317	5.318	5.318
($\lambda = 1$, $s = 6$)	W_q	2.94	1.95	1.08	0.21	0.02	0.00
	W_s	5.00	9.39	12.81	16.53	17.03	17.34
	$W_q + W_s$	7.94	11.34	13.89	16.74	17.05	17.34
($\lambda = 1.15$, $s = 6$)	W_q	17.75	16.38	15.03	11.98	6.87	0.00
	W_s	5.00	9.90	15.09	24.51	43.66	64.71
	$W_q + W_s$	22.75	26.28	30.12	36.49	50.53	64.71

this model the productivity is better for a simultaneous treatment than for a successive one when the number of simultaneous tasks is smaller than a limit (determined with t). After this limit the productivity is worse. Roughly speaking the agent is overwhelmed. In Table B.2 we present the simulated results for different values of t. We observe an opportunity in working simultaneously on different tasks when t is small. For $t = 0.1$, the average waiting time decreases until $k = 10$ and the overall spent time also decreases until $k = 2$. For $t = 0.5$ or $t = 1$ there is no benefit in simultaneity for the overall spent time and neither for the average waiting time when $t = 1$. This implies that simultaneity in the treatments is an opportunity only if working simultaneously on more than one task induces a huge increase in productivity (t small). Moreover it is necessary to control the number of possible tasks done simultaneously (k) to avoid important decrease in the performance measures.

B.2 Computation of $P(W < t)$

This section is related to Sections 2.5.1 and 2.5.2. We give details on the derivation of the quantities $\mathrm{P}(W < t, (A, n))$, $\mathrm{P}(W < t, (B, n))$, $\mathrm{P}(W < t, (B', n))$, $\mathrm{P}(W < t, (C, n))$ and $\mathrm{P}(W < t, (M, n))$,

Table B.2: Performance Measures in function of the level of simultaneity ($\lambda = 1$,$\mu = 0.2$, $s = 6$)

	$t = 0.1$			$t = 0.5$			$t = 1$		
k	W_q	W_s	$W_q + W_s$	W_q	W_s	$W_q + W_s$	W_q	W_s	$W_q + W_s$
1	2.94	5.00	7.94	2.94	5.00	7.94	2.94	5.00	7.94
2	0.12	5.50	5.62	0.76	7.50	8.26	5.88	11.00	16.87
3	0.03	6.05	6.08	1.14	11.25	12.39	8.81	15.33	24.14
4	0.01	6.66	6.67	3.23	16.88	20.11	11.75	21.73	33.48
5	0.01	7.32	7.33	16.48	25.31	41.79	29.38	52.10	81.47
10	0.00	11.79	11.79	25.48	38.31	63.79	-	-	-
50	3.32	106.72	110.04	-	-	-	-	-	-

for $n \geq 0$. These quantities are involved in the computation of the cdf of the call waiting time distribution, $P(W < t)$ for $t \geq 0$. They are convolutions of independent exponential random variables with arbitrarily rates.

Consider a distinct sums of exponential random variables with rates $\mu_1, \mu_2, ..., \mu_a$ and a number of terms in each sum equals to $r_1, r_2, ..., r_a$, respectively. Let $F(t)$, for $t \geq 0$, denote the cdf of the summation of all the random variables. From Amari and Misra (1997), we have

$$F(t) = 1 - \left(\prod_{j=1}^{a} \mu_j^{r_j} \right) \sum_{k=1}^{a} \sum_{l=1}^{r_k} \frac{\Psi_{k,l}(-\mu_k) t^{r_k - 1} \exp(-\mu_k t)}{(r_k - l)!(l-1)!}, \tag{B.1}$$

for $t \geq 0$, with

$$\Psi_{k,l}(x) = -\frac{\partial^{l-1}}{\partial x^{l-1}} \left(\prod_{j=1, j \neq k}^{a} (\mu_j + x)^{-r_j} \right),$$

for $x \in \mathbb{R}$. For example for Model 1, we obtain

$$\mathrm{P}(W < t, (A, n-1)) = 1 - (\mu_1 \mu_2 \mu_3)^n \sum_{k=1}^{3} \sum_{l=1}^{n} \frac{\Psi_{k,l}(-\mu_k) t^{n-1} \exp(-\mu_k t)}{(n-l)!(l-1)!},$$

for $n \geq 1$. Using the Leibnitz formulae, we have

$$\Psi_{k,l}(x) = (-1)^l (l-1)! \sum_{i=0}^{l-1} \binom{n}{i} \binom{n}{l-1-i} (\mu_k' + x)^{-(n+i)} (\mu_k'' + x)^{-(n+l-1-i)},$$

with $\mu'_k, \mu''_k \in \{\mu_1, \mu_2, \mu_3 \setminus \mu_k\}$ and $\mu'_k \neq \mu''_k$. We then deduce, for $n \geq 1$, that

$$
\begin{aligned}
&\mathrm{P}(W < t, (A, n-1)) = \\
&1 - (\mu_1\mu_2\mu_3)^n \sum_{l=1}^{n}\sum_{i=0}^{l-1} \left(\frac{(-1)^l \binom{n}{i}\binom{n}{l-1-i}(\mu_3-\mu_1)^{-(n+i)}(\mu_2-\mu_1)^{-(n+l-1-i)} t^{n-1}\exp(-\mu_1 t)}{(n-l)!} \right. \\
&+ \frac{(-1)^l \binom{n}{i}\binom{n}{l-1-i}(\mu_3-\mu_2)^{-(n+i)}(\mu_1-\mu_2)^{-(n+l-1-i)} t^{n-1}\exp(-\mu_2 t)}{(n-l)!} \\
&\left. + \frac{(-1)^l \binom{n}{i}\binom{n}{l-1-i}(\mu_1-\mu_3)^{-(n+i)}(\mu_2-\mu_3)^{-(n+l-1-i)} t^{n-1}\exp(-\mu_3 t)}{(n-l)!} \right).
\end{aligned}
$$

For Model PM, let us take for example the case of a new call that arrives to a system with n ($n \geq 1$) calls (excluding the new call): $n-1$ calls are waiting in the queue, and assume that the one in service is being in the first stage of service. The waiting time in the queue of the new call is the time it takes to clean the system from the n customers ahead of her. Recall that in Model PM, an agent works on emails during the second stage of a call service with probability q. Then, the waiting time of our new call can be represented by Figure B.1. Each branch in B.1 is a possible scenario with a given summation of exponential random variables (with rates inside the circles). The subscript n or k in Figure B.1 indicates that the corresponding random variable is summed n or k times, respectively. A scenario with n exponential stages with rate μ_1, n exponential stages with rate μ_2, n exponential stages with rate μ_3 and k exponential stages of rate μ_0, occurs with probability $\binom{n}{k} q^k (1-q)^{n-k}$, for $0 \leq k \leq n$ and $n \geq 1$.

The cdf of the waiting time of our new call corresponds to the quantity $\mathrm{P}(W < t, (A, n-1))$ ($n \geq 1$). Using Equation B.1, we obtain

$$
\begin{aligned}
P_{(A,n-1)}(W < t) = 1 - (\mu_1\mu_2\mu_3)^n \sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} \mu_0^k \\
\times \left(\sum_{i=1}^{3}\sum_{l=1}^{n} \frac{\Psi_{i,l}(-\mu_i) t^{n-1}\exp(-\mu_i t)}{(n-l)!(l-1)!} + \sum_{l=1}^{k} \frac{\Psi_{0,l}(-\mu_0) t^{k-1}\exp(-\mu_0 t)}{(k-l)!(l-1)!} \right),
\end{aligned}
$$

for $n \geq 1$, with

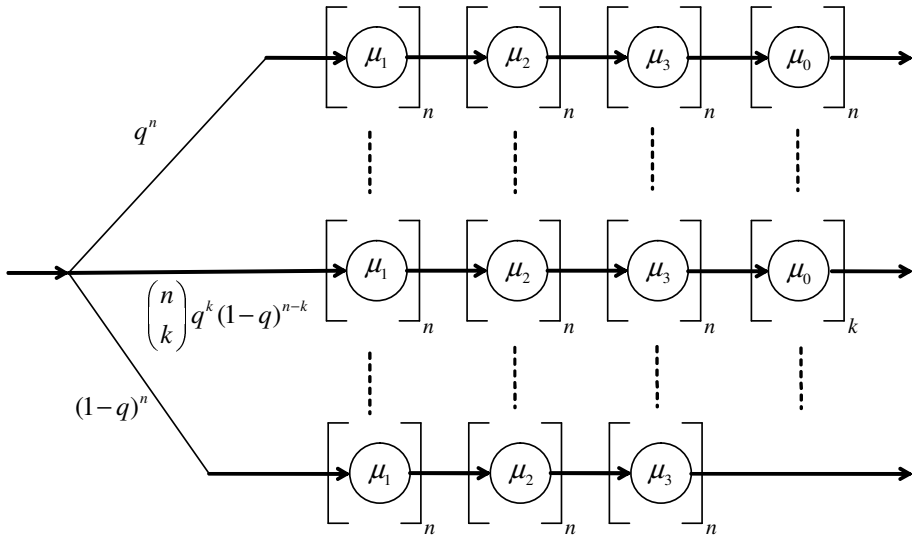


Figure B.1: Waiting time of the new call

$$\Psi_{0,l}(x) = -\frac{\partial^{l-1}}{\partial x^{l-1}} \left(((\mu_1 + x)(\mu_2 + x)(\mu_3 + x))^{-n} \right),$$
$$\Psi_{1,l}(x) = -\frac{\partial^{l-1}}{\partial x^{l-1}} \left((\mu_0 + x)^{-k} ((\mu_2 + x)(\mu_3 + x))^{-n} \right),$$
$$\Psi_{2,l}(x) = -\frac{\partial^{l-1}}{\partial x^{l-1}} \left((\mu_0 + x)^{-k} ((\mu_1 + x)(\mu_3 + x))^{-n} \right),$$
$$\Psi_{3,l}(x) = -\frac{\partial^{l-1}}{\partial x^{l-1}} \left((\mu_0 + x)^{-k} ((\mu_1 + x)(\mu_2 + x))^{-n} \right).$$

We may also write

$$P_{(A,n-1)}(W < t) = 1 - (\mu_1\mu_2\mu_3(q\mu_0 + 1 - q))^n \left(\sum_{i=1}^{3} \sum_{l=1}^{n} \frac{\Psi_{i,l}(-\mu_i) t^{n-1} \exp(-\mu_i t)}{(n-l)!(l-1)!} \right)$$
$$- (\mu_1\mu_2\mu_3)^n \sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} \mu_0^k \left(\sum_{l=1}^{k} \frac{\Psi_{0,l}(-\mu_0) t^{k-1} \exp(-\mu_0 t)}{(k-l)!(l-1)!} \right).$$

for $n \geq 1$. Finally, note that all the other quantities can be computed in the same way.

B.3 Method for Deriving $P(W < t)$ in a Three-Stages Hypoexponential Distribution Service

In this section we propose a method to derive $P(W < t)$ for an arriving customer in a system with a single agent and distributed according to a three-stages distribution with rates μ_1, μ_2 and μ_3. Recall that the queue is infinite, we do note consider any abandonment or retrial and customers are served according to a FCFS rule. The method is based on the Dunford decomposition for the sub-generator matrix. This method conducts to close form formulas which are easier to compute than the general formula of a distribution (when there are distinct sums of exponential distributions with rates λ_1,λ_2,$\cdots$,λ_a and a number of terms in each sum equals to $r_1, r_2, \cdots, r_a$ respectively. The cumulative distribution function for $t \geq 0$ is given by $F(t) = 1 - \left(\prod_{j=1}^{a} \lambda_j^{r_j}\right) \sum_{k=1}^{a} \sum_{l=1}^{r_k} \frac{\Psi_{k,l}(-\lambda_k) t^{r_k - l} \exp(-\lambda_k t)}{(r_k - l)!(l-1)!}$, with $\Psi_{k,l}(x) = -\frac{\partial^{l-1}}{\partial x^{l-1}} \left(\prod_{j=0, j\neq k}^{a} (\lambda_j + x)^{-r_j}\right)$). Yet this method can hardly be generalized to other phase-type distribution which explain why we do not present it as a main contribution.

The distribution is a phase-type distribution. If we denote by M the sub-generator matrix of this distribution. The cumulative distribution function is given by $F(x) = 1 - \boldsymbol{\alpha} e^{xM} \mathbf{1}$ where $\mathbf{1}$ is a column vector of ones of the size k (k stages), e^A is the matrix exponential of A and $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$. The difficulty is to evaluate e^{xM}. We propose a method based on the Dunford decomposition. We consider an arriving customer when n ($n \geq 1$) customers are already in the system, one in service in stage 1 and $n - 1$ is the queue. Thus the distribution is composed with n exponential distributions of rate μ_1, n exponential distributions of rate μ_2 and n exponential distributions of rate μ_3. Note that when the customer in service is in stage 2 or 3 we find one stage less of rate μ_1 and/or μ_2. The sub-generator matrix of this distribution is of dimension $3n \times 3n$. We denote by $m_{i,j}$ its coefficient on line i and column j for $i, j \in \{1, 2 \cdots, 3n\}$. For $i \in \{1, 2, \cdots, n\}$ we have $m_{i,i} = -\mu_1$ and $m_{i,i+1} = \mu_1$. For $i \in \{n+1, n+2, \cdots, 2n\}$ we have $m_{i,i} = -\mu_2$ and $m_{i,i+1} = \mu_2$. For $i \in \{2n+1, 2n+2, \cdots, 3n\}$ we have $m_{i,i} = -\mu_3$ and $m_{i,i+1} = \mu_3$.

When $\mu_1 = \mu_2 = \mu_3$, we have the well known Erlang distribution. Thus we consider here

only cases when at least we find $i, j = 1, 2, 3$ which verify $\mu_i \neq \mu_j$. We present here the most general case is when $\mu_i \neq \mu_j$ for $i, j \in \{1, 2, 3\}$. The other cases can be easily derived from this method. The sub-generator matrix has three different eigenvalues: $-\mu_1$, $-\mu_2$ and $-\mu_3$. Each of these eigenvalues is associated to an eigenspace of dimension 1. The eigenvector associated with $-\mu_1$ is $v_1 = (1, 0, \ldots, 0)$. The eigenvector associated with $-\mu_2$ is $v_2 = (v_{2,1}, v_{2,2}, \ldots, v_{2,3n})$ for which $v_{2,i} = \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{i-1}$ for $i \in \{1, \ldots, n+1\}$ and $v_{2,i} = 0$ else. The eigenvector associated with $-\mu_3$ is $v_3 = (v_{3,1}, v_{3,2}, \ldots, v_{3,3n})$ for which $v_{3,i} = \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{i-1}$ for $i \in \{1, \ldots, n+1\}$, $v_{3,i} = \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{i-(n+1)}$ for $i \in \{n+1, \ldots, 2n+1\}$ and $v_{2,i} = 0$ else. If we denote by e_1, e_2, ..., e_{3n} the usual basis B. In the new basis B' in which we change e_1 with v_1, e_{n+1} with v_2 and e_{2n+1} with v_3 and keep the other vector. We denote by P the transition matrix from basis B to basis B'. We can write the sub generator matrix in this new basis as

$$\begin{bmatrix} E_1 & O & O \\ O & E_2 & O \\ O & O & E_3 \end{bmatrix}, \tag{B.2}$$

in which E_i $(i = 1, 2, 3)$ is a sub-matrix of dimension $n \times n$ equals to

$$\begin{bmatrix} -\mu_i & \mu_i & 0 & \ldots & 0 & 0 \\ 0 & -\mu_i & \mu_i & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & -\mu_i & \mu_i & 0 \\ 0 & 0 & \ldots & 0 & -\mu_i & \mu_i \\ 0 & 0 & \ldots & 0 & 0 & -\mu_i \end{bmatrix}, \tag{B.3}$$

and O is as sub-matrix of dimension $n \times n$ only composed with 0. We can decompose E_i into a

sum of a diagonal matrix (denoted by D_i) and a nilpotent one (denoted by N_i). We have

$$D_i = \begin{bmatrix} -\mu_i & 0 & 0 & \dots & 0 & 0 \\ 0 & -\mu_i & 0 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & -\mu_i & 0 & 0 \\ 0 & 0 & \dots & 0 & -\mu_i & 0 \\ 0 & 0 & \dots & 0 & 0 & -m_i \end{bmatrix}, \tag{B.4}$$

and

$$N_i = \begin{bmatrix} 0 & \mu_i & 0 & \dots & 0 & 0 \\ 0 & 0 & \mu_i & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & \mu_i & 0 \\ 0 & 0 & \dots & 0 & 0 & \mu_i \\ 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix}. \tag{B.5}$$

We denote by D the matrix

$$\begin{bmatrix} D_1 & O & O \\ O & D_2 & O \\ O & O & D_3 \end{bmatrix}, \tag{B.6}$$

and by N the matrix

$$\begin{bmatrix} N_1 & O & O \\ O & N_2 & O \\ O & O & N_3 \end{bmatrix}. \tag{B.7}$$

We have $M = P^{-1}(N + D)P$. Then $e^{xM} = P^{-1}e^{x(N+D)}P$ and because N and D are commuting (Dunford decomposition) we have $e^{x(N+D)} = e^{x(N+D)} = e^{xD} \times e^{xN}$. Because xD is a diagonal

matrix we know that

$$e^{xD} = \begin{bmatrix} e^{xD_1} & O & O \\ O & e^{xD_2} & O \\ O & O & e^{xD_3} \end{bmatrix}, \tag{B.8}$$

and

$$e^{xD_i} = \begin{bmatrix} e^{-x\mu_i} & 0 & 0 & \dots & 0 & 0 \\ 0 & e^{-x\mu_i} & 0 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & e^{-x\mu_i} & 0 & 0 \\ 0 & 0 & \dots & 0 & e^{-x\mu_i} & 0 \\ 0 & 0 & \dots & 0 & 0 & e^{-x\mu_i} \end{bmatrix}, \tag{B.9}$$

for $i = 1, 2, 3$. Since $N_i^n = O$ for $i = 1, 2, 3$, we have $N^n = O$. For $k \geq n$, $N^k = O$. We can write $e^{xN} = \sum_{k=0}^{+\infty} \frac{(xN)^k}{k!}$. The difficulty is to evaluate an infinite sum but because for $k \geq n$ we have $N^k = O$ this sum is in fact a finite one and $e^{xN} = \sum_{k=0}^{n-1} \frac{(xN)^k}{k!}$. We still need to evaluate the coefficients of the matrix N^k for $k \in 0, 1, 2, \cdots, n-1$. We have

$$N^k = \begin{bmatrix} N_1^k & O & O \\ O & N_2^k & O \\ O & O & N_3^k \end{bmatrix}. \tag{B.10}$$

The coefficients on line i and column $i+k$ for $i \in 1, 2, \cdots, n-1$ of N_1^k, N_2^k and N_3^k are μ_1^k, μ_2^k and μ_3^k respectively. The other coefficient of N^k are null. Then

$$\begin{bmatrix} e^{xN_1} & O & O \\ O & e^{xN_2} & O \\ O & O & e^{xN_3} \end{bmatrix}, \tag{B.11}$$

and for $i = 1, 2, 3$

$$e^{xN_i} = \begin{bmatrix} 1 & x\mu_i & \frac{x^2}{2}\mu_i^2 & \dots & \frac{x^{n-2}}{(n-2)!}\mu_i^{n-2} & \frac{x^{n-1}}{(n-1)!}\mu_i^{n-1} \\ 0 & 1 & x\mu_1 & \ddots & \frac{x^{n-3}}{(n-3)!}\mu_i^{n-3} & \frac{x^{n-2}}{(n-2)!}\mu_i^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 1 & x\mu_i & \frac{x^2}{2}\mu_i^2 \\ 0 & 0 & \dots & 0 & 1 & x\mu_i \\ 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}. \tag{B.12}$$

Then for i=1,2,3 we have

$$e^{x(N_i+D_i)} = \begin{bmatrix} e^{-x\mu_i} & e^{-x\mu_i}x\mu_i & e^{-x\mu_i}\frac{x^2}{2}\mu_i^2 & \dots & e^{-x\mu_i}\frac{x^{n-2}}{(n-2)!}\mu_i^{n-2} & e^{-x\mu_i}\frac{x^{n-1}}{(n-1)!}\mu_i^{n-1} \\ 0 & e^{-x\mu_i} & e^{-x\mu_i}x\mu_i & \ddots & e^{-x\mu_i}\frac{x^{n-3}}{(n-3)!}\mu_i^{n-3} & e^{-x\mu_i}\frac{x^{n-2}}{(n-2)!}\mu_i^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & e^{-x\mu_i} & e^{-x\mu_i}x\mu_i & e^{-x\mu_i}\frac{x^2}{2}\mu_i^2 \\ 0 & 0 & \dots & 0 & e^{-x\mu_i} & e^{-x\mu_i}x\mu_i \\ 0 & 0 & \dots & 0 & 0 & e^{-x\mu_i} \end{bmatrix}. \tag{B.13}$$

We need to evaluate $e^{xM} = P^{-1}e^{x(N+D)}P$. In other words we need to write the matrix $e^{x(N+D)}$ in the basis B. In fact the only differences between the basis B and the basis B' are the $n+1$ and $2n+1$ vectors. We denote by f the linear application associated to the matrix $e^{x(N+D)}$. The target is to calculate

$$F(x) = 1 - \boldsymbol{\alpha}e^{xM}\mathbf{1}. \tag{B.14}$$

In this formula we note that we need to sum in fact the coefficients of the first line of e^{xM}. In the basis B we have $\boldsymbol{\alpha}e^{xM}\mathbf{1} = \sum_{k=1}^{3n} < f(e_k)|e_1 >$. We have $f(e_k) = \sum_{i=1}^{k} e^{-x\mu_1}\frac{(x\mu_1)^{k-i}}{(k-i)!}e_i$ for $1 \leq k \leq n$. Then for $1 \leq k \leq n$, $< f(e_k)|e_1 >= e^{-x\mu_1}\frac{(x\mu_1)^{k-1}}{(k-1)!}$. For $2 \leq k \leq n$, $f(e_{k+n}) = \sum_{i=2}^{k} e^{-x\mu_2}\frac{(x\mu_2)^{k-i}}{(k-i)!}e_{n+i} + e^{-x\mu_2}\frac{(x\mu_2)^{k-1}}{(k-1)!}v_2$. Because $v_2 = \sum_{k=1}^{n+1}\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1}e_k$, then $< v_2|e_1 >= 1$ and for $2 \leq k \leq n$, $< f(e_{k+n})|e_1 >= e^{-x\mu_2}\frac{(x\mu_2)^{k-1}}{(k-1)!}$. For $2 \leq k \leq n$, $f(e_{k+2n}) = \sum_{i=2}^{k} e^{-x\mu_3}\frac{(x\mu_3)^{k-i}}{(k-i)!}e_{2n+i} + e^{-x\mu_3}\frac{(x\mu_3)^{k-1}}{(k-1)!}v_3$.

Because $v_3 = \sum_{k=1}^{n} \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1} e_k + \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \sum_{k=n+1}^{2n+1} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-(n+1)} e_k$, then $< v_3|e_1 >= 1$ and for $2 \leq k \leq n$, $< f(e_{k+2n})|e_1 >= e^{-x\mu_3} \frac{(x\mu_3)^{k-1}}{(k-1)!}$. We still need to evaluate $< f(e_{n+1})|e_1 >$ and $< f(e_{2n+1})|e_1 >$. We have $v_2 = \sum_{k=1}^{n+1} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1} e_k$ and $f(v_2) = e^{-x\mu_2} v_2$ then,

$$\sum_{k=1}^{n+1} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1} f(e_k) = \sum_{k=1}^{n+1} e^{-x\mu_2} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1} e_k. \tag{B.15}$$

This equality is equivalent to

$$\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{n} f(e_{n+1}) + \sum_{k=1}^{n} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1} \left(\sum_{i=1}^{k} e^{-x\mu_1} \frac{(x\mu_1)^{k-i}}{(k-i)!} e_i\right) = \sum_{k=1}^{n+1} e^{-x\mu_2} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1} e_k. \tag{B.16}$$

Then $< f(e_{n+1})|e_1 >= -e^{-x\mu_1} \left(\sum_{k=1}^{n} \frac{(x\mu_1)^{k-1}}{(k-1)!} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1-n}\right) + e^{-x\mu_2} \left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{-n}$. We have

$$v_3 = \sum_{k=1}^{n} \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1} e_k + \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \sum_{k=n+1}^{2n+1} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-(n+1)} e_k \tag{B.17}$$

and $f(v_3) = e^{-x\mu_3} v_3$. Thus

$$\begin{aligned} &\sum_{k=1}^{n} \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1} f(e_k) + \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \sum_{k=n+1}^{2n+1} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-(n+1)} f(e_k) \\ &= e^{-x\mu_3} \left(\sum_{k=1}^{n} \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1} e_k + \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \sum_{k=n+1}^{2n+1} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-(n+1)} e_k\right). \end{aligned} \tag{B.18}$$

This implies that

$$\begin{aligned} &\sum_{k=1}^{n} \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1} < f(e_k)|e_1 > + \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} < f(e_{n+1})|e_1 > \\ &+ \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \sum_{k=2}^{n} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-1} < f(e_{k+n})|e_1 > \\ &+ \left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{n} \left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{n} < f(e_{2n+1})|e_1 >= e^{-x\mu_3}, \end{aligned} \tag{B.19}$$

and

$$< f(e_{2n+1})|e_1 > = e^{-x\mu_1}\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n}\sum_{k=1}^{n}\frac{(x\mu_1)^{k-1}}{(k-1)!}\left(\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1-n}-\left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1-n}\right) \tag{B.20}$$

$$- e^{-x\mu_2}\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n}\left(\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{-n}+\sum_{k=2}^{n}\frac{(x\mu_2)^{k-1}}{(k-1)!}\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-1}\right)$$

$$+ e^{-x\mu_3}\left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{-n}\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n}.$$

Consequently when n customers are in the system and one customer is in the first stage of the service we have,

$$\begin{aligned}
&P(W<x) = 1- \\
&[e^{-x\mu_1}\sum_{k=1}^{n}\frac{(x\mu_1)^{k-1}}{(k-1)!}\{1-\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1-n} \\
&+\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n}\left(\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{k-1-n}-\left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{k-1-n}\right)\} \\
&+e^{-x\mu_2}(\sum_{k=1}^{n}\frac{(x\mu_2)^{k-1}}{(k-1)!}(1-\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{k-1-n})-1+\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{-n}+\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n} \\
&-\left(-\frac{\mu_2-\mu_1}{\mu_1}\right)^{-n}\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n}) \\
&+e^{-x\mu_3}\left(\left(-\frac{\mu_3-\mu_1}{\mu_1}\right)^{-n}\left(-\frac{\mu_3-\mu_2}{\mu_2}\right)^{-n}-1+\sum_{k=1}^{n}\frac{(x\mu_3)^{k-1}}{(k-1)!}\right)].
\end{aligned} \tag{B.21}$$

This finishes the demonstration of the method.

B.4 Reminder for the Cardan-Ferrari Method

In Sections 3.4.1 and 3.4.2, we use the Cardan-Ferrari method to solve cubic and quadric equations. Below, we describe the principles of these methods. We also refer the reader to the textbook Gourdon (1994) for more details.

Solution of a cubic equation: Consider the general cubic equation written as

$$x^3 + ax^2 + bx + c = 0, \tag{B.22}$$

with x as variable and the real parameters a, b and c. Changing the variable x into $z = x - a/3$ leads to $z^3 + (b - \frac{a^2}{3})z + c - \frac{ab}{3} + \frac{2a^3}{27} = 0$. Define now p and q as $p = b - \frac{a^2}{3}$ and $q = c - \frac{ab}{3} + \frac{2a^3}{27}$. Then, Equation (B.22) becomes

$$z^3 + pz + q = 0. \tag{B.23}$$

The discriminant denoted by Δ of this equation is $\Delta = q^2 + \frac{4}{27}p^3$. Let us define u, v and j by $u = (\frac{1}{2}(-q + \sqrt{\Delta}))^{1/3}$, $v = (\frac{1}{2}(-q - \sqrt{\Delta}))^{1/3}$ and $j = e^{i\frac{2\pi}{3}}$. Three situations are possible for Δ: $\Delta > 0$, $\Delta < 0$ or $\Delta = 0$.

- If $\Delta > 0$, Equation (B.23) has one real solution $z_1 = u + v$ and two complex solutions $z_2 = ju + \overline{j}v$ and the conjugate of z_2 denoted by $\overline{z_2}$.
- If $\Delta < 0$, then $v = \overline{u}$ and Equation (B.23) has three real solutions: $z_1 = u + \overline{u}$, $z_2 = ju + \overline{ju}$ and $z_3 = j^2u + \overline{j^2u}$.
- If $\Delta = 0$, Equation (B.23) has two real solutions: a simple one z_1 and a double one z_2.

Solution of a cubic equation: Consider the general cubic equation written as

$$x^4 + ax^3 + bx^2 + cx + d = 0,$$

with x as variable and the real parameters a, b, c and d. Changing the variable x into $z = x - a/4$

leads to

$$z^4 + pz^2 + qz + r = 0,$$

with $p = b - \frac{3a^2}{8}$, $q = c - \frac{ab}{2} + \frac{a^3}{8}$ and $r = d - \frac{ac}{4} + \frac{a^2b}{16} - \frac{3a^4}{256}$. The solving method is based on computing the three real numbers α, β and δ such that

$$z^4 + pz^2 + qz + r = (z^2 + \delta)^2 - (\alpha z + \beta)^2. \tag{B.24}$$

Equation (B.24) holds only if

$$(z^2 + \delta)^2 - (z^4 + pz^2 + qz + r) = (2\delta - p)z^2 - qz + (\delta^2 - r), \tag{B.25}$$

is a square. In other words, the discriminant of the right hand side of Equation (B.25) is equals to 0. Then

$$q^2 - 4(\delta^2 - r)(2\delta - p) = -8\delta^3 + 4p\delta^2 + 8c\delta + q^2 - 4pr = 0. \tag{B.26}$$

Equation (B.26), with δ as variable is a cubic equation that can be solved using the method described in the first part of this section. Having in hand δ, we easily deduce α and β. It finally remains to note that the solutions in z of $(z^2 + \delta)^2 - (\alpha z + \beta)^2 = 0$ are $-\frac{\sqrt{-4\delta+4\beta+\alpha^2}-\alpha}{2}$, $\frac{\sqrt{-4\delta+4\beta+\alpha^2}+\alpha}{2}$, $-\frac{\sqrt{-4\delta-4\beta+\alpha^2}+\alpha}{2}$ and $\frac{\sqrt{-4\delta-4\beta+\alpha^2}-\alpha}{2}$.

B.5 Proof of Proposition 2

We start by proving the third statement. It consists on solving the inequality $\overline{\lambda}_2 < \overline{\lambda}_3$ as a function of w^*. This inequality is equivalent to

$$\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}\right)+\frac{\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}\right)^2+\frac{1}{\mu_1^2}+\frac{1}{\mu_2^2}+\frac{1}{\mu_3^2}}{2\left(w^*-\frac{1}{\mu_0}\right)}-\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}+\frac{1}{\mu_0}\right)$$
$$-\frac{\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}+\frac{1}{\mu_3}+\frac{1}{\mu_0}\right)^2+\frac{1}{\mu_1^2}+\frac{1}{\mu_2^2}+\frac{1}{\mu_3^2}+\frac{1}{\mu_0^2}}{2w^*}>0,$$

or

$$-\frac{\frac{1}{\mu_0}\left(w^{*2}+w^*\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)-\sum_{i,j=0}^{3}\frac{1}{\mu_i\mu_j}\right)}{w^*\left(w^*-\frac{1}{\mu_0}\right)}>0. \tag{B.27}$$

We then need to know the sign of $w^{*2}+w^*\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)-\sum_{i,j=0}^{3}\frac{1}{\mu_i\mu_j}$. The equation $w^{*2}+w^*\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)-\sum_{i,j=0}^{3}\frac{1}{\mu_i\mu_j}=0$ with w^* as variable has two real solutions denoted by w_1^* and w_2^*. They are given by

$$w_1^*=-\frac{\sqrt{\frac{4}{\mu_0^2}+\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)\frac{4}{\mu_0}+5\sum_{i=1}^{3}\frac{1}{\mu_i^2}+6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}}+\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}}{2},$$

$$w_2^*=\frac{\sqrt{\frac{4}{\mu_0^2}+\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)\frac{4}{\mu_0}+5\sum_{i=1}^{3}\frac{1}{\mu_i^2}+6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}}-\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)}{2}.$$

Then, the numerator in Inequality (B.27) is strictly negative if and only if w^* is strictly between w_1^* and w_2^*. It is easy to see that $w_1^*<0$. We next prove that the second solution w_2^* is strictly

higher than $\frac{1}{\mu_0}$. We have

$$w_2^* - \frac{1}{\mu_0} = \frac{\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3}\frac{1}{\mu_i^2} + 6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}} - \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1} + \frac{2}{\mu_0}\right)}{2}. \quad \text{(B.28)}$$

Multiplying the right hand side of Equation (B.28) by the following positive quantity

$$2\left(\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3}\frac{1}{\mu_i^2} + 6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1} + \frac{2}{\mu_0}\right)\right)$$

leads to

$$\begin{aligned}
&\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3}\frac{1}{\mu_i^2} + 6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j} - \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1} + \frac{2}{\mu_0}\right)^2 \\
&= 4\sum_{i=1}^{3}\frac{1}{\mu_i^2} + 4\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j} > 0.
\end{aligned}$$

Since $w_2^* > \frac{1}{\mu_0}$ and defining $\overline{w}^*$ as $\overline{w}^* = w_2^*$, Inequality (B.27) holds if and only if $\frac{1}{\mu_0} < w^* < \overline{w}^*$. This finishes the proof of the third statement.

Let us now prove the first statement. It suffices to prove that Model 2 is not optimal for all $\lambda > 0$ if and only if $\overline{\lambda}_2 \leq 0$.

$\boxed{\Leftarrow}$ If $\overline{\lambda}_2 \leq 0$, Model 2 is obviously not optimal for all $\lambda > 0$.

$\boxed{\Rightarrow}$ Assume Model 2 is not optimal for all $\lambda > 0$. Then there are two possible reasons: $\overline{\lambda}_2 \leq 0$ or $R \leq \overline{\lambda}_4 \leq \overline{\lambda}_2 \leq \overline{\lambda}_3$. We prove next that the second inequality is impossible. The inequality $R \leq \overline{\lambda}_4 \leq \overline{\lambda}_2 \leq \overline{\lambda}_3$ is equivalent to $R \leq \overline{\lambda}_4$ and $\overline{\lambda}_2 \leq \overline{\lambda}_3$, because the inequality $\overline{\lambda}_4 \leq \overline{\lambda}_2$ is always true. From the third statement of Proposition 3, we have: $\overline{\lambda}_2 \leq \overline{\lambda}_3$ is equivalent to $\frac{1}{\mu_0} \leq w^* \leq \overline{w}^*$. The inequality $R \leq \overline{\lambda}_4$ is equivalent to

$$\frac{1}{\frac{1}{\mu_{eq}} + \frac{1}{\mu_2}} \leq \frac{2\left(w^* - \frac{1}{\mu_0}\right)}{2\left(w^* - \frac{1}{\mu_0}\right)\left(\frac{1}{\mu_{eq}}\right) + \left(\frac{1}{\mu_{eq}}\right)^2 + \sum_{i=0}^{3}\frac{1}{\mu_i^2}}.$$

So,

$$\frac{2\left(w^* - \frac{1}{\mu_0}\right)}{\mu_2} \geq \frac{1}{\mu_{eq}^2} + \sum_{i=0}^{3} \frac{1}{\mu_i^2},$$

or equivalently

$$w^* \geq \frac{1}{\mu_0} + \frac{\mu_2}{2}\left(\frac{1}{\mu_{eq}^2} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right).$$

Having $R \leq \overline{\lambda}_4$ and $\overline{\lambda}_2 \leq \overline{\lambda}_3$ implies

$$\frac{1}{\mu_0} + \frac{\mu_2}{2}\left(\frac{1}{\mu_{eq}^2} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right) \leq w^* \leq \overline{w}^*.$$

This previous inequality is impossible because $\overline{w}^* < \frac{1}{\mu_0} + \frac{\mu_2}{2}\left(\frac{1}{\mu_{eq}^2} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right)$. The proof is as follows.

We have

$$\begin{aligned}
&\overline{w}^* - \left(\frac{1}{\mu_0} + \frac{\mu_2}{2}\left(\frac{1}{\mu_{eq}^2} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right)\right) \\
&= \frac{\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3} \frac{1}{\mu_i^2} + 6 \sum_{i,j=1; i\neq j}^{3} \frac{1}{\mu_i \mu_j}} - \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)}{2} - \left(\frac{1}{\mu_0} + \frac{\mu_2}{2}\left(\frac{1}{\mu_{eq}^2} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right)\right) \\
&= \frac{1}{2}\left(\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3} \frac{1}{\mu_i^2} + 6 \sum_{i,j=1; i\neq j}^{3} \frac{1}{\mu_i \mu_j}}\right) \\
&- \frac{1}{2}\left(\left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1} + \frac{2}{\mu_0} + \mu_2\left(\sum_{i=0}^{3} \frac{1}{\mu_i} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right)\right)\right)
\end{aligned}$$

Multiplying the last right hand side by the following positive quantity

$$\begin{aligned}
&2\left(\sqrt{\frac{4}{\mu_0^2} + \left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1}\right)\frac{4}{\mu_0} + 5\sum_{i=1}^{3} \frac{1}{\mu_i^2} + 6 \sum_{i,j=1; i\neq j}^{3} \frac{1}{\mu_i \mu_j}}\right) \\
&+ 2\left(\left(\frac{1}{\mu_3} + \frac{1}{\mu_2} + \frac{1}{\mu_1} + \frac{2}{\mu_0} + \mu_2\left(\sum_{i=0}^{3} \frac{1}{\mu_i} + \sum_{i=0}^{3} \frac{1}{\mu_i^2}\right)\right)\right),
\end{aligned}$$

leads to

$$
\begin{aligned}
&\frac{4}{\mu_0^2}+\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}\right)\frac{4}{\mu_0}+5\sum_{i=1}^{3}\frac{1}{\mu_i^2}+6\sum_{i,j=1;i\neq j}^{3}\frac{1}{\mu_i\mu_j}\\
&-\left(\frac{1}{\mu_3}+\frac{1}{\mu_2}+\frac{1}{\mu_1}+\frac{2}{\mu_0}+\mu_2\left(\sum_{i=0}^{3}\frac{1}{\mu_i}+\sum_{i=0}^{3}\frac{1}{\mu_i^2}\right)\right)^2\\
&=-4\mu_2^2(\frac{1}{\mu_0^4}+2\frac{1}{\mu_3}\frac{1}{\mu_0^3}+4\frac{1}{\mu_2}\frac{1}{\mu_0^3}+2\frac{1}{\mu_1}\frac{1}{\mu_0^3}+3\frac{1}{\mu_3^2}\frac{1}{\mu_0^2}+7\frac{1}{\mu_2}\frac{1}{\mu_3}\frac{1}{\mu_0^2}+4\frac{1}{\mu_1^2}\frac{1}{\mu_3}\frac{1}{\mu_0^2}+6\frac{1}{\mu_2^2}\frac{1}{\mu_0^2}\\
&+7\frac{1}{\mu_1}\frac{1}{\mu_2}\frac{1}{\mu_0^2}+3\frac{1}{\mu_1^2}\frac{1}{\mu_0^2}+2\frac{1}{\mu_3^3}\frac{1}{\mu_0}+7\frac{1}{\mu_2}\frac{1}{\mu_3^2}\frac{1}{\mu_0}+4\frac{1}{\mu_1}\frac{1}{\mu_3^2}\frac{1}{\mu_0}+8\frac{1}{\mu_2^2}\frac{1}{\mu_3}\frac{1}{\mu_0}+10\frac{1}{\mu_1}\frac{1}{\mu_2}\frac{1}{\mu_3}\frac{1}{\mu_0}\\
&+4\frac{1}{\mu_1^2}\frac{1}{\mu_3}\frac{1}{\mu_0}+5\frac{1}{\mu_2^3}\frac{1}{\mu_0}+8\frac{1}{\mu_1}\frac{1}{\mu_2^2}\frac{1}{\mu_0}+7\frac{1}{\mu_1^2}\frac{1}{\mu_2}\frac{1}{\mu_0}+2\frac{1}{\mu_1^3}\frac{1}{\mu_0}+\frac{1}{\mu_3^4}+3\frac{1}{\mu_2}\frac{1}{\mu_3^3}+2\frac{1}{\mu_1}\frac{1}{\mu_3^3}+4\frac{1}{\mu_2^2}\frac{1}{\mu_3^2}\\
&+6\frac{1}{\mu_1}\frac{1}{\mu_2}\frac{1}{\mu_3^2}+3\frac{1}{\mu_1^2}\frac{1}{\mu_3^2}+3\frac{1}{\mu_2^3}\frac{1}{\mu_3}+6\frac{1}{\mu_1}\frac{1}{\mu_2^2}\frac{1}{\mu_3}+6\frac{1}{\mu_1^2}\frac{1}{\mu_2}\frac{1}{\mu_3}+2\frac{1}{\mu_1^3}\frac{1}{\mu_3}+\frac{1}{\mu_2^4}+3\frac{1}{\mu_1}\frac{1}{\mu_2^3}+4\frac{1}{\mu_1^2}\frac{1}{\mu_2^2}\\
&+3\frac{1}{\mu_1^3}\frac{1}{\mu_2}+\frac{1}{\mu_1^4})<0.
\end{aligned}
$$

This proves that the inequality $R \leq \overline{\lambda}_4 \leq \overline{\lambda}_2 \leq \overline{\lambda}_3$ is impossible and finishes the proof of the first statement.

The proof of the second statement of Proposition 3 is straightforward. The details are then omitted. It suffices to see that we choose Model 3 if $\overline{\lambda}_2 < \lambda < \overline{\lambda}_3$. If not, Model 3 is chosen if $R \leq \lambda < \overline{\lambda}_3$. This finishes the proof of the second statement and completes the proof of the proposition. □

B.6 Proof of Lemma 1

We want to solve the following inequality in ρ_0:

$$\frac{\partial T}{\partial p}=\mu_0\frac{(1-\rho_1-\rho_2-q\rho_0-\rho_3)(1+\rho_0)}{(1+p\rho_0)^2}>\frac{\partial T}{\partial q}=\mu_0\frac{\rho_0(1-p)}{1+p\rho_0}.$$

This is equivalent to $(1-\rho_1-\rho_2-\rho_3-q\rho_0)(1+\rho_0)-\rho_0(1-p)(1+p\rho_0)>0$, or also

$$(p^2-p-q)\rho_0^2+(p-q-\rho_1-\rho_2-\rho_3)\rho_0+1-\rho_1-\rho_2-\rho_3>0. \tag{B.29}$$

The discriminant for this inequality is $\Delta=(p-q-\rho_1-\rho_2-\rho_3)^2-4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)>0$. Equation (B.29) has then the two following solutions:

$$\frac{\sqrt{(p-q-\rho_1-\rho_2-\rho_3)^2-4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)}-q+p-(\rho_1+\rho_2+\rho_3)}{2(q-p^2+p)}$$

and

$$-\frac{\sqrt{(p-q-\rho_1-\rho_2-\rho_3)^2-4(p^2-p-q)(1-\rho_1-\rho_2-\rho_3)}+q-p+\rho_1+\rho_2+\rho_3}{2(q-p^2+p)}.$$

Since the first solution is positive (denoted by $\overline{\rho}_0$) and the second one is negative, $\rho_0\in[0;\rho_0^*)$, which finishes the proof of the lemma. □

B.7 Expression of the Probability q_2

This section is related to the proof of Theorem 1. Equation (B.30) gives the expression of the probability q_2.

$$
\begin{aligned}
q_2 = &- (((\rho_3+\rho_2+\rho_1-1)r+(\rho_2+\rho_1-1)\rho_3+(\rho_1-1)\rho_2-\rho_1) \\
&\sqrt{(}((\rho_2-1)r^2+(\rho_2-1)r)S\mu_0+\rho_2 r^2+((\rho_2-1)\rho_3+\rho_2^2 \\
&+(\rho_1+1)\rho_2-\rho_1)r+(\rho_2-1)\rho_3+\rho_2^2+\rho_1\rho_2-\rho_1)+((\rho_2-1)\rho_3 \\
&+\rho_2^2+(\rho_1-2)\rho_2-\rho_1+1)r^2S\mu_0^2+((\rho_2\rho_3+\rho_2^2+(\rho_1-1)\rho_2)r^2 \\
&+((2\rho_2-2)\rho_3^2+(3\rho_2^2+(3\rho_1-2)\rho_2-3\rho_1)\rho_3+2\rho_2^3+(3\rho_1-1)\rho_2^2 \\
&+(2\rho_1^2-2\rho_1-2)\rho_2-2\rho_1^2+1)r)S\mu_0+(\rho_2\rho_3^2+(\rho_2^2+(\rho_1+1)\rho_2)\rho_3 \\
&+\rho_2^3+(\rho_1+1)\rho_2^2+(\rho_1^2+\rho_1-1)\rho_2)r+(\rho_2-1)\rho_3^3 \\
&+(2\rho_2^2+2\rho_1\rho_2-2\rho_1-1)\rho_3^2+(2\rho_2^3+(3\rho_1+1)\rho_2^2 \\
&+(2\rho_1^2-2)\rho_2-2\rho_1^2-2\rho_1+1)\rho_3+\rho_2^4+(2\rho_1+1)\rho_2^3 \\
&+(2\rho_1^2+\rho_1-1)\rho_2^2+(\rho_1^3-2\rho_1)\rho_2-\rho_1^3-\rho_1^2+\rho_1)/ \\
&((\rho_2-1)r^3S\mu_0^2+((2\rho_2-1)r^3+((2\rho_2-2)\rho_3+2\rho_2^2+2\rho_1\rho_2-2\rho_1-1)r^2)S\mu_0 \\
&+\rho_2 r^3+((2\rho_2-1)\rho_3+2\rho_2^2+(2\rho_1+1)\rho_2-\rho_1)r^2 \\
&+((\rho_2-1)\rho_3^2+(2\rho_2^2+2\rho_1\rho_2-2\rho_1-1)\rho_3+\rho_2^3+(2\rho_1+1)\rho_2^2 \\
&+\rho_1^2\rho_2-\rho_1^2-\rho_1)r).
\end{aligned}
\tag{B.30}
$$

Appendix C

Appendix of Chapter 4

C.1 Results for the performance Measures

From the expressions of the performance measures found in Section 4.3.1, we can derive some monotonicity results. We conjecture that the email throughput is increasing and neither convex nor concave in u and that the call service level is decreasing and concave in u (for $0 \leq u \leq s$). In the case of equal service rates we can prove Proposition 5.

Proposition 5 *The following holds:*

1. *The email throughput T is strictly increasing and neither convex nor concave in u, for $0 \leq u \leq s$. However the end of the email throughput, for $s-2 \leq u \leq s$ and $s \geq 2$, is concave in u.*

2. *The call service level $P(W < \tau)$ is strictly decreasing and concave in u, for $0 \leq u \leq s$ and $a < 1$ or $u+1 \leq a < s$.*

Proof. Let us prove the first statement. From Equation (4.2), we have

$$T(s,u,a) = \mu \frac{\frac{1}{(u-1)!} + \frac{a}{u!} + \frac{a^2}{(u+1)!} + \cdots + \frac{a^{s-u}}{(s-1)!} + \frac{a^{s-u+1}}{(s-1)!(s-a)}}{\frac{1}{u!} + \frac{a}{(u+1)!} + \frac{a^2}{(u+2)!} + \cdots + \frac{a^{s-u-1}}{(s-1)!} + \frac{a^{s-u}}{(s-1)!(s-a)}} - \lambda,$$

for $0 \leq u \leq s$. Thus $T(s,u,a) = \mu\left(a + \frac{1}{g_u}\right) - \lambda$, with

$$g_u = \frac{1}{u} + \frac{a}{u(u+1)} + \cdots + \frac{a^{s-u-1}}{u(u+1)(u+2)\cdots(s-1)} + \frac{a^{s-u}}{u(u+1)(u+2)\cdots(s-1)(s-a)},$$

for $0 < u \leq s$ (and $T(s,0,a) = 0$). We may write for $0 < u < s$

$$\begin{aligned} g_{u+1} - g_u = &\left(\frac{1}{u+1} - \frac{1}{u}\right) + \left(\frac{a}{(u+1)(u+2)} - \frac{a}{u(u+1)}\right) + \cdots \\ &+ \left(\frac{a^{s-u-1}}{(u+1)(u+2)\cdots(s-1)(s-a)} - \frac{a^{s-u-1}}{u(u+1)\cdots(s-1)}\right) + \left(-\frac{a^{s-u}}{u(u+1)\cdots(s-1)(s-a)}\right). \end{aligned} \tag{C.1}$$

Since each term of the summation in the right hand of Equation (C.1) is strictly negative, $g_{u+1} < g_u$ for $0 < u < s$. Then, g_u is strictly decreasing in u for $0 < u \leq s$. We also have $T(s,1,a) > 0 = T(s,0,a)$. This implies that $T(s,u,a)$ is strictly increasing in u, for $0 \leq u \leq s$. Figure C.1 illustrates that in general the throughput is neither convex nor concave. Let us now prove that the end of the email throughput, for $s-2 \leq u \leq s$, is concave in u. For $s \geq 2$, we have $T(s,s,a) = s\mu - \lambda$, $T(s,s-1,a) = \frac{\mu}{s}(s^2 - s + a) - \lambda$, and $T(s,s-2,a) = \frac{\mu}{s^2-s+a}(s^3 - 3s^2 + 2(a+1)s - 2a + a^2) - \lambda$. This implies $T(s,s-1,a) - T(s,s-2,a) = \frac{\mu(s-1)}{s(s^2-s+a)}(s^2 - a^2)$, and $T(s,s,a) - T(s,s-1,a) = \frac{\mu}{s}(s-a)$, which yields to $T(s,s-1,a) - T(s,s-2,a) = (T(s,s,a) - T(s,s-1,a))\frac{(s-1)(s+a)}{s^2-s+a}$. Since for $s \geq 2$ that $(s-1)(s+a) - (s^2 - s + a) = a(s-2) \geq 0$, we may write $T(s,s-1,a) - T(s,s-2,a) \geq T(s,s,a) - T(s,s-1,a)$. Then the end of the throughput is concave, which finishes the proof of the first statement of the proposition.

In what follows, we prove the second statement of the proposition. Let us define the sequence f_u as $f_u = s!(1 - a/s)C(s,u,a)$, for $0 \leq u \leq s$. Using Equation (4.3), it suffices then to prove that f_u is strictly increasing and convex in u. We have

$$f_u = \left(\frac{a}{s!(s-a)} + \sum_{k=0}^{s-u} \frac{a^{k+u-s}}{(u+k)!}\right)^{-1},$$

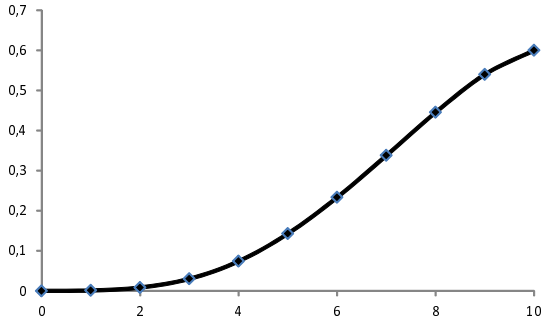


Figure C.1: Email Throughput ($s = 10$, $\mu_0 = \mu = 0.2$, $\lambda = 1.4$)

for $0 \leq u \leq s$. Since for $0 \leq u < s$ we have $\sum_{k=0}^{s-u} \frac{a^{k+u-s}}{(u+k)!} = a^{u-s}\left(\sum_{k=0}^{s-u} \frac{a^k}{(u+k)!}\right)$ and $\sum_{k=0}^{s-(u+1)} \frac{a^{k+u+1-s}}{(u+1+k)!} = a^{u-s}\left(\sum_{k=1}^{s-u} \frac{a^k}{(u+k)!}\right)$, we deduce that $\sum_{k=0}^{s-u} \frac{a^{k+u-s}}{(u+k)!} > \sum_{k=0}^{s-(u+1)} \frac{a^{k+u+1-s}}{(u+1+k)!}$. This implies that $f_u < f_{u+1}$, for $0 \leq u < s$. Then, $P(W < \tau)$ is strictly decreasing in u, for $0 \leq u \leq s$.

We next focus on the proof of convexity of f_u in u (for $s \geq 2$). We prove that $f_u + f_{u+2} - 2f_{u+1} > 0$ for $0 \leq u \leq s-2$. Since $f_u + f_{u+2} - 2f_{u+1} = 2f_u f_{u+1} f_{u+2}\left(\frac{1}{2}\frac{1}{f_{u+1}}\left(\frac{1}{f_u} + \frac{1}{f_{u+2}}\right) - \frac{1}{f_u}\frac{1}{f_{u+2}}\right)$, it suffices to show that $\frac{1}{2}\frac{1}{f_{u+1}}\left(\frac{1}{f_u} + \frac{1}{f_{u+2}}\right) - \frac{1}{f_u}\frac{1}{f_{u+2}} > 0$, for $0 \leq u \leq s-2$. We have

$$\begin{aligned}
&\frac{1}{2}\frac{1}{f_{u+1}}\left(\frac{1}{f_u} + \frac{1}{f_{u+2}}\right) - \frac{1}{f_u}\frac{1}{f_{u+2}} = \\
&\left(\frac{a}{s!(s-a)} + a^{u-s}\sum_{k=1}^{s-u}\frac{a^k}{(u+k)!}\right)\left(\frac{a}{s!(s-a)} + a^{u-s}\left(\frac{1}{2}\left(\frac{1}{u!} + \frac{a}{(u+1)!}\right) + \sum_{k=2}^{s-u}\frac{a^k}{(u+k)!}\right)\right) \\
&- \left(\frac{a}{s!(s-a)} + a^{u-s}\sum_{k=0}^{s-u}\frac{a^k}{(u+k)!}\right)\left(\frac{a}{s!(s-a)} + a^{u-s}\sum_{k=2}^{s-u}\frac{a^k}{(u+k)!}\right),
\end{aligned}$$

or equivalently

$$\begin{aligned}
&\frac{1}{2}\frac{1}{f_{u+1}}\left(\frac{1}{f_u} + \frac{1}{f_{u+2}}\right) - \frac{1}{f_u}\frac{1}{f_{u+2}} = \\
&\frac{1}{2}\frac{a}{s!(s-a)}a^{u-s}\left(-\frac{1}{u!} + \frac{a}{(u+1)!}\right) + \frac{a^{2u-2s}}{2}\left(\frac{a}{u!(u+1)!} + \left(\frac{a}{(u+1)!}\right)^2 - \frac{1}{u!}\sum_{k=2}^{s-u}\frac{a^k}{(u+k)!}\right),
\end{aligned}$$

for $0 \leq u \leq s-2$. We thus need to show that

$$\frac{a}{s!(s-a)}\left(-1+\frac{a}{u+1}\right)+a^{u-s}\left(\frac{a}{(u+1)!}+\frac{a^2}{(u+1)!(u+1)}-\sum_{k=2}^{s-u}\frac{a^k}{(u+k)!}\right)>0. \tag{C.2}$$

for $0 \leq u \leq s-2$.

Case 1: $a \geq u+1$. We have $\frac{a}{s!(s-a)}\left(-1+\frac{a}{u+1}\right) > 0$ and $\frac{a}{(u+1)!}+\frac{a^2}{(u+1)!(u+1)}-\sum_{k=2}^{s-u}\frac{a^k}{(u+k)!} > 0$. The result then follows.

Case 2: $a < 1$. Consider the two last terms in the right hand side of Equation (C.2). We may write

$$\begin{aligned}
&\frac{a^2}{(u+1)!(u+1)}-\left(\sum_{k=2}^{s-u}\frac{a^k}{(u+k)!}\right) \\
&=\frac{a^2}{(u+1)!}\left(\frac{1}{u+1}-\frac{1}{u+2}-\frac{a}{(u+2)(u+3)}-\cdots-\frac{a^{s-u-2}}{(u+2)(u+3)(u+4)\cdots s}\right) \\
&>\frac{a^2}{(u+1)!}\left(\frac{1}{u+1}-\frac{1}{u+2}-\frac{1}{(u+2)(u+3)}-\cdots-\frac{1}{(u+2)(u+3)(u+4)\cdots s}\right).
\end{aligned} \tag{C.3}$$

for $0 \leq u \leq s-2$. We next consider the last line of Equation (C.3) and prove the following statement:

$$\begin{aligned}
P(k): &\frac{1}{u+1}-\frac{1}{u+2}-\frac{1}{(u+2)(u+3)}-\cdots-\frac{1}{(u+2)(u+3)(u+4)\cdots(u+k)} \\
&=\frac{Q_k(u)}{(u+1)(u+2)(u+3)(u+4)\cdots(u+k)},
\end{aligned}$$

for $k \geq 2$, where $Q_k(u)$ is a polynomial in u, for $0 \leq u \leq s-2$, with all coefficients higher than or equal to 1. For $k=2$, we have $\frac{1}{u+1}-\frac{1}{u+2}=\frac{1}{(u+1)(u+2)}$. So, $Q_1(u)=1$ and $P(1)$ is true. Assume

now that $P(k)$ is true for $k \geq 2$. We have

$$\begin{aligned}
&\frac{1}{u+1} - \frac{1}{u+2} - \frac{1}{(u+2)(u+3)} - \cdots - \frac{1}{(u+2)(u+3)(u+4)\cdots(u+k)} \\
&- \frac{1}{(u+2)(u+3)(u+4)} - \cdots - \frac{1}{(u+2)(u+3)(u+4)\cdots(u+k+1)} \\
&= \frac{Q_k(u)}{(u+1)(u+2)(u+3)(u+4)\cdots(u+k)} - \frac{1}{(u+2)(u+3)(u+4)\cdots(u+k+1)} \\
&= \frac{(u+k+1)Q_k(u) - (u+1)}{(u+1)(u+2)(u+3)(u+4)\cdots(u+k)(u+k+1)}.
\end{aligned} \tag{C.4}$$

Then $Q_{k+1}(u) = (u+k+1)Q_k(u) - (u+1)$. Using in addition that $k \geq 2$ and the fact that all the coefficients of $Q_k(u)$ are higher than or equal to one, we deduce that $P(k+1)$ is also true. As a conclusion, the statement $P(k)$ is true for all $k \geq 2$. Using the statement $P(k)$, we state that the last line of Equation (C.3) is strictly positive. Then $\frac{a^2}{(u+1)!(u+1)} - \left(\sum_{k=2}^{s-u} \frac{a^k}{(u+k)!}\right) > 0$, for $0 \leq u \leq s-2$. Recall that we are trying to prove Inequality (C.2). It then remains to prove that

$$\frac{a}{s!(s-a)}\left(-1 + \frac{a}{u+1}\right) + a^{u-s}\frac{a}{(u+1)!} > 0, \tag{C.5}$$

for $0 \leq u \leq s-2$. One may write

$$\begin{aligned}
\frac{1}{s!(s-a)}\left(-1 + \frac{a}{u+1}\right) + a^{u-s}\frac{1}{(u+1)!} &> \frac{1}{s!(s-1)}\left(-1 + \frac{a}{u+1}\right) + a^{u-s}\frac{1}{(u+1)!} \\
&= \frac{1}{s!(s-1)}\frac{a}{u+1} + \frac{a^{u-s}}{(u+1)!} - \frac{1}{s!(s-1)}.
\end{aligned}$$

for $0 \leq u \leq s-2$. Since $u \leq s-2$ and $0 \leq a < 1$, $a^{u-s} > 1$ and $(u+1)! < s!(s-1)$. Therefore $\frac{a^{u-s}}{(u+1)!} - \frac{1}{s!(s-1)} > 0$. We then obtain the convexity result of f_u for $0 \leq u \leq s-2$ and $a < 1$. This completes the proof of the proposition. □

From an extensive numerical study, we observe for both cases (equal or unequal service rates) that $P(W < \tau)$ is also concave in u for $1 \leq a < u+1$. Also, the results of Proposition 5 still hold

Table C.1: Comparison under steady-states assumption (θ=15min)

λ	μ	μ_0	s	γ	Optimal c	Optimal T	ATP T	Difference
0.005	0.2	0.2	1	0.22	0.80	0.04	0.04	0.00%
0.02	0.2	0.2	1	0.23	0.13	0.02	0.02	0.00%
0.05	0.2	0.2	1	0.35	0.06	0.01	0.01	0.00%
0.02	0.2	1	1	1.68	0.20	0.18	0.18	0.00%
0.02	1	0.2	1	0.22	0.21	0.04	0.04	0.00%
0.1	0.2	0.2	5	0.32	4.34	0.76	0.75	1.33%
0.3	0.2	0.2	5	1.25	3.93	0.55	0.50	10.00%
0.5	0.2	0.2	5	1.65	2.75	0.23	0.21	9.52%
0.5	0.2	1	5	6.21	4.29	2.03	1.88	7.98%
0.5	1	0.2	5	0.39	4.17	0.78	0.73	6.85%
0.1	0.2	0.2	10	0.52	9.51	1.80	1.72	4.65%
1	0.2	0.2	10	0.99	8.72	0.80	0.77	3.90%
1.45	0.2	0.2	10	5.28	1.09	0.01	0.01	0.00%
8.2	1	1	10	17.89	5.94	0.38	0.36	5.56%
0.1	0.05	0.05	10	0.06	9.94	0.36	0.33	9.09%
1.3	0.2	1	10	18.67	9.30	3.46	3.41	1.46%
1	1	0.2	10	0.66	9.37	1.78	1.77	0.05%
2	0.2	0.2	28	2.20	27.37	3.29	2.99	10.03%
4	0.2	0.2	28	1.99	25.49	1.39	1.37	1.46%
1	0.1	0.1	28	0.24	27.19	1.74	1.71	1.75%
4	0.27	0.15	28	0.93	26.63	1.89	1.89	0.00%
4	0.17	1	28	196.31	23.21	2.00	1.79	11.73%
17.5	0.2	0.2	100	3.83	97.77	2.37	2.16	9.72%
18	0.2	0.2	100	7.73	93.91	1.65	1.58	4.43%
18.5	0.2	0.2	100	2.30	85.61	0.58	0.54	7.41%
95	1	0.2	100	11.13	88.28	0.82	0.73	12.89%

for the case of different service rates.

C.2 Evaluation of The ATP

In Table C.1 we propose scenarios to compare the throughput found with the ATP method and the optimal throughput. We consider various situations by changing the service rates, the workload on calls and the size of the call center.

Bibliography

Akşin, O., Armony, M., and Mehrotra, V. (2007). The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16:665–688.

Akşin, O. Z. and Karaesmen, F. (2007). Characterizing the performance of process flexibility structures. *Operations Research Letters*, 35:477–484.

Albino, V. and Garavelli, A. (1999). Limited flexibility in cellular manufacturing systems: A simulation study. *International Journal of Production Economics*, 60-61(0):447–455.

Amari, S. and Misra, R. (1997). Closed-form expressions for distribution of sum of exponential random variables. *IEEE Transactions on Reliability*, 46:519–522.

Armony, M. and Maglaras, C. (2004a). Contact Centers with a Call-Back Option and Real-Time Delay Information. *Operations Research*, 52:527–545.

Armony, M. and Maglaras, C. (2004b). On Customer Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules and System Design. *Operations Research*, 52(2):271–292.

Armony, M. and Ward, A. (2010). Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems. *Operations Research*, 58(3):624–637.

Atlason, J., Epelman, M., and Henderson, S. (2008). Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. *Management Science*, 54:295–309.

Bassamboo, A., Randhawa, R., and Van Mieghem, J. (2010). Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing. *Management Science*, 56:1285–1303.

Benjaafar, S. (1995). Performance Bounds for the Effectiveness of Pooling in Multi-Processing Systems. *European Journal of Operational Research*, 87:375–388.

Bernett, H., Fischer, M., and Masi, D. (2002). Blended call center performance analysis. *IT Professional*, 4(2):33–38.

Bhulai, S. and Koole, G. (2003). A Queueing Model for Call Blending in Call Centers. *IEEE Transactions on Automatic Control*, 48:1434–1438.

Bolotin, V. (1994). Telephone circuit holding time distributions. *Proceedings of the 14th International Teletraffic Conference. Labetoulle J.and Roberts J.W., editors*, pages 125–134.

Borst, S., Mandelbaum, A., and Reiman, M. (2004). Dimensioning Large Call Centers. *Operations Research*, 52:17–34.

Borst, S. and Seri, P. (2000). Robust Algorithms for Sharing Agents with Multiple S kills. Working Paper. CWI, Amsterdam, The Netherlands.

Brandt, A. and Brandt, M. (1999). A two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability*, 1:191–210.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association*, 100:36–50.

Charron, S. and Koechlin, E. (2010). Divided Representation of Concurrent Goals in the Huma Frontal Lobes. *Science*, 328:360–363.

Chevalier, P., Shumsky, R., and Tabordon, N. (2004). Routing and Staffing in Large Call Centers with Specialized and Fully Flexible Servers. Université catholique de Louvain, University of Rochester and Belgacom Mobile/Proximus. Working paper.

Choudhury, G., Lucantoni, D., and Whitt, W. (1995). Numerical Solution of Mt/Gt/1 Queues. *Operations Research*, 45:451–463.

Daigle, J. and Lucantoni, D. (1991). Queueing systems having phase-dependant arrival and service rates. Chapter 10 of Numerical Solutions of Markov Chains, Editor: W.J. Stewart, Marcel Dekker, INC., 161-202.

Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., and Avramidis, A. (2007). Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645.

Dux, P., Tombu, M., Harrison, S., Rogers, B., Tong, F., and Marois, R. (2009). Training improves Multitasking Performance by Increasing the Speed of Information Precessing in Human Prefontal Cortex. *Neuron*, 63:127–138.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5:73–141.

Gans, N. and Zhou, Y. (2003a). A Call-Routing Problem with Service-Level Constraints. *Operations Research*, 51:255–271.

Gans, N. and Zhou, Y.-P. (2003b). A call-routing problem with service-level constraints. *Operations Research*, 51:255–271.

Garavelli, A. (2001). Performance analysis of a batch production system with limited flexibility. *International Journal of Production Economics*, 69(1):39 – 48.

Garavelli, A. (2003). Flexibility configurations for the supply chain management. *International Journal of Production Economics*, 85(2):141 – 153.

Garnett, O. and Mandelbaum, A. (2001). An Introduction to Skills-Based Routing and its Operational Complexities. Teaching notes, Technion.

Gladstones, W., Regan, M., and Lee, R. (1989). Division of Attention: The Single-Channel Hypothesis Revisited. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 41:1–17.

Gourdon, X. (1994). *Les maths en tête : Algèbre*. Ellipses, Paris.

Guo, P. and Zipkin, P. (2008). The Effects of Information on a Queue with Balking and Phase-Type Service Times. *Naval Research Logistics*, 55:406–411.

Gurumurthi, S. and Benjaafar, S. (2004). Modeling and Analysis of Flexible Queueing Systems. *Naval Research Logistics*, 51:755–782.

Henderson, S. and Mason, A. (1998). Rostering by Iterating Integer Programming and Simulation. *Winter Simulation Conference*, 1:677–683.

Holman, D., Batt, R., and U., H. (2007). The Global Call Center Report: International Perspectives on Management and Employment. Global Call Centre Research Network.

Hopp, W., Tekin, E., and Van Oyen, M. (2004). Benefits of Skill Chaining in Production Lines with Cross-Trained Workers. *Management Science*, 50:83–98.

Hopp, W. and van Oyen, M. (2004). Agile Workforce Evaluation: A Framework for Cross-Training and Coordination. *IIE Transactions*, 36:919–940.

ICMI (2013). Extreme Engagement in the Multichannel Contact Center: Leveraging the Emerging Channels Research Report and Best Practices Guide. ICMI Research Report.

Jordan, W. and Graves, S. (1995). Principles on the Benefits of Manufacturing Process Flexibility. *Management Science*, 41:577–594.

Jordan, W., Inman, R., and Blumenfeld, D. (2004). Chained Cross-Training of Workers for Robust Performance. *IIE Transactions*, 36:953–967.

Keblis, M. and Chen, M. (2006). Improving customer service operations at amazon.com. *Interfaces*, 36:433–445.

Keilson, J., Sumita, U., and Zachmann, M. (1987). Row-Continuous Finite Markov Chains: Structure and Algorithms. *Journal of the Operations Research Society of Japan*, 30:291–314.

Kleinrock, L. (1975). *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication.

Koole, G. (2013). *Call Center Optimization*. MG Books.

Legros, B., Jouini, O., and Dallery, Y. (2012). A Flexible Architecture for Call Centers with Skill-Based Routing. Working paper. Ecole Centrale Paris.

Legros, B., Jouini, O., and Koole, G. (2013a). Adaptative Call Center Blending. Working paper. Ecole Centrale Paris.

Legros, B., Jouini, O., and Koole, G. (2013b). Call centers with a call back option. Working paper. Ecole Centrale Paris.

Legros, B., Jouini, O., and Koole, G. (2013c). Optimal Email routing in a Multi-Channel Call Center. Working paper. Ecole Centrale Paris.

Mandelbaum, A. and Reiman, M. (1998). On Pooling in Queueing Networks. *Management Science*, 44:971–981.

Manitz, M. and Stolletz, R. (2013). The impact of a waiting-time threshold in overflow systems with impatient customers. *Omega*, 41:280–286.

Marengo, W. (2004). Skill based routing in multi-skill call center. Working Paper. Vrije universiteit, The Netherlands.

Milner, J. and Olsen, T. (2008). Service-Level Agreements in Call Centers: Perils and Prescriptions. *Management Science*, 54:238–252.

Mitrani, I. and Chakka, R. (1995). Spectral Expansian Solution of a Class of Markov Models: Application and Comparison with the Matrix-Geometric Method. *Performance Evaluation*, 23:241–260.

Neuts, M. (1981). *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*. Johns Hopkins University Press, Mineola.

Neuts, M. (1982). Explicit Steady-State Solutions to Some Elementary Queueing Models. *Operations Research*, 30:480–489.

Nomden, G. and van der Zee, D. (2008). Virtual cellular manufacturing: Configuring routing flexibility. *International Journal of Production Economics*, 112(1):439–451.

Pichitlamken, J., A., D., P., L., and Avramidis, A. (2003). Modeling and simulation of a telephone call center. *Proceedings of the 37th Conference on Winter Simulation, New Orleans, LA*, pages 1805–1812.

Pinedo, M., S., S., and J.G., S. (1999). 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*. Kluwer.

Pinker, E. and Shumsky, R. (2000). The Efficiency-Quality Trade-Off of Cross-Trained Workers. *Manufacturing and Service Operations Management*, 2:32–48.

Pollaczeck, F. (1930). Uber eine Aufgabe der Wahrscheinlichkeitstheorie. *Mathematische Zeitschrift*, 32:64–100.

Ren, Z. and Zhou, Y. (2008). Call Center Outsourcing: Coordinating Staffing Level and Service Quality. *Management Science*, 54:369–383.

Robbins, T. and Harrison, T. (2010). Cross Training in Call Centers with Uncertain Arrivals and Global Service Level Agreements. *International Journal of Operations and Quantitative Management*, 16:307–329.

Seelen, L. (1986). An Algorithm for Ph/Ph/c Queues. *European Journal of Operational Reseearch*, 23:118–127.

Sheikhzadeh, M., Benjaafar, S., and Gupta, D. (1998). Machine Sharing in Manufacturing Systems: Total Flexibility versus Chaining. *International Journal of Flexible Manufacturing Systems*, 10:351–378.

Smith, D. and Whitt, W. (1981). Resource Sharing for Efficiency in Traffic Systems. *The Bell System Technical Journal*, 60:39–55.

Sze, D. (1984). A queueing model for telephone operator staffing. *Operations Research*, 32:229–249.

Tekin, E., Hopp, W., and van Oyen, M. (2009). Pooling Strategies for Call Center Agent Cross-Training. *IIE Transactions*, 41:546–561.

Tomlin, B. and Wang, Y. (2005). On the Value of Mix Flexibility and Dual Sourcing in Unreliable Newsvendor Networks. *Manufacturing & Service Operations Management*, 7:37–57.

van Dijk, N. and van Der Sluis, E. (2008). To Pull or not to Pull in Call Centers. *Production and Operations Management*, 17:1–10.

Wallace, R. and Whitt, W. (2005). A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*, 7:276–294.

Index

Abstract

Call centers have been introduced with great success by many service-oriented 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 PhD thesis, we focus on various operations management issues of multi-skill and multi-channel call centers. The objective of our work is to derive, both qualitative and quantitative, results for practical management.

In the first part, we focus on architectures with limited flexibility for multi-skill call centers. The context is that of call centers with asymmetric parameters: unbalanced workload, different service requirements, a predominant customer type, unbalanced abandonments and high costs of cross-training. The most knowing architectures with limited flexibility such as chaining fail against such asymmetry. We propose a new architecture referred to as single pooling with only two skills per agent and we demonstrate its efficiency under various situations of asymmetry.

In the second part, we focus on routing problems in multi-channel call centers. In the first study, we consider a blended call center with calls arriving over time and an infinitely backlogged queue of emails. The call service is characterized by three successive stages where the second one is a break. We focus on the optimization of the email routing to agents. The objective is to maximize the throughput of emails subject to a constraint on the call waiting time. Various guidelines to call center managers are provided. In particular, we prove for the optimal routing that all the time at least one of the two email routing parameters has an extreme value.

In the second study, we examine a threshold policy on the reservation of agents for the inbound calls. We study a general non-stationary model where the call arrival follows a non-homogeneous Poisson process. The optimization problem consists on maximizing the throughput of outbound tasks under a constraint on the waiting time of inbound calls. We propose an efficient adaptive threshold policy easy to implement. This scheduling policy is evaluated through a comparison with the optimal performance measures found in the case of a constant stationary arrival rate, and also a comparison with other intuitive adaptive threshold policies in the general non-stationary case.

In the third study, we consider a call center model with a call back option, which allows to transform an inbound call into an outbound one. The optimization problem consists on minimizing the expected waiting time of the outbound calls while respecting a service level constraint on the inbound ones. We propose a routing policy with two thresholds, one on the reservation of the agents for inbound calls, and another on the number of waiting outbound calls. A curve relating the two thresholds is determined.

Keywords call centers, stochastic models, queuing systems, Markov chains, simulation, scheduling policies, quality of service, transient analysis, skill-based routing, multi-channel call centers

Résumé

Les centres d'appels connaissent un grand succès depuis leur introduction dans les entreprises de service. Ils sont le principal point de contact avec les clients, et une composante essentielle de la majorité des entreprises. L'émergence des centres d'appels à grande échelle a suscité de nombreuses problématiques de management. Dans cette thèse, nous considérons des problématiques de management orientées sur les centres multi-canaux et multi-compétences. L'objectif de notre travail est de trouver des résultats qualitatifs et quantitatifs utiles pour le management.

Dans la première partie, nous considérons les architectures de centres multi-compétences à flexibilité limitée. Le contexte est celui de centres d'appels avec des paramètres asymétriques : charge de travail non équilibrée, différents temps de services, prédominance d'une catégorie de clients, taux d'abandons variables et coûts élevés de la multi-compétence. Les architectures les plus connues avec flexibilité limitée comme chaining ne résistent pas à de telles asymétries. Nous proposons une nouvelle architecture, appelée Single Pooling avec seulement deux compétences par agent et nous démontrons son efficience dans diverses situations d'asymétrie.

Dans la seconde partie, nous nous intéressons aux problèmes de routage dans les centres d'appels multi-canaux. Dans le premier projet, nous considérons un centre avec des appels arrivant au fil du temps et des emails présents en nombre illimité. Le service des appels se fait en trois étapes dont la seconde est une pause pour l'agent. Nous cherchons à optimiser le routage des emails. L'objectif est de maximiser le débit d'emails traités sous contrainte de temps d'attente pour les appels. De nombreuses recommandations sont proposées au manager. En particulier, nous démontrons que pour obtenir un routage optimal il est nécessaire de fixer à une valeur extrême au moins l'un des deux paramètres définissant le routage des emails.

Dans le second projet, nous étudions une politique de seuil de réservation d'agents pour les appels en réception. Nous considérons un cas général de modèle non stationnaire où le processus d'arrivée des appels est Poisson non homogène. Le problème d'optimisation est la maximisation du débit de tâches en émission sous contrainte de qualité de service sur les appels en réception. Nous proposons une méthode efficiente et facile à implémenter de changement adaptatif de seuil. Cette politique est évaluée en comparaison avec les performances optimales trouvées dans le cas particulier de taux d'arrivée constant, et en comparaison avec d'autres méthodes intuitives de changement adaptatif de seuil dans le cas général non stationnaire.

Dans le troisième projet, nous considérons un modèle de centre avec option de rappels. Cette option permet de transformer un appel en réception en un appel en attente d'émission. Le problème d'optimisation est la minimisation du temps d'attente des appels en émission sous contrainte de qualité de service pour les appels en réception. Nous proposons une politique de routage à deux seuils, un sur la réservation d'agents pour les appels en réception et un sur le nombre d'appels en attente d'émission. Nous déterminons une courbe optimale entre ces deux seuils.

Mots clefs centres d'appels, modèles stochastiques, files d'attente, chaine de Markov, simulation, politique d'ordonnancement, qualité de service, analyse transitoire, routage par compétences, centres d'appels multicanaux

Abstract

Call centers have been introduced with great success by many service-oriented 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 PhD thesis, we focus on various operations management issues of multi-skill and multi-channel call centers. The objective of our work is to derive, both qualitative and quantitative, results for practical management.

In the first part, we focus on architectures with limited flexibility for multi-skill call centers. The context is that of call centers with asymmetric parameters: unbalanced workload, different service requirements, a predominant customer type, unbalanced abandonments and high costs of cross-training. The most knowing architectures with limited flexibility such as chaining fail against such asymmetry. We propose a new architecture referred to as single pooling with only two skills per agent and we demonstrate its efficiency under various situations of asymmetry.

In the second part, we focus on routing problems in multi-channel call centers. In the first study, we consider a blended call center with calls arriving over time and an infinitely backlogged queue of emails. The call service is characterized by three successive stages where the second one is a break. We focus on the optimization of the email routing to agents. The objective is to maximize the throughput of emails subject to a constraint on the call waiting time. Various guidelines to call center managers are provided. In particular, we prove for the optimal routing that all the time at least one of the two email routing parameters has an extreme value.

In the second study, we examine a threshold policy on the reservation of agents for the inbound calls. We study a general non-stationary model where the call arrival follows a non-homogeneous Poisson process. The optimization problem consists on maximizing the throughput of outbound tasks under a constraint on the waiting time of inbound calls. We propose an efficient adaptive threshold policy easy to implement. This scheduling policy is evaluated through a comparison with the optimal performance measures found in the case of a constant stationary arrival rate, and also a comparison with other intuitive adaptive threshold policies in the general non-stationary case.

In the third study, we consider a call center model with a call back option, which allows to transform an inbound call into an outbound one. The optimization problem consists on minimizing the expected waiting time of the outbound calls while respecting a service level constraint on the inbound ones. We propose a routing policy with two thresholds, one on the reservation of the agents for inbound calls, and another on the number of waiting outbound calls. A curve relating the two thresholds is determined.

Keywords call centers, stochastic models, queuing systems, Markov chains, simulation, scheduling policies, quality of service, transient analysis, skill-based routing, multi-channel call centers

Résumé

Les centres d'appels connaissent un grand succès depuis leur introduction dans les entreprises de service. Ils sont le principal point de contact avec les clients, et une composante essentielle de la majorité des entreprises. L'émergence des centres d'appels à grande échelle a suscité de nombreuses problématiques de management. Dans cette thèse, nous considérons des problématiques de management orientées sur les centres multi-canaux et multi-compétences. L'objectif de notre travail est de trouver des résultats qualitatifs et quantitatifs utiles pour le management.

Dans la première partie, nous considérons les architectures de centres multi-compétences à flexibilité limitée. Le contexte est celui de centres d'appels avec des paramètres asymétriques : charge de travail non équilibrée, différents temps de services, prédominance d'une catégorie de clients, taux d'abandons variables et coûts élevés de la multi-compétence. Les architectures les plus connues avec flexibilité limitée comme chaining ne résistent pas à de telles asymétries. Nous proposons une nouvelle architecture, appelée Single Pooling avec seulement deux compétences par agent et nous démontrons son efficience dans diverses situations d'asymétrie.

Dans la seconde partie, nous nous intéressons aux problèmes de routage dans les centres d'appels multi-canaux. Dans le premier projet, nous considérons un centre avec des appels arrivant au fil du temps et des emails présents en nombre illimité. Le service des appels se fait en trois étapes dont la seconde est une pause pour l'agent. Nous cherchons à optimiser le routage des emails. L'objectif est de maximiser le débit d'emails traités sous contrainte de temps d'attente pour les appels. De nombreuses recommandations sont proposées au manager. En particulier, nous démontrons que pour obtenir un routage optimal il est nécessaire de fixer à une valeur extrême au moins l'un des deux paramètres définissant le routage des emails.

Dans le second projet, nous étudions une politique de seuil de réservation d'agents pour les appels en réception. Nous considérons un cas général de modèle non stationnaire où le processus d'arrivée des appels est Poisson non homogène. Le problème d'optimisation est la maximisation du débit de tâches en émission sous contrainte de qualité de service sur les appels en réception. Nous proposons une méthode efficiente et facile à implémenter de changement adaptatif de seuil. Cette politique est évaluée en comparaison avec les performances optimales trouvées dans le cas particulier de taux d'arrivée constant, et en comparaison avec d'autres méthodes intuitives de changement adaptatif de seuil dans le cas général non stationnaire.

Dans le troisième projet, nous considérons un modèle de centre avec option de rappels. Cette option permet de transformer un appel en réception en un appel en attente d'émission. Le problème d'optimisation est la minimisation du temps d'attente des appels en émission sous contrainte de qualité de service pour les appels en réception. Nous proposons une politique de routage à deux seuils, un sur la réservation d'agents pour les appels en réception et un sur le nombre d'appels en attente d'émission. Nous déterminons une courbe optimale entre ces deux seuils.

Mots clefs centres d'appels, modèles stochastiques, files d'attente, chaine de Markov, simulation, politique d'ordonnancement, qualité de service, analyse transitoire, routage par compétences, centres d'appels multicanaux